**Prediction Of CO2 Emissions By Country Using IBM Watson**

1. **Introduction**

**Overview**

Carbon emissions and environmental protection issues have brought pressure from the international community during Chinese economic development. Recently, Chinese Government announced that carbon emissions per unit of GDP would fall by 60–65% compared with 2005 and non-fossil fuel energy would account for 20% of primary energy consumption by 2030.The Beijing-Tianjin-Hebei region is an important regional energy consumption center in China, and its energy structure is typically coal-based which is similar to the whole country.Therefore, forecasting energy consumption related to carbon emissions is of great significance to emissions reduction and upgrading of energy supply in the Beijing-Tianjin-Hebei region. Thus, this study thoroughly analyzed the main energy sources of carbon emissions including coal, petrol, natural gas, and coal power in this region.

Carbon footprint has become a widely used term and concept in the public debate on responsibility and abatement action against the threat of global climate change. It had a tremendous increase in public appearance over the last few months and years and is now a buzzword widely used across the media, the government and in the business world. It is a measure of the total amount of carbon dioxide released into the atmosphere in the given time frame that is directly or indirectly caused by an activity to provide service or product . A carbon footprint is a measure of the amount of carbon dioxide emitted through the combustion of fossil fuels. In the case of a business organization, it is the amount of CO2 emitted either directly or indirectly as a result of its everyday operations. It also might reflect the fossil energy represented in a product or commodity reaching the market. The carbon footprint of U.S. households is About 5 times greater than the global average, which is approximately 10 tons CO2 per household per year. For most U.S. households, the single most important action to reduce their carbon footprint is driving less or switching to a more efficient vehicle

**Purpose**

ML depends heavily on data, without data, it is impossible for an “AI” to learn. It is the most crucial aspect that makes algorithm training possible. In Machine Learning projects, we need a training data set. It is the actual data set used to train the model for performing various actions.

There are many features which are responsible for CO-2 Emission of Countries, e.g. Country Name, Country Code, Indicator Name etc. For better prediction of the CO2 Emission of Countries, we should consider as many relevant features as possible.

**2. Literature Survey**

**Existing Problems**

Research on CO2 emissions is highly renewable, qualitative and quantitative researches that have been done until now is still in the process of the discussion, it makes the theory and approaching method in order to calculate the emitter of CO2 emissions has not become one unit . Mitigation of Carbon Dioxide emission is the challenge of the future in order to stabilize global warming . CO2 prediction

using a computational intelligence approach . An adaptive neuro-fuzzy

interference system (ANFIS) and multi-layer perceptron artificial neural network (MLP-ANN) have been developed to estimate CO2. The proposed model of ANFIS and MLP-ANN demonstrates that both methods can solve CO2 prediction problem. There’s another method for prediction namely, support vector machine (SVM). Recently, several applications of SVM can be found both for classification and for regression problems . SVM is more accurate than semi empirical equations for predict solubility

of different solutes in supercritical carbon dioxide . On the other hand, SVM was implements the structural risk minimization principle. SVM has successfully solved a prediction of CO2 exchange rate . In their study, different tests were performed along the North Atlantic oceanic region with data obtained during 2009 and the proposed model of SVM demonstrates that SVM can solve CO2 prediction

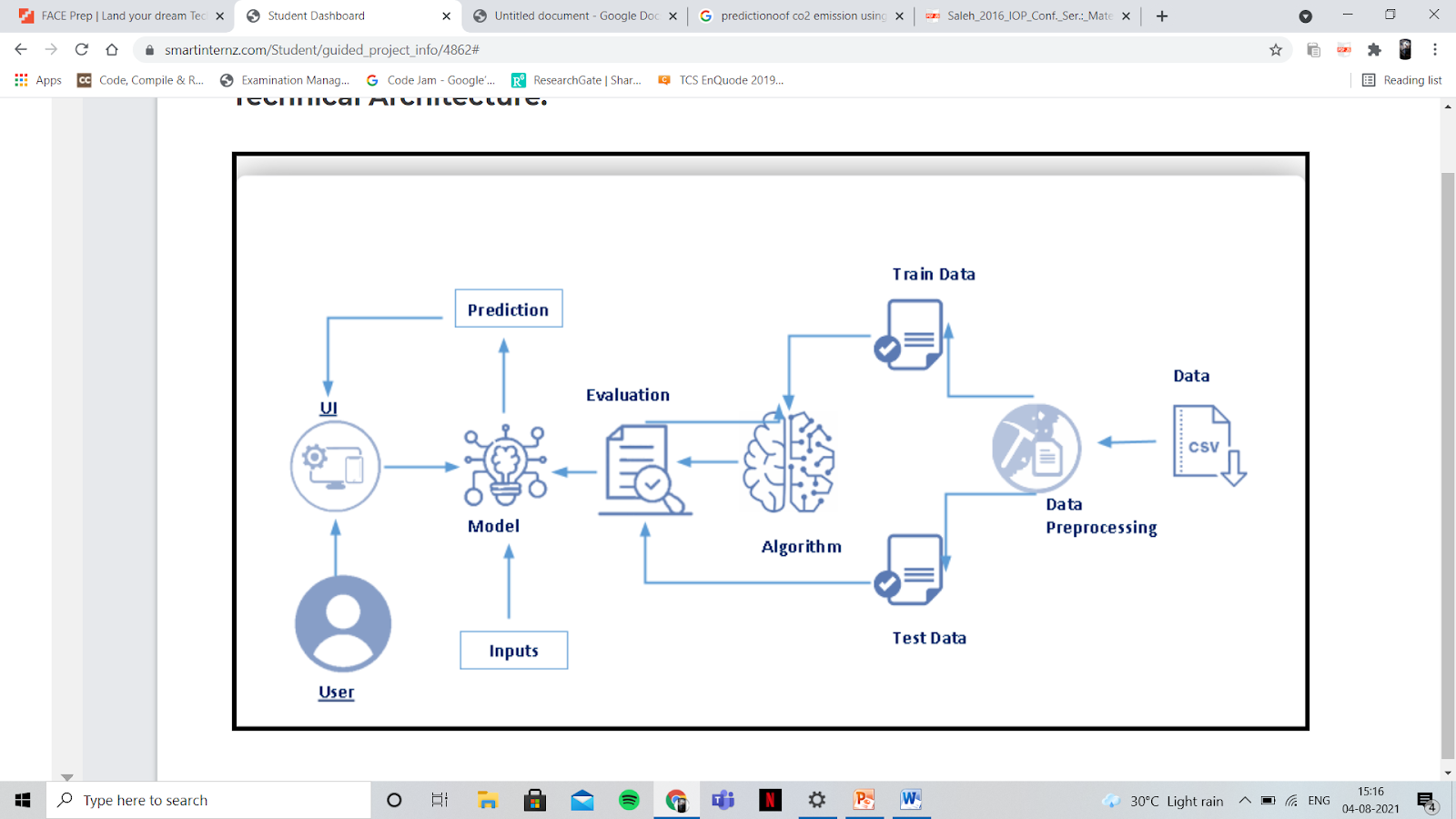
problem. SVM is powerful machine learning tool that can be used for time-series prediction .

**Proposed Solution**

A Machine Learning Model for calculating CO2 emission by country, Due to the increasingly deteriorating environment, it is time for the government to upgrade the energy consumption structure.

**3. Theoretical Analysis**

**Block Diagram**



**Hardware/Software design**

i) Install Anaconda Software

To install Anaconda on your local system, go through the below links according to your system requirements. After Anaconda is installed, run the .exe folder.

ii) Run Jupyter

Search Anaconda Navigator and open a Jupyter notebook.

iii) Numpy

Using Anaconda Navigator: conda install numpy

iv) Pandas

Using Anaconda Navigator: conda install pandas

v) Matplotlib

Using Anaconda Navigator: conda install matplotlib

OR

Using command prompt: pip install matplotlib

vi) Scikit-Learn

Using Anaconda Navigator: conda install -c conda-forge scikit-learn

OR

Using command prompt: pip install -U scikit-learn

**4. Experimental Investigations**

i) Dataset Collection

Collect the dataset or create the dataset.

* ML depends heavily on data, without data, it is impossible for an “AI” to learn. It is the most crucial aspect that makes algorithm training possible. In Machine Learning projects, we need a training data set. It is the actual data set used to train the model for performing various actions.

* There are many features which are responsible for CO-2 Emission of Countries, e.g. Country Name, Country Code, Indicator Name etc. For better prediction of the CO2 Emission of Countries, we should consider as many relevant features as possible.

* We can collect dataset from different open sources like kaggle.com, data.gov, UCI machine learning repository etc.

The kaggle repository link is : https://www.kaggle.com/ashukr/exploring-co2-emission?select=Indicators.csv

ii ) Data Preprocessing :

* Here, we are reading the dataset(.csv) from the system using pandas and storing it in a variable ‘df’. It's time to begin building your text classifier! The data has been loaded into a DataFrame called df. The .head() method is particularly informative.
* We might have your data in .csv files, .excel files or .tsv files or something else. But the goal is the same in all cases. If you want to analyse that data using pandas, the first step will be to read it into a data structure that’s compatible with pandas.

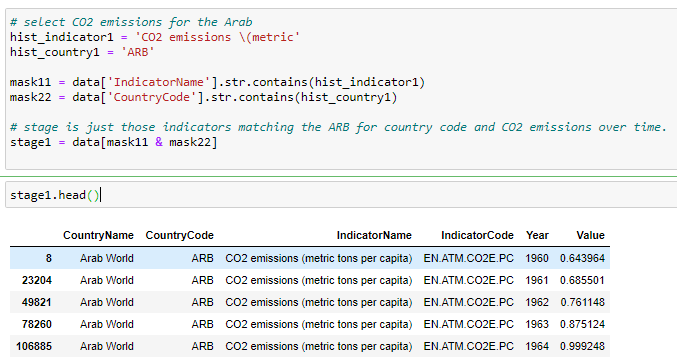
load a .csv data file into pandas. There is a function for it, called read\_csv().We will need to locate the directory of the CSV file at first (it’s more efficient to keep the dataset in the same directory as your program).

iii) Check Unique Values In Dataset

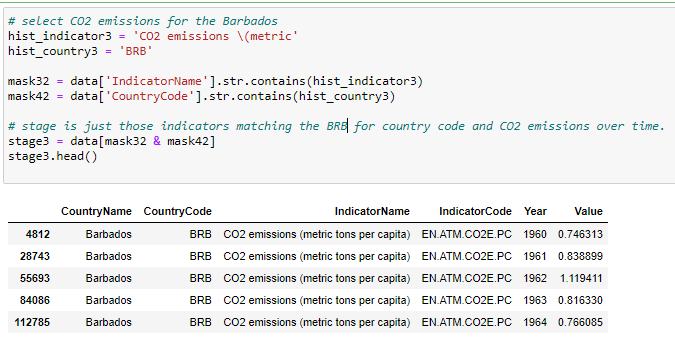
* Often, a DataFrame will contain columns that are having some unique values from which we can find out the unique records which are present in the dataset.
* find the min year and max year for count and also,find how many years of data we have.

iv) CO-2 Emission Of The Countries.

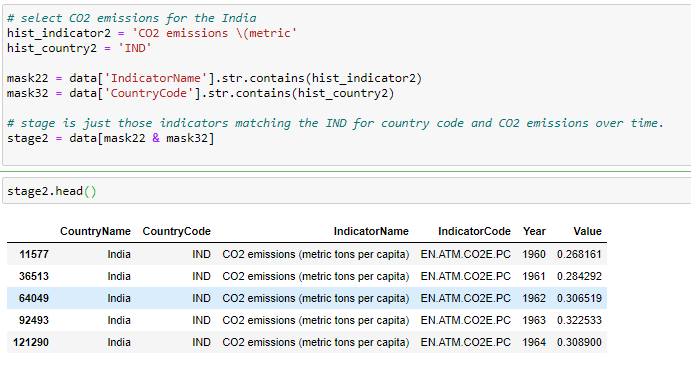
* As our dataset is very huge, we are dealing with a few countries like the USA,SGP,IND,BRB,ARB. Here we are selecting the indicator name which is Co2-Emission (metric tons per capita) and also take the country code to find the co2 emission over the time.
* Select CO2 emissions for the Arab Country and stage is just those indicators matching the ARB for country code and CO2 emissions over time.



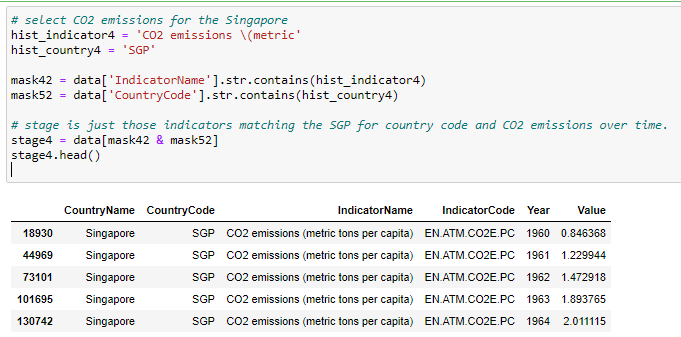
* Select CO2 emissions for the Barbados Country and stage is just those indicators matching the BRB for country code and CO2 emissions over time.



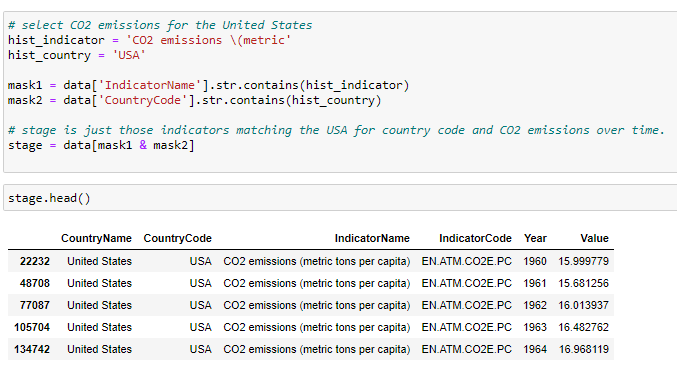
* Select CO2 emissions for the India Country and stage is just those indicators matching the IND for country code and CO2 emissions



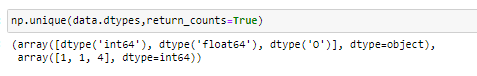
* Select CO2 emissions for the Singapore Country and stage is just those indicators matching the SGP for country code and CO2 emissions over time.



* Select CO2 emissions for the United States Country and stage is just those indicators matching the USA for country code and CO2 emissions over time.



v) Observing Target,Numerical And Categorical Columns



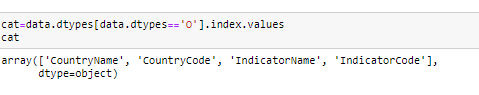
Here, data.dtypes will return us all the different types of data present in our data and return\_counts will give us the count.

Let’s start observing the categorical and numerical columns and check whether the above

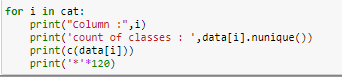
count is correct or not.

Categorical Columns:

* The below code is used for fetching all the object or categorical type of columns from our data and we are storing it as set in a variable cat.



* As you can observe, it gives us the same count of columns which we found previously.



* In the above we are looping with each categorical column and printing the classes of each categorical column using the counter function so that we can detect which columns are categorical and which are not.

* If you observe, some columns have a few classes and some have many, but in some columns there are similar types of classes present so let’s make that similar class into a single class using numpy .

Numerical Columns



Same as we did with categorical columns, we are making use of dtypes for finding the continuous columns.

vi) Data Visualization :

* Data visualization is where a given data set is presented in a graphical format. It helps the detection of patterns, trends and correlations that might go undetected in text-based data.Understanding your data and the relationship present within it is just as important as any algorithm used to train your machine learning model. In fact, even the most sophisticated machine learning models will perform poorly on data that wasn’t visualized and understood properly.
* To visualize the dataset we need libraries called Matplotlib and Seaborn.The Matplotlib library is a Python 2D plotting library which allows you to generate plots, scatter plots, histograms, bar charts etc.
* Let’s visualize our data using Matplotlib and the Seaborn library.

* Get the years and co2 emission value in the particular year.

**5. Flow chart**

### Train And Test The Model Using Random Forest Regressor

* There are several Machine learning algorithms to be used depending on the data you are going to process such as images, sound, text, and numerical values. The algorithms that you can choose according to the objective that you might have may be Classification algorithms are Regression algorithms.

Example: 1. Linear Regression.

2. Logistic Regression.

3. Random Forest Regression / Classification.

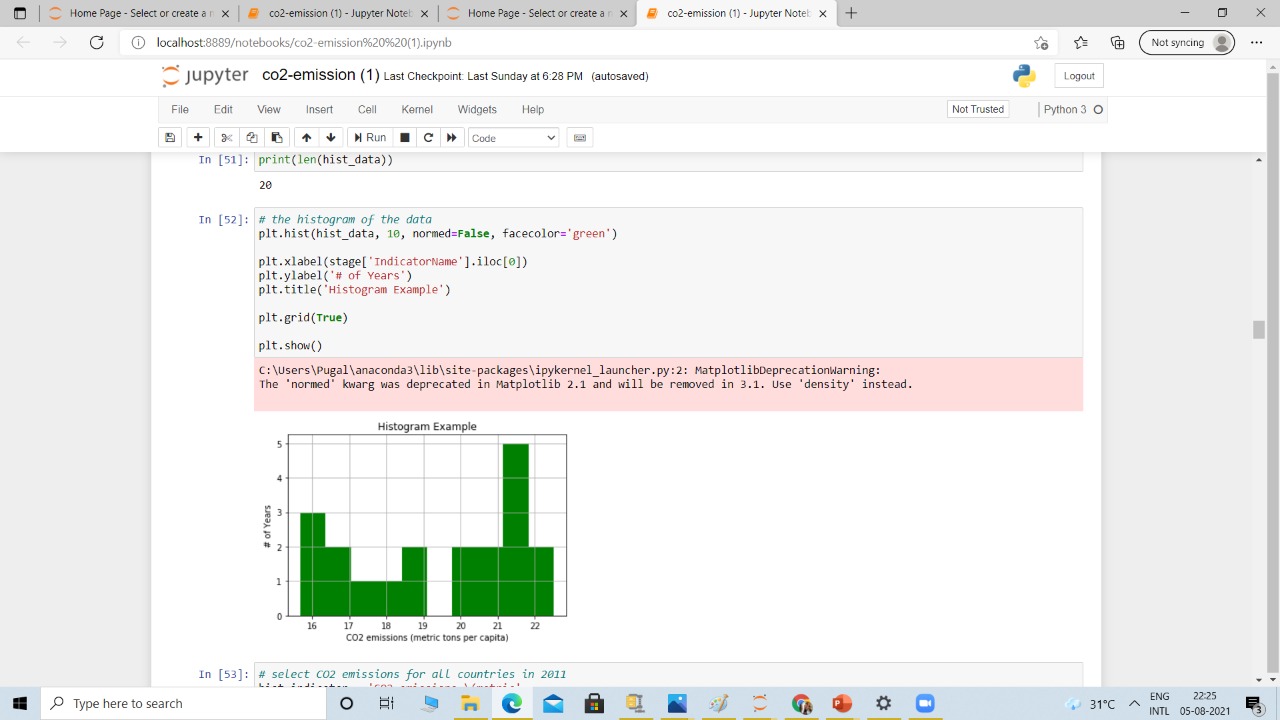
4. Decision Tree Regression / Classification.

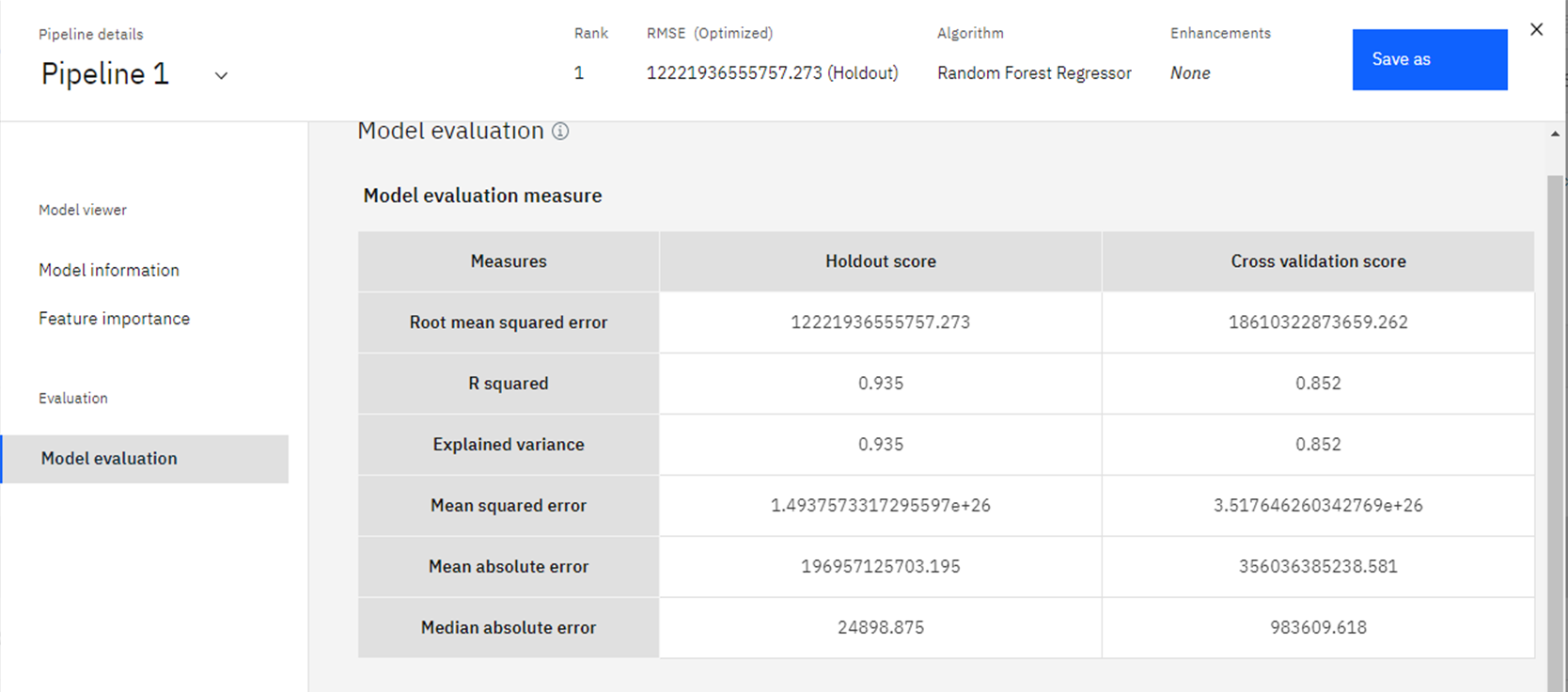
* You will need to train the datasets to run smoothly and see an incremental improvement in the prediction rate.
* Now we apply the Random forest regressor algorithm on our dataset.
* A Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap and Aggregation, commonly known as bagging. The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees.

**Build the model with the Random Forest Regressor**.

* We’re going to use x\_train and y\_train obtained above in train\_test\_split section to train our Random forest regression model. We’re using the fit method and passing the parameters as shown below.

**6. Result**





**7.Advantages and Disadvantages**

With the advantages such as fast convergence speed, high training accuracy

and no manual tuning, the ELM model has been successfully applied to forecasting problems in many fields, such as wind speed,electricity load, oil price

and so on.However,ELM is based on empirical risk minimization principle which easily causes over-fitting phenomenon.

Therefore, in order to guarantee the global optimization and generalization ability,RELM model, in which the calculation process of Moore-Penrose generalized inverse and the introduction of the regularization factor are added to ELM, is used for CO2 emission

prediction in this paper.

In general, based on the aforementioned studies, it can be found that the appropriate selection of influential factors has a momentous influence on the prediction results of CO2 emissions. However, most studies only put emphasis on the impact of these factors on the total CO2 emissions and ignore the correlation to each other. In reality, there exist overlaps of information contained in the data, thus, the computational efficiency is greatly depressed due to the complex network. Therefore,

principal component analysis (PCA) is employed in this paper to reduce the dimension of pre-select influential factors with retention of information to the utmost so that the network structure can be simplified and operation efficiency and prediction accuracy can be significantly improved.

Therefore, compared with past works, there exist two main differences:

(i) RELM, a new kind of neural network , is firstly introduced into CO2 emission prediction, which overcomes the disadvantages of slow learning speed, the need of numerous training samples, overfitting and so on in the previous research.

(ii) The correlations among influential factors are paid close attention in this paper, thus PCA is utilized to manipulate them for dimension reduction to improve the computational efficiency and forecasting precision.

**8.Applications**

1. Machine Learning models such as SVM, tuned SVM, linear model etc are used in Estimating CO2 emissions from Electricity generation

2. banking and Financial services industry uses Machine Learning to detect and reduce fraud, measure market riska dn identify oppurtunities.

3. Machine Learning play a key part in security as they typically use predictive analysis to improve services and performance, but also to detect anamolies, fraud, understand consumer behavior and enhance data security.

4. ML Algorithms are also known for it's recommendation algorithms like in Facebook or YouTube where similar content will be suggested to engage the user

**9.Conclusion**

The Support Vector Machine model can be applied to predict CO2 emissions from energy consumption which can give us more accuracy . This model is used to monitor electrical energy and burning coal which affect the amount of CO2 emitted. Trial and error approach was applied in order to obtain a better prediction model with a lower error. The results obtained show that the lower error (RMSE) value was 0.004 with optimal parameters for the SVM model of 0.1 for the C parameter and 0 for Epsilon. Prediction with high accuracy can give information concerning about CO2 emissions Furthermore, the main objective in this work is to achieve the lower RMSE when designing the model prediction. It can be concluded that with the high accuracy of the prediction model, then the lower RMSE value must be obtained .

**10.Future Scope**

With increasing investment in adaptation, one important challenge for the future will be to establish the right sets of indicators for identifying priorities, as well as monitoring and evaluation frameworks for adaptation. While progress on mitigation can be interpreted from trends in national GHG emissions,comparable measurable outcomes do not yet exist for adaptation. The difficulties in monitoring and evaluating adaptation range from the ambiguous definition of adaptation to the identification of targets and the choice of indicators used to monitor performance. Consequently, while international discussions on adaptation have focused on implementation of adaptation and the associated costs, systematic evaluation of

how much progress is being made in this direction is generally lacking and needs further development.

**11.Bibliography**

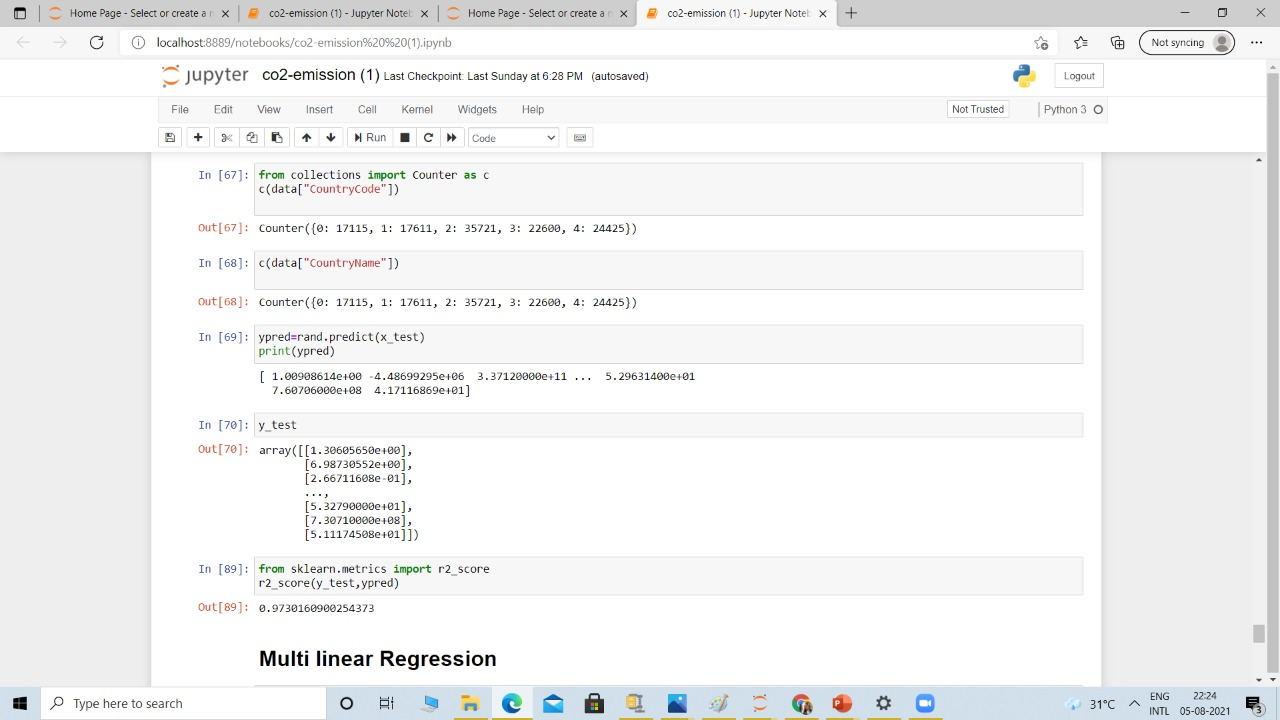
1.https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html#:~:text=A%20random%20forest%20regressor.,accuracy%20and%20control%20over%2Dfitting.

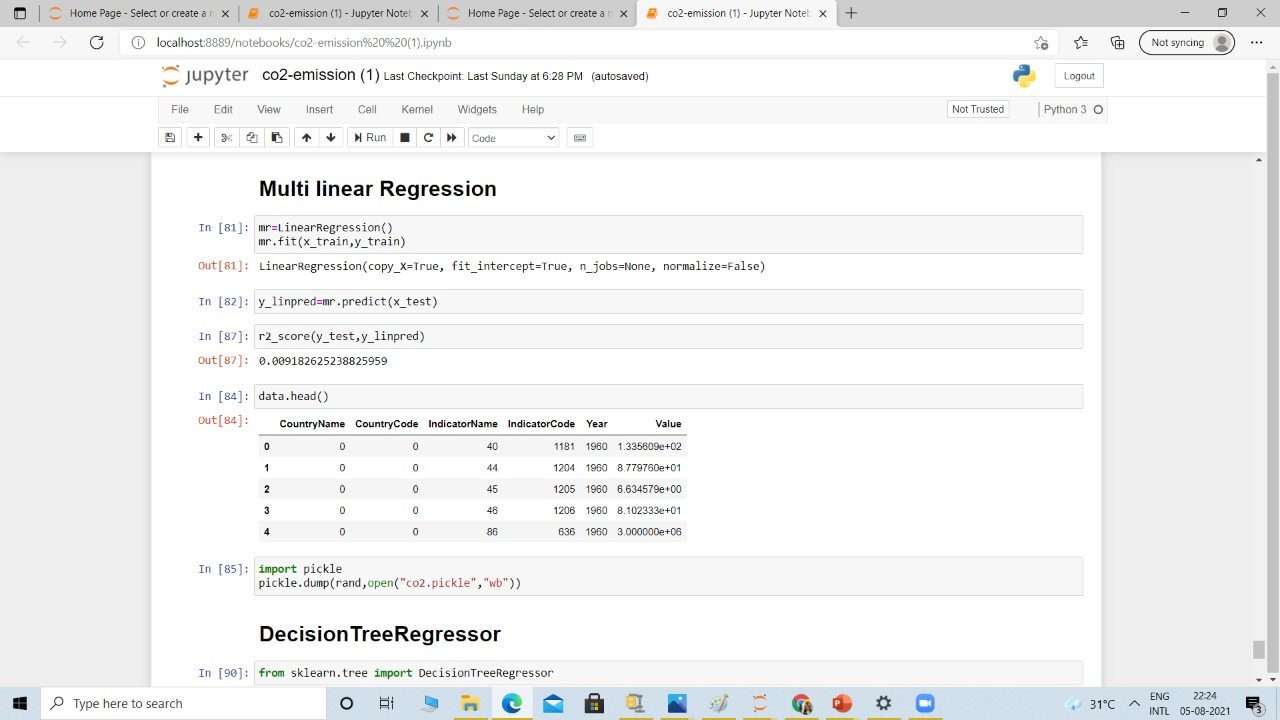
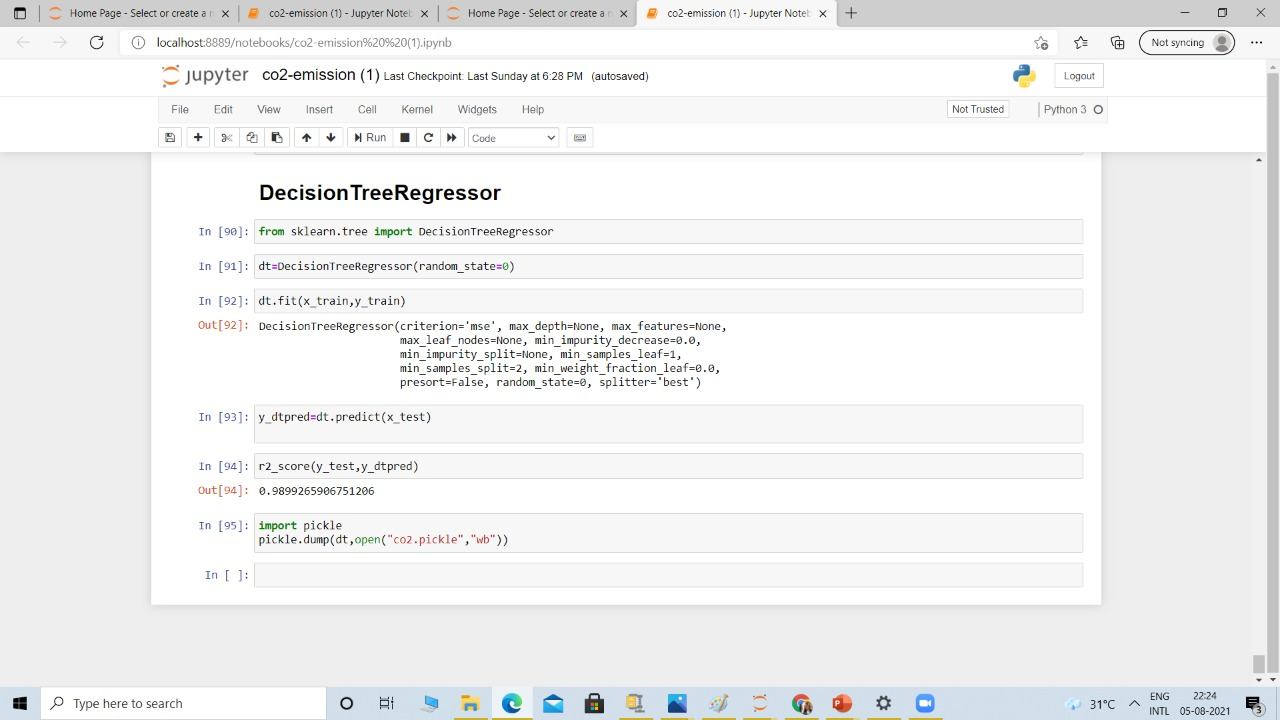
2.https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html#:~:text=A%20random%20forest%20regressor.,accuracy%20and%20control%20over%2Dfitting.

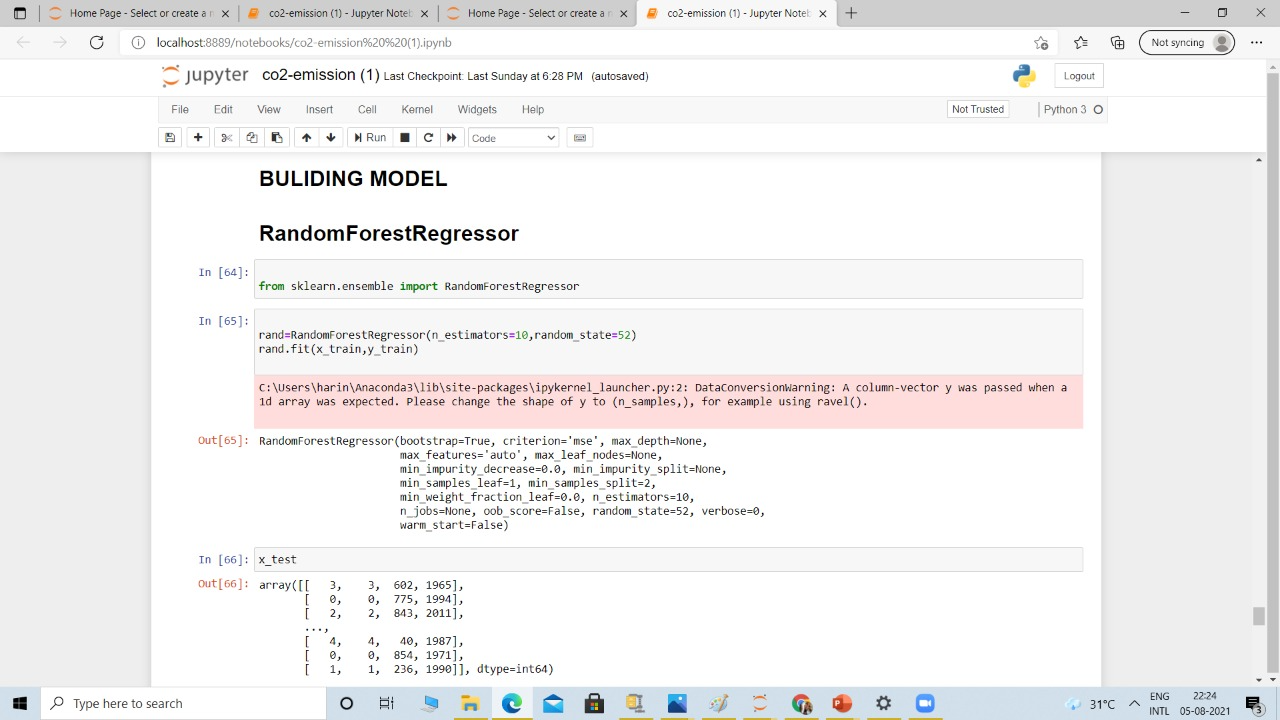
3.http://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html

**12.Appendix**

1. **Source code :**





**2. UI output screen shots :**

