National College of Ireland

Analysis and predictions of CO2 emissions using Neural Networks

MSc Research Project
Data Analytics

Jeet Jaikishan Vyas
Student ID: x19197161

School of Computing
National College of Ireland

Supervisor: Prof. Jorge Basilio

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Jeet Jaikishan Vyas
Student ID:	X19197161
Programme:	Data Analytics **Year:** 2021
Module:	Research Project
Supervisor:	……
Submission Due Date:	16/12/2021
Project Title:	Analysis and predictions on CO2 emissions using Neural Network
Word Count:	6431 **Page Count**: 16

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Jeet Jaikishan Vyas

Date: 16/12/2021

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Analysis and predictions of CO2 emissions using Neural Network

Jeet Jaikishan Vyas
X19197161

Abstract
The research is broadly focused on analyzing how the CO2 emissions and Age dependency ratio indicators showcase the effects on an overall development of certain countries. The data has been fetched from World Bank and world health organisation (WHO) originally and manually was put together to be available on Kaggle. The aim is to understand how this can give a broad insight on the historical and trends on the development prospects of the particular country. The indicators targeted here are in the employment and environment sectors. The indicators are "Age Dependency Ratio" & "Carbon-dioxide Emissions" quoted as CO2 emissions. Research purely focuses on achieving a comparative study on the models and analysis on the behavior and trends of the indicators for a span of 55 years. Models created using neural networks and time series predictions. Analyzing the indicators and predicting the CO2 emissions using economic and CO2 emission indicators. A comparative, analytical and predictive study will be achieved.

1 Introduction

1.1 Background

There are 195 countries in this world, every country has their own pace of development. The countries are categorized as developed, developing and under-developed countries. Each country is driven away from the progress of development because of environmental, political, financial factors. Governments play an important role in gaining the prospective and ensuring the country runs smoothly & tackling the problems with certain policies. It is an effective step towards gaining an understanding how these policies contribute towards development of the country. This will lead them towards the strategy to adapt techniques in order to have control in the future. A small drawback over this can be the comparison might trigger some countries which are being compared to a country having a bigger economy or population. Every research has a productive outcomes and drawbacks, the aim is to accept the changes. For example, comparing India with Australia cannot be considered valid taking into consideration of certain points like population but comparing Australia with Ireland can make more sense. The World Bank and WHO have been collecting data for different indicators. The data allows us to understand and make use of the factors that aim to show how a certain country sustains itself, may it be the economic, financial or environmental crisis. The human and industrial activities give rise to the greenhouse effect and further leads to climate change. Humans need to make it essential to minimize activities that may increase the effects on the economy. As the world is developing at a great pace, the CO2 emissions have been increasing drastically. Carbon dioxide is an important GHG gas and mainly the emissions sources are vehicles, fossil fuel combustion like natural gas, oil. Different industrial sectors

are also responsible for these emissions. The rate on which emissions are more than the rate at which the gas is being absorbed due to deforestation.

1.2 Research Question

How CO2 emissions and Age dependency ratio affects a particular country using neural network models and time series?

1.3 Research Objectives

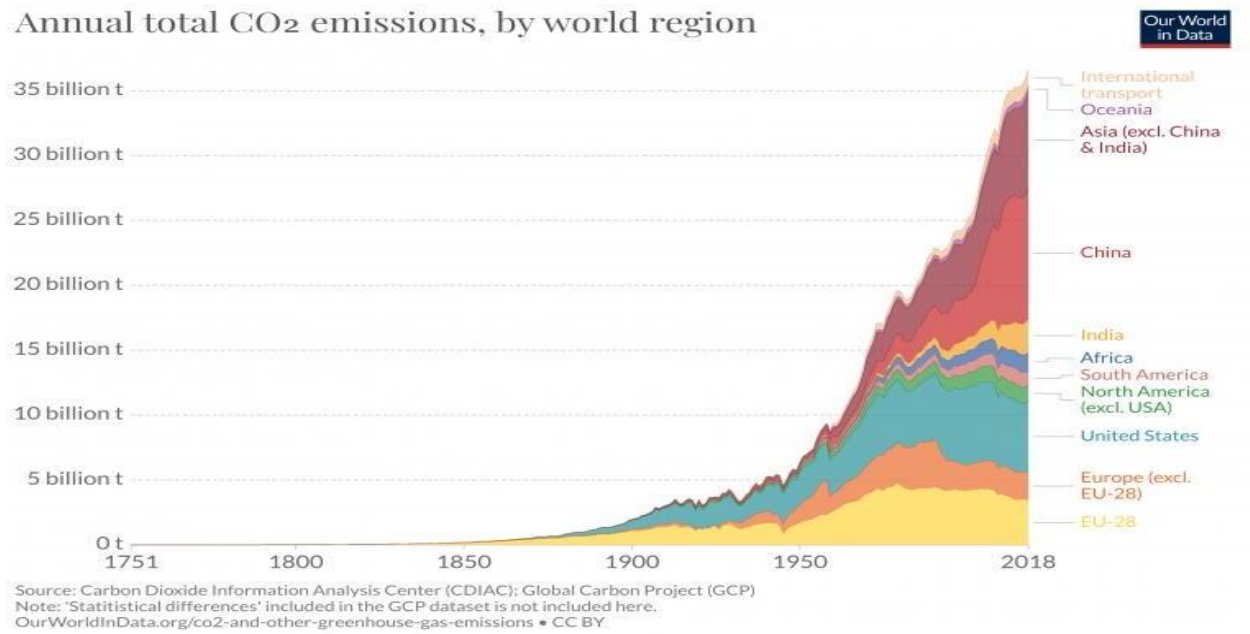

Figure 1: CO2 emissions trend annually by world region

According to Our world in data the CO2 emissions have been manipulating every country which is increasing every year. The statistics shows the continent of Asia excluding China and India having an increase in CO2 emissions every year since 1950 to 2018 from 5 billion tones to reaching nearly 35 billion tons in the span of 68 years. India and China are the most populated countries in the Asian continent and have been experiencing an increase in CO2 emissions compared to any other country in the world. If the two countries are listed in Asia the number is quite worrying. The rest of the continents have a slow but a gradual increase in the CO2 emissions with Africa being on the top followed by South America, North America excluding USA, the list ends with USA and Europe having a steep increase in the emissions. EU or countries which are members of the European Union has the lowest emissions since 1950 to 2018. Europe excluding the EU states has a falling trend in the emissions as well.

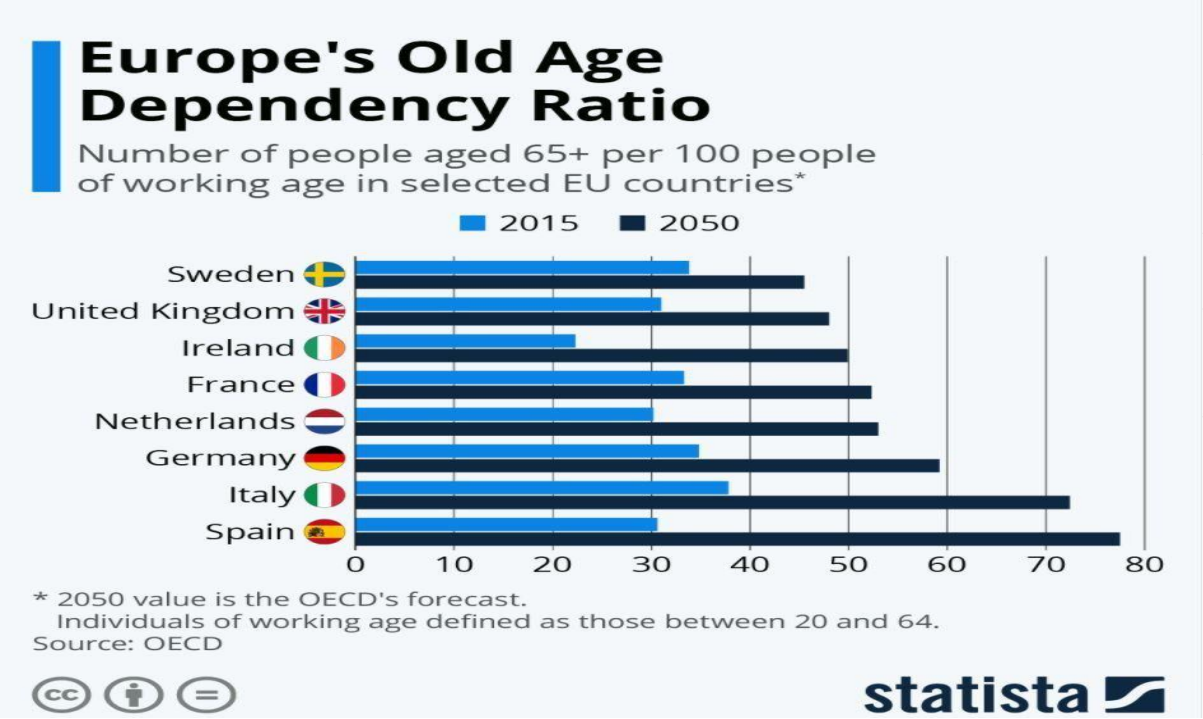

Figure 2: Europe's Old age dependency ratio between 2015 - 2050

The old dependency ratio defines the number of old individuals which is above the age of 59 being dependent on the population engaging in the working class.
According to Statista, the old age dependency ratio is expected to increase in most of the countries in the EU. With Spain predicted to have the highest dependency ratio in 2050 and Sweden having the lowest.

Following objectives are aimed to be achieved:

- Data Analysis and EDA on CO2 emissions and age dependency ratio indicators for the world and 6 countries.
- Visualizing the trend of CO2 emissions metric tons per capita, CO2 emissions from liquid fuel consumption and Age dependency ratio.
- Preparing clusters of the countries and categorizing into developed, under-developed and developing countries using indicators.
- Neural Network model implementation and prediction for CO2 emissions and time series analysis.

2 Related Work

Industrialization and the progress human society is making is imminent that there can be potential risks that can occur. Global warming is the highest risk being considered by experts world-wide. The most harmful and dangerous gases are realized due to human activities and industrial productions. Fossil fuels combustions are increasing at an alarming rate. The major of these due to the greenhouse effect is the CO2 emissions due to different activities.

2.1 Linear Regression

[18] A simple and effective research on analyzing and predicting CO2 emissions in India. The data was statistically calculated and multiple linear regression was applied. The choice of variable was quite clear for the research. They also considered the indicators connected to energy and fuel consumption like electricity consumption. The methodology approach being linear regression the independent and dependent variables were chose effectively. Data for the relevant indicators were fetched between the years of 1995 – 2018. Research conducted was quite simple, the problem objectified and to be solved was quite on point. Predictions for the year 2019 were obtained which were practically fitting the research objectives. There could be different results on the regression models by splitting the data equally or having a 60 – 40 splits. Regression results are quite bias if any one of the variables are omitted, in this case omitting one variable can lead to it being biased. The conclusion of the research clearly stated the increasing risk of the CO2 emissions and how they can keep growing in India.

2.2 Clustering

The agricultural and food industry has kept intact with the introduction of policies in the EU. Keeping the sector regulated the policies have been created. Agricultural subsidies are focused under the CAP policy. Cost reduction and boosting the sector was the priority of the policy. The research conducted [3] on sustainable development and the intellectual growth of countries in the EU, especially in the agricultural sector. Many countries have joined the EU and the CAP policy. Research adapted the cluster analysis approach. The performance has been conducted for the span of 11 years between 2002 – 2013. The analysis involves large groups of observations into smaller homogenous groups. The clusters performed with a split of every 3 years from 2002 – 2013, which suits the method well. Executing the ward method clusters were formed. The features targeted for clustering were livestock production index, food production index, agricultural value and GDP value and lastly the agricultural value per worker. Every feature chosen for a particular aspect, food production index focused on crops that are edible and considered having nutrients. The livestock production focuses on dairy and meat production and distribution. The agricultural sector includes forest, fishing, hunting and fishing. Each cluster created with test. The agricultural sector per worker aims productivity. The Kruskal-Wallis rank test applied for comparing the indicators being different from each other in the clusters obtained. The conclusions for this were focused on the p-value compared to the level of significance. Considering the statistical significance, the sustainable development of the countries in the EU depends on the food sector as a big contribution.

2.3 ANN

Machine learning and deep learning can be two different but co-related artefacts. The models created with the algorithms for both are quite reliable and can help in solving unexpected problems. Machines are bound to outlearn humans therefore there can be a possibility that many practical problems which humans cannot solve may be solved. Complexities from different problems require effective and smooth solutions. [14] have conducted research on a detailed comparison between the machine learning model and ANN models. As the research sounds quite predictable with ANN models performing better but they've effectively proposed. Research focuses on heating and cooling demands, energy consumption and CO2 emissions from offices in Chile. To determine the data, it was considered to keep in mind the

structure of the buildings, conditions of the building, required energy consumption according to the structure. The cooling and heating consumption were gained from the building's heat balance. The approach made the research quite more reliable. The dataset was divided into three parts in order to carry out the predictions, training and validation which was considered a part of data pre-processing. It was observed the research was too focused on getting an adjustment on measuring the error rate and modelling. The values play an important role. There was no mention about any data being null or missing it should've been a part of the pre-processing. The predictor variables aim to creating regression models to gain the cooling and heating demands. The emission factors were added to determine the cooling and heating emission models. Two models were created multi-linear regression and multi-layer perceptron. The results interpreted each indicator as a layer and accordingly models were created.

2.4 ELM and ANN

The relation between CO2 emissions and GDP (Gross Domestic Product) have quite an empirical relationship with the environment. There are changes over time in the economic growth, regulatory policies and technologies. The effects of CO2 emissions from every sector does give a downfall to the progress of a particular country. The greenhouse effect does emit the most harmful gas which is CO2, these emissions are generally by burning of fossil fuels. [13] Research has taken an interesting and simple approach. ELM and ANN which can be described as Extreme Learning Machine which was briefed as a method developed by the research referenced by Huang in their research and ANN a deep learning methodology. The Gaussian process model gives us the probabilities and statistics, the finite collection of random variables giving us a multi-variant normal distribution. Parameters taken into account were sub - sections of CO2 emissions liquid, gaseous and solid fuel consumptions. These were the input parameters. Results will be the GDP growth rate. ELM has been proven to be easy to use and producing good generalizing performances. There is a statistical approach involved in the research that shows the effect of CO2 emissions affecting the GDP. There was room for more methodologies that could have been applied to the research. The results yield satisfaction using ELM approach but it only could show the RSME and co-efficient. Comparing the results to ANN there was a minor difference in the results and both seemed quite satisfactory. A broader approach could have brought some complexity and interesting facts that would have been missed. Research progresses with time as there are changes happening in the source or data that has been selected to be worked on. Gaussian results showed the lowest results. Comparison in all of the three models showcased ELM model being most efficient followed by the ANN and the least progress shown by GP model. To showcase the concerns of global warming and the environmental and economic change the computational models have been developed. The evidence based economic measures and analysis using these models and the evident policies are limited all over the world. Accuracies and results will be subject to change with every year in the future.

2.5 ARMA

Energy based carbon emissions have been a concerning aspect for almost all countries around the world. Forecasting the same helps in getting a productive and analytical knowledge on how there has been significant changes. The number of years that are needed to be researched on can be made through statistical forecasting methods and time series modelling. [10] have conducted the research on a similar concept. The research broadly focuses on the CO2 emissions forecasting for the period of 43 years (1975 - 2018). The countries with the large

emissions have been considered. Data for forecasting was taken for the following countries: China, United States and India. The data was distributed on a timely basis over the past 4 decades. The time series model used for this research was ARMA. The models were created using the MATLAB language. Using the optimum level, the models focused on maximizing model fitting percentages and minimize the final estimation error. These factors enhance the time series modelling and predictions. The predictions were carried out for the following years from 2019 – 2023. The aim of getting the predictions for the proposed forecast the models gave some accurate values. The ARMA model focused on getting the upcoming series of values. The auto regressive part of the research showed regressive values on the existing past values. MA involving linear error terms kept repeating and, in the past, as well. ARMA modelling proved better at forecasting for the research. Linear modelling shows the results of the CO2 emissions increasing constantly up to 2023. This shows the clear picture of the research being showcasing the forecasts are quite accurate. A few points were considered missing from the research, the description of parameters not given briefly. Only the actual data being used and the estimated forecast data was visualized. The methodology of data being split using the model was shown being neutral in order to achieve the minimum estimation error in order to get higher predictions.

2.6 AI and other forecasting models

Artificial intelligence and machine learning has been proving to be a progressive concept in the field of technology. Today AI has been manipulating and predicted to be proving much of human activities. [2] conducted research taking into consideration the use of AI and machine learning in order to get profound results. There has been a great use of machine learning models for the research. The focus is on creating an analysis system predicting the CO2 emissions in smart cities around the world. The list for smart cities weren't mentioned for the research but the perfect definition of being a smart city was showcased. Focusing on the cars and vans having their fuel consumption and emissions from the cargo vehicles. Modelling carried out were multi-variable linear regression modelling backed with supervised machine learning to get the desired output. The training and test data was divided with an 80-20 split, the co-efficient helping to understand the input variable having a great impact. The most interesting part of the research was the prototype build to calculate the CO2 emissions by putting the input of the destination, further getting the arrival time and the amount of emission expected. As the research suggest it was a smart move to creating the prototype to understand the practicality of the research and the predictions helped in taking the necessary measures to be carried out to overcome the excessive emissions.

3 Methodology

The research focuses on getting the most accurate results by carrying out some effective strategies and methodologies. For this research, an inclusive approach of statistical and analytical processes has been used. The combination of analysis and deep learning approach has been carried out for the same. The methods will be giving a boost with different visualizations and algorithms to achieve the research objectives. The different tools have been used to give a broader insight about how the research has been giving an effective impact.

3.1 Data Selection

The data has been fetched from kaggle.com. The data contains CSV (comma separated) files which includes data containing the indicators, countries, series. The data has been verified for public use, which means the data can be used for academic and research purposes by any individual. Link to dataset: https://www.kaggle.com/kaggle/world-development-indicators.

3.2 Data cleaning, processing and Exploratory Data Analysis

This stage includes exploring the data and carrying out the necessary cleaning and EDA (Exploratory Data Analysis). The focus is to work on countries, indicators and series. Each of the data has been used for appropriate steps in order to calculate the number of countries having the appropriate target indicators. The file indicators.csv contains the data of all countries with the respective country code, the various indicators with their indicators code, and the values of each indicator persisting the same, the data had 5656499 rows and 6 columns. The country.csv file contained 247 rows and 31 columns here 247 being the countries around the world from each continent and 31 are the feature information about the countries. The series.csv contained 1345 rows and 20 columns. Null, NaN and duplicated values were calculated and represented. The indicators data did not have any null or duplicated values. In the country data there were only 3 columns with null values other columns had many null values, the columns were filled with 0 and stored in a new data frame named countries. The series data has not been considered for the research but with the aim to get some general information about the values it was seen that there were too many columns with missing data. No processes to clean were performed as the data is not being used for future purposes.

3.3 Visualizations

To understand the basic evaluation and information about certain indicators manipulating some countries over the span of 50 years. The bar graphs are representations to gain the understanding of the indicators having the number of changes every year and the highest and lowest time the indicators were affecting these countries. An overall representation of the world obtained. The countries included for the visualizations were USA, India, Ireland, China, Australia, and Germany. All the countries are from different continents therefore the graphs give us some interesting information about the indicator's behaviors over time. The trends of three indicators were considered for this purpose. As part of the research, age dependency ratio has been considered for visualizations and EDA purposes.

3.4 Clustering

The data of all countries over the span of years were taken into account in order to develop clusters. The clustering achieved using the K-means clustering. This algorithm is simple and unsupervised but quick to use and understand. As the clustering suggest n observations were set in place to define the number of clusters. The clustering was executed with certain indicator codes. Three clusters were formed using K-means clustering described as target k or the number of centroids, the clusters obtained justified the countries into three categories developed, undeveloped and developing.

3.5 Data Transformation & Modelling

As the deep learning approach has been taken into account for the research, the models were 1 layer LSTM with feedforward, GRU, Random Forest with Gradient boost. For neural networks the original LSTM model comprises of a hidden layer and a standard output layer of feedforward. A recurrent lstm model was build where dimensions were matched with the requirements of the lstm model. The last hidden state was obtained and was reshaped. The GRU model constructed. The flattened feed forward algorithm was developed with 5 layers of linear functions with the input dimensions and the time step. For the xgboost model, stacks were formed from the training and validation defined using numpy and the model fitting.

3.6 Evaluation and Results

The bar graphs will show the nature and behavior of the indicators. This helps in analyzing the trend of the indicators for the span of 50 years. The comparison of countries having a fluctuating and the results were evaluated for the models using box plots. Box plots help in representing the validation results and according to the epochs behaving according to the validation loss.

4 Implementation

4.1 Introduction

The aim of the research is to get the overall effects of CO2 emissions and age dependency ratio on the economy of a certain country. The most dangerous effect from the greenhouse are carbon-dioxide emissions. The data for the indicators and value for every country have been taken into consideration for this research. The predictions and modelling for the CO2 emissions will be considered with taking into account the economic indicators in the data.

4.2 Data Cleaning, Pre-processing & EDA

For the research first all the necessary datasets were loaded on python in the anaconda environment. The first step was to observe the nature of the data. The data seemed to be quite clear but a room for doubt for some missing or NaN values was present. To understand this the necessary steps were carried out. Firstly, all the necessary libraries required for the appropriate data analysis, cleaning, visualizing were loaded. The libraries were pandas, numpy, seaborn, matplotlib. Pandas help us to create the data frames and carry out necessary steps of cleaning and EDA (Exploratory Data Analysis) on the datasets. Using pandas, three data frames were created. The three datasets were indicators, country and series. All three data are already formed in comma separated versions (CSV) formats. The basic analysis of the data about the shape, description and columns of the dataset were performed to get appropriate information about all three datasets. The data in indicators.csv was from the year of 1960 – 2015 that is 55 years. For this research, all years were considered for modelling and for the visualizations it was considered from 1960 – 2010. After getting the data loaded and creation of data frames the steps were performed to identify the null and duplicated values which are included in the pandas library. During the implementation, it was found that the indicators data didn't have any null or duplicated values. The countries data which includes information of countries having 31 columns showed NaN values in 27 columns. The NaN

values were filled with 0 and stored in a new data frame. The series data has not been considered for the research after a long observation and understanding the relevance and correlation with other two datasets. There is no data that is useful to carry out any implementation for this research from the series data. The duplicates for indicators and country data showed no duplicates in the data.
To get a clear brief about the target data which is indicators a new data frame was created and the pivot table function was carried out. The reason for this was to get a sophisticated view of the data, the overview of 6 countries was briefed by printing the outputs. India, Germany, United States of America, Ireland, China and Australia.

4.3 Visualizations

The indicators chosen for basic visualizations were CO2 emissions per capita, CO2 emissions from liquid fuel consumptions and another indicator is age dependency ratio. There has been evaluation on age dependency ratio and only general visualizations of bar graphs. The visualizations have been carried out to get information on how the indicators have been behaving over the span of 50 years from 1960 – 2010. The indicators showed us that there has been a different flow of CO2 emissions metric ton per capita in different countries, firstly the emissions for all over the world was visualized and the results showed that there has been a sequential increase in the emissions beginning from 3 per metric tons in 1960 and increasing up to 5 metric tons in 2010. With every country having their own share of emissions and the broadness of their economy shows how the emissions will be increasing. It was observed that there was an increase in every country. India, Australia, Ireland and China had a great increase in the emissions between 1960 and 2010. USA had a great jump in emissions in early 1960 – 1980 and but as time it was observed a slight fall in the emissions between 2000 and 2010. The slow and gradual increase was seen in India, Australia and China. In case of Ireland, there was an increase in the emissions until 2000 and there was steep fall until 2010. The indicator for age dependency ratio was being reviewed about how age dependency has been affecting the economy and increasing in the specific countries. The dependency ratio basically shows the population engaged into work labor and the population not engaged into working class. The pressure of the population is calculated with the dependency ratio formula. For this research this indicator has been only used for the descriptive purposes on how long the indicator has been affecting the countries. As the graphical visualizations first showed the world having a high level of dependency of 80% in 1960, there was deep fall in the dependency ratio up to 60-65% in 2010. Australia, USA and Ireland had quite correlated rise and drop. The percentage was quite high in these three countries in the year 1960 of about 80%, but there was a fall seen between 2000 – 2005 going to 40%. The lowest fall was seen in China followed by India. The third indicator visualized is CO2 emissions from liquid fuel consumptions. This indicator states the number of emissions from vehicles. The bar graphs for the world showed rise in the CO2 emissions from liquid fuel consumptions. The overall graph shows quite low effects in 1960 but a rise in the emissions every 10 years, until 2010 the consumption was the highest comparing all the years. USA, China, Australia, India and Ireland all the countries showed an overall increase in the consumption. India and China had a low level of consumption between 1960 – 1980, whereas the other countries had a significant high level of consumption during the same years.

4.4 Clustering

	Cluster_No	CountryName	GB.XPD.RSDV.GD.ZS	EG.ELC.ACCS.ZS	SL.UEM.TOTL.ZS	FB.CBK.BRCH.P5	EN.ATM.CO2E.PC	BX.KLT.DINV.WD.GD.ZS	NY.GD
1	3	Albania	-0.748576590814535	0.62968400745767	1.03148696542579	-0.0373742858317785	-0.506222572777328	0.413647390643372	-0.62
2	1	Algeria	-0.650714231057733	0.552874333734435	1.82097933959629	-0.826990805588521	-0.408183429981856	-0.625845549033526	-0.61
3	3	Argentina	-0.380961215029682	0.402288088156737	0.452429137533181	-0.394019663783785	-0.252480826456477	-0.465848726521276	-0.28
4	3	Armenia	-0.613156854721408	0.559524706188324	2.2832096904968	-0.185667146504454	-0.563021818630181	0.345044840847068	-0.65
5	2	Australia	1.39240943032681	0.62968400745767	-0.362460523721972	0.522178354020915	1.18567246594506	-0.285849412083147	1.268
6	2	Austria	1.77956178550916	0.62968400745767	-0.815983258496209	-0.369682375698632	0.231841260579462	-0.358821810477908	1.317
7	3	Azerbaijan	-0.574533373425018	0.52834572067694	-0.334160707732188	-0.618523814058039	-0.160025772473086	3.03185694104305	-0.54
8	3	Belarus	-0.0464430347786825	0.62968400745767	-0.445183064632168	-0.880426881185994	0.0978569684570612	-0.348312153622441	-0.47
9	2	Belgium	1.36924109944893	0.62968400745767	-0.146946520719176	1.40967206887037	0.787215941182725	2.34363529033112	1.250
10	1	Bolivia	-0.56116694130892	-0.224223815621695	-0.807275614689108	-0.664064240477338	-0.606248778949638	-0.147208191488676	-0.73
11	1	Botswana	-0.430975792659213	-1.38660694818389	2.02851134658874	-0.654123264972739	-0.5133707192891	-0.0343344065606719	-0.57
12	3	Brazil	0.345295583348744	0.522391331868258	-0.196289788914677	1.25352024883268	-0.539570438107248	-0.385429446075053	-0.46
13	3	Brunei Darussalam	-0.861864685299337	-0.400543633073901	-0.88419308394863	0.0973384141389343	2.13535765982875	-0.142918009657846	1.013
14	3	Bulgaria	-0.300732553724213	0.62968400745767	0.770983501425606	2.99079300640579	0.229311579703704	0.79837796547433	-0.46
15	1	Burkina Faso	-0.650210659178231	-2.5695885010266	-1.05544326856063	-0.993670982191342	-0.71885741074651	-0.679823980123767	-0.80
16	1	Burundi	-0.712516871749429	-2.77708012158794	-0.295701974832479	-0.973631309100731	-0.722581982288473	-0.777032131816474	-0.81
17	1	Cabo Verde	-0.80478152182017	-0.661815662938619	0.109928771769289	0.287851208365461	-0.683258577459784	0.300995457801896	-0.65
18	1	Cambodia	-0.831467326293067	-2.04786465389981	-1.34206963440904	-0.880827737224098	-0.713411160360755	0.611031901710481	-0.79
19	2	Canada	1.26667484088268	0.62968400745767	-0.138964521248721	0.185463695571777	1.36853382791948	-0.292777071476183	1.291
20	3	Chile	-0.48313228616491	0.562293566591593	-0.186130872004752	-0.268941564905175	-0.359586417352913	0.242881798477439	-0.37
21	3	China	0.573975306110912	0.558194631697547	-0.776798893370217	-0.673130930909506	-0.426339635699707	-0.105272394945545	-0.71
22	3	Colombia	-0.663565245216988	0.476282082854627	0.638917702700222	2.7074909551312	-0.5321461769174	-0.317178638706879	-0.60
23	3	Costa Rica	-0.441533905553102	0.512866047738369	-0.482190529619208	-0.0321473703586585	-0.58257169617321	-0.0284483407022163	-0.51
24	3	Croatia	0.11586626977231	0.62968400745767	0.636015146114262	0.664145242833117	-0.113215953491682	0.214927569387205	0.239

Figure 1: Clusters Formation using KMeans Algorithm

The next part of the research is creating the cluster of all the countries. The K-means clustering algorithm was applied on the data. Loading the libraries is the first step. Three libraries loaded for carrying out the clustering were readr, reshape and cluster. The indicators data was loaded. 31 indicators were considered to in order to create the clusters. They were declared in the form of Indicator code. As the algorithm suggests the value for k taken is 3. The countries will be categorized as developed, developing and under-developed. The transformation of the data was done by achieving an aggregated form using the cast function. The rows were transformed to columns and the null values were omitted if there were any. The data was standardized after conversion. The sapply function was applied to get an output of the data in a vector. This makes the data standardized and ready for clustering. The k means algorithm applied with 42 seeds to creating a uniform distribution with the number of seeds declared. The algorithm was applied declaring the transformed and standardized form of the data and mentioning 3 clusters. The clusters plot was carried out by declaring the transformed data into a data frame. Three subsets created declaring one cluster for each category from first to third cluster. The output gave the cluster number in which the country was placed and alternatively the country name. All the three categories were printed in the output achieved. The clusters formed were (38, 29, 50). All the countries were assumed to be in a certain category of cluster. As the total number of countries were 247, the indicators data consisted records from all overall world and also from certain continents like Arab world which explains the countries in the Middle East. The indicators chosen at the beginning to perform the clustering didn't have the data for this type of country description. Therefore, all the countries containing these indicators got selected.

4.5 Modelling

The models were created in neural networks. Using pytorch time series analysis and deep learning models created using CO2 emissions and economic indicators. The libraries for deep learning modelling and time series which in our case is torch loaded in python. Some anonymous packages were loaded for creation of progress bars. The first step

was to create a data of the country code along with the country names and both are part of the features for creation and training of the models. This was preparing the data, after this is the pre-processing stage. For pre-processing a dictionary created for each of the countries and indicators as it is a dataframe with shape. A statistical analysis of mean and standard deviation by using the country and indicator code. For the research the target features are the indicator name, indicator code, country name, country code and values for every indicator. An exploratory data analysis was carried out using the countries and indicators data. For EDA first the total countries were shown then countries having a no values for currency unit were filtered out. The results showed 214 countries out of the total 247 countries. Countries with population of more than 3 million were calculated. The output shows 134 countries have population more than 3 million. As the NaN values in countries data were filled with 0, the results were changed. Before carrying out the time series implementation the data is prepared to have range, time stamp and valid prediction. The features and target values are economic indicators and of CO2 emissions indicators respectively.

The creation of the training and validation of the data which are the target values and features. Three models will be implemented for the research. 1D LSTM (Long short-term memory) with shallow feed forward, GRU (Gated Recurrent Unit), Random Forest with Gradient Boosting. A Feed Forward neural network algorithm is being implemented but the model is not being trained for this research. The training step includes declaration of train, validation, loss function, optimizer and device. The train and validation function will calculate the features of the target features which is the economic and CO2 emissions indicators. The features are used in the form of indicator codes. The torch.optim module was loaded in the beginning in order to implement various optimizer algorithms, the optimize function was used to optimize the models that are being implemented. The train_dl will be having all the observations which will be trained. The loss_fn is the function used to calculate the average loss after running the model's algorithm. The LSTM model from the nn module was fetched, a recurrent model will be created. After creating the LSTM model, the GRU model was created this is also a recurrent model. After these models a Flattened Feed Forward model was created with 5 layers. The three recurrent models were created. After this, The CO2_cols were categorised which had three indicators which are the total CO2 emission per kilotons, CO2 emission metric ton per capita. The economic indicators chosen for the research were mentioned and the ones not used are commented off. The number of training and validation observations for the time series came out as 2211 for training and 565 for validations. Next, the models were trained, optimized and loss function was calculated. The models were implemented LSTM at first, then GRU followed by Feed Forward and random forest with gradient boosting. The model was fit with the training loop. The epochs were set to 250 with an interval of 50. The history declares to show the training and validation loss. After getting the training loop ready, the model fit was implemented.

4.6 Conclusion

The Final step for the implementation were creation of validation box plots to calculate the validation loss for the models. First the learning curve was achieved drawing the most 50 epochs returning the train and validation loss. The second representation to show the validation loss of the models which are stored in history. An average loss of was achieved, which is not fixed. Changes in the number of epochs and the features led to difference in getting the average loss. As the data being quite huge and all the countries not carrying all the necessary indicators few countries may have missing values. The features chosen for this research and modelling carried all the necessary data. The algorithms used have handled to

look if there are any missing values of indicators or the countries. A validation loss of was achieved for the gradient boosted model. The validation loss achieved are low.

5 Evaluation & Results

```
1 xgb_model = XGBRegressor()
2 xgb_model.fit(train_np_x, train_np_y)
```

```
XGBRegressor(base_score=0.5, booster='gbtree', colsample_bylevel=1,
             colsample_bynode=1, colsample_bytree=1, enable_categorical=False,
             gamma=0, gpu_id=-1, importance_type=None,
             interaction_constraints='', learning_rate=0.300000012,
             max_delta_step=0, max_depth=6, min_child_weight=1, missing=nan,
             monotone_constraints='()', n_estimators=100, n_jobs=8,
             num_parallel_tree=1, predictor='auto', random_state=0, reg_alpha=0,
             reg_lambda=1, scale_pos_weight=1, subsample=1, tree_method='exact',
             validate_parameters=1, verbosity=None)
```

Figure 1: Random Forest with Gradient Boosted Model Fit

The above figure shows the gradient boosting model fitting with the training splits observations.

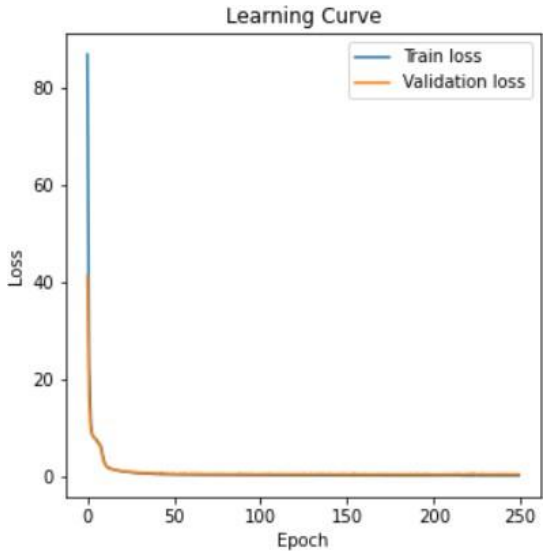

Figure 2: Learning Curve for the training and validation

The training and validation loss with 250 epochs with an interval of 50 for the training and validation observations for the neural network models. The curve shows the train and validation loss for the epochs being implemented,
The 1layer LSTM model and random forest with gradient boosting model showed satisfactory outcomes.

```
Gradient boosted model's validation loss: 0.4001006794492151
```

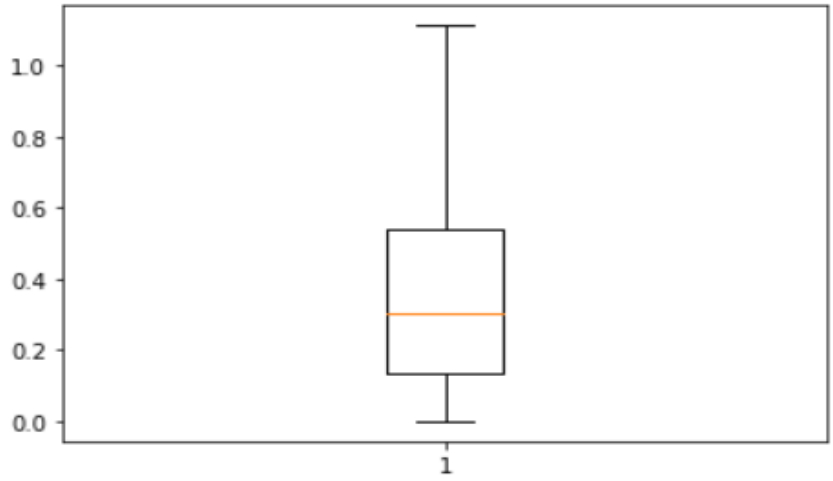

Figure 3: Gradient Boosted Model

The Figure shows the validation loss calculated on the gradient boosted model, the validation loss

```
1 print("Average validation loss of final 5 epochs: {:2f}".format(np.mean(history['val_loss'][-5:])))
Average validation loss of final 5 epochs: 0.470243
```

Figure 4: Average loss for last 5 epochs

```
Gradient boosted model's validation loss: 0.4740619944533844
```

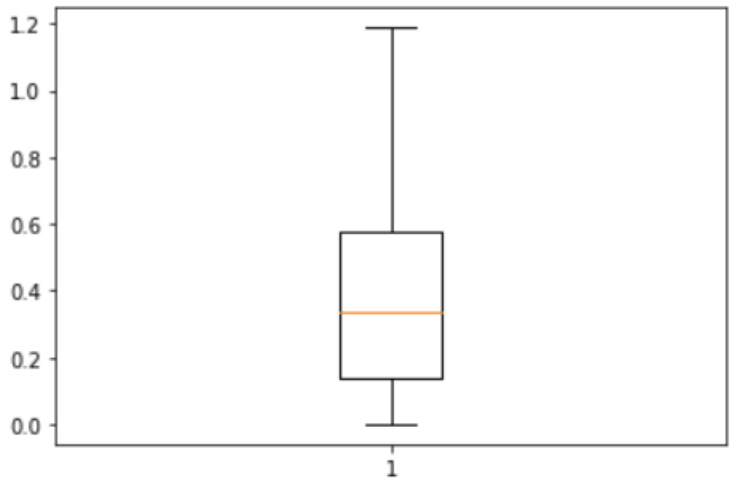

Figure 5: Gradient Boosted Model's validation loss

The validation for the last 5 epochs in history which is the train and validation loss for the models shows an average loss of 0.47 and the validation loss achieved in the gradient boosted model is 0.47.

The validation loss being less tells the models are under fitting or there might be more room for the neural networks to run the epochs. There is a possibility of having fluctuating validation losses due to this kind of fitting.

6 Acknowledgement

My sincere thanks to Prof. Jorge Basilio for his constant support and guidance through the journey of conducting this research. Lastly, I want to thank my family and friends for their motivation and encouragement for conducting this research.

7 Conclusion & Discussion

The overall research was focused on CO2 emissions and Age dependency ratio affecting the particular country. There was all the necessary data for the indicators as well as the countries. The methodology applied for analysis and determining the trends of the indicators were quite useful. It gave a great insight of the indicators behavior and how it has been changing in span of 50 years. The neural network models are mainly focused on simplifying the methods through which the brain processes any information. It contains processing units which contain layers. The models used for the research are recurrent, but for this research these models are not effective. Simpler machine learning models may give a better understanding and accuracies for the objective we are aiming to achieve. As there needs to be more data to let these models show their efficiency the data is not measured to be enough therefore there are no better results. GRU and 1d convolution model might work better taking the better advantage of the time series and the relationship with the time. As every country's development is unique due to different factors there can be a reason for the models showing low efficiency. The estimated prediction for CO2 emissions ranged between 0.4 and 1.2.

8 References

[1] Ho, T.C., Mat, S.C.K.M.Z. and San, L.H., 2015, August. A prediction model for CO2 emission from manufacturing industry and construction in Malaysia. *In 2015 International Conference on Space Science and Communication (IconSpace)* (pp. 469-472). IEEE.

[2] Yeasmin, S., Syed, S.N.J., Shmais, L.A. and Al Dubayyan, R., 2020, November. Artificial Intelligence-based CO2 Emission Predictive Analysis System. *In 2020 International Conference on Artificial Intelligence & Modern Assistive Technology (ICAIMAT)* (pp. 1-6). IEEE.

[3] Reiff, M., Ivanicova, Z. and Surmanova, K., 2018. Cluster analysis of selected world development indicators in the fields of agriculture and the food industry in European Union countries. *Agricultural Economics, 64(5)*, pp.197-205.

[4] Zhu, H.M., You, W.H. and Zeng, Z.F., 2012. Urbanization and CO2 emissions: A semi-parametric panel data analysis. *Economics Letters*, 117(3), pp.848-850.

[5] Wang, Z., Zhang, B. and Wang, B., 2018. The moderating role of corruption between economic growth and CO2 emissions: evidence from BRICS economies. *Energy*, 148, pp.506-513.

[6] Maestas, N., Mullen, K.J. and Powell, D., 2016. *The effect of population aging on economic growth, the labor force and productivity (No. w22452)*. National Bureau of Economic Research.

[7] Antonakakis, N., Chatziantoniou, I. and Filis, G., 2017. Energy consumption, CO2 emissions, and economic growth: An ethical dilemma. *Renewable and Sustainable Energy Reviews*, 68, pp.808-824.

[8] Park, C.Y. and Mercado, R., 2015. *Financial inclusion, poverty, and income inequality in developing Asia*. Asian Development Bank Economics Working Paper Series, (426).

[9] Martínez-Zarzoso, I. and Maruotti, A., 2011. *The impact of urbanization on CO2 emissions: evidence from developing countries*. Ecological Economics, 70(7), pp.1344-1353.

[10] Al-Haija, Q.A. and Smadi, M.A., 2020, June. *Parametric prediction study of global energy-related carbon dioxide emissions*. In 2020 International Conference on Electrical, Communication, and Computer Engineering (ICECCE) (pp. 1-5). IEEE.

[11] Kasman, A. and Duman, Y.S., 2015. CO2 emissions, economic growth, energy consumption, trade and urbanization in new EU member and candidate countries: a panel data analysis. *Economic modelling*, 44, pp.97-103.

[12] Cogoljević, D., Alizamir, M., Piljan, I., Piljan, T., Prljić, K. and Zimonjić, S., 2018. A machine learning approach for predicting the relationship between energy resources and economic development. *Physica A: Statistical Mechanics and its Applications*, 495, pp.211-214.

[13] Marjanović, V., Milovančević, M. and Mladenović, I., 2016. Prediction of GDP growth rate based on carbon dioxide (CO2) emissions. *Journal of CO2 Utilization*, 16, pp.212-217.

[14] Pino-Mejías, R., Pérez-Fargallo, A., Rubio-Bellido, C. and Pulido-Arcas, J.A., 2017. Comparison of linear regression and artificial neural networks models to predict heating and cooling energy demand, energy consumption and CO2 emissions. *Energy*, 118, pp.24-36.

[15] Kadam, P. and Vijayumar, S., 2018, April. *Prediction Model: CO 2 Emission Using Machine Learning*. In 2018 3rd International Conference for Convergence in Technology (I2CT) (pp. 1-3). IEEE.

[16] Colby, S.L. and Ortman, J.M., 2015. *Projections of the Size and Composition of the US Population: 2014 to 2060*. Population Estimates and Projections. Current Population Reports. P25-1143. US Census Bureau.

[17] Magazzino, C., 2016. The relationship between real GDP, CO2 emissions, and energy use in the GCC countries: A time series approach. *Cogent Economics & Finance*, 4(1), p.1152729.

[18] Tanania, V., Shukla, S. and Singh, S., 2020, January. *Time Series Data Analysis and Prediction of CO2 Emissions*. In 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence) (pp. 665-669). IEEE.

[19] Sadorsky, P., 2014. The effect of urbanization on CO2 emissions in emerging economies. *Energy Economics*, 41, pp.147-153.

[20] Traver, M.L., Atkinson, R.J. and Atkinson, C.M., 1999. Neural network-based diesel engine emissions prediction using in-cylinder combustion pressure. *SAE transactions*, pp.1166-1180.