# PREDICTION OF $CO_2$ EMISSION LEVELS IN INDIA USING MACHINE LEARNING TECHNIQUES

**Research Project**

*Submitted by*

**Aishitha Pachipala**

**2120425**

*Under the guidance of*

**Dr.Manu K S**

**Associate Professor**

**School of Business and Management**

**Christ (Deemed to be University)**

*In Partial Fulfilment of the Requirements for the Award of the Degree of*

**BACHELORS OF BUSINESS ADMINISTRATION**

**SCHOOL OF BUSINESS AND MANAGEMENT**

**CHRIST (Deemed to be University)**

**BENGULURU**

**2024**

# CERTIFICATE

This is to certify that the project submitted by Aishitha Pachipala (2120425) titled "Prediction of CO2 emission level in India using Machine Learning Techniques" submitted to CHRIST (Deemed to be University), in partial fulfilment of the requirements for the award of the Degree of Bachelor of Business Administration, is a record of original study undertaken by Aishitha Pachipala, during the period 2023 – 2024 in the School of Business and Management at CHRIST (Deemed to be University), Bangalore, under my supervision and guidance. The project has not formed the basis for award of any Degree / Diploma / Associate ship / Fellowship or other similar title of recognition to any other University.

Place: Bengaluru

Date:

Dr. Manu K S
Associate Professor,
School of Business and Management

Dr. Anuradha R
Head - School of Business and Management

# DECLARATION

I, Aishitha Pachipala, hereby declare that the project, titled "Prediction of CO2 emission level in India using Machine Learning Techniques" submitted to CHRIST (Deemed to be University), in partial fulfilment of the requirements for the award of the Degree of Bachelor of Business Administration is a record of original and independent study undertaken by me during 2023–2024 under the supervision and guidance of Dr. Manu K S, School of Business and Management. I also declare that this dissertation has not been submitted for the award of any degree, diploma, associate ship, fellowship or other title to any other Institution/University.

Place: Bengaluru

Date: Aishitha Pachipala (2120425)

# ACKNOWLEDGEMENT

I, Aishitha Pachipala would like to express my profound gratitude to all those who have been instrumental in the preparation of this research Project Report. I wish to place on records, our deep gratitude to my project guide, Dr. Manu K S, for guiding me through this project with valuable and timely advice. I would like to thank Dr. (Fr). Joseph C Cm Vice Chancellor, CHRIST (Deemed to be University), Dr. Jain Mathew, Dean and Dr. Anuradha, Head, School of Business and Management for their encouragement. Last but not the least; I would like to thank my parents and friends for their constant help and support.

Aishitha Pachipala

# TABLE OF CONTENTS

# CHAPTER 1

# INTRODUCTION

An economy grows when there are men and women employed in the country and earning money. Now cities and towns start to grow as people migrate for work and this rapid urbanization causes increased environmental pollution. Even though urbanization leads to a better living standard and job opportunities, it eventually leads to increase in $CO_2$ emissions. Cities around the world account for more than two thirds of global energy use, leading to 70% of energy-related carbon dioxide emissions (IRENA 2016). In an ideal situation, urbanization promotes productivity, opens doors for more economic gains, and creates more wealth and creativity to redesign science, arts, politics, and other human endeavours (Stewart and Lee 1986; Bloom et al. 2008; Glaeser 2011). However, urbanization leads to the spread of illness and inflicts other social problems of exclusion, crime, and poverty and ultimately leads to the degradation of environmental quality (Bloom et al. 2008).

The focus at large when it comes to global anthropogenic $CO_2$ emissions has always been on the developed world and emerging economies in Asia because they jointly contribute to about 80% of the global anthropogenic $CO_2$ emissions. For instance, the top ten emitting countries in the world  in the year 2012 which were all developed economies and emerging economies in Asia accounted  for about two-thirds of the global anthropogenic $CO_2$ emissions (International Energy Agency, 2014). The rate of urbanization in Africa and Asia regions is observed relatively fast, where the percentage of urban population is projected to be doubled between the year 2000 and 2030. Overall, the global urban population, which was 1.52 billion in 1970, is projected to reach 4.6 billion people by 2030, and much of its proportion will be in the Asian and African cities (World Urbanisation Prospects, 2011). Shahbaz M et.al, (2016) states that urban areas are expected to be energy intensive with high tendency of economic activities (i.e. industrial manufacturing and transportation) that are mainly fossil fuel driven and cause environmental degradation.

A country's growth is observed to be determined by the nexus of urbanization (the pace of urban population growth), energy consumption (the rate at which this growing populace relies on fossil fuel-driven energy), and trade openness (the speed of importing and exporting various goods and services). These components are interdependent and are leading to the destruction of more natural land at the expense of urban and industrial development. India being 7th largest country in the world with the highest population succeeding China, has been experiencing rapid economic growth, increased trade openness and additionally, has been urbanizing at fast pace since its liberalization. The rapid growth of cities in India over the decades has been driven by a large influx of people moving from rural to urban areas in search of better economic opportunities. This trend has been further accelerated by the rise of the country's IT sector. Cities like Mumbai, Delhi, Bangalore, and Kolkata have experienced development that has expanded areas and metropolitan regions. However, this urbanization has brought about challenges such as infrastructure, housing shortages, environmental degradation, and social inequalities within communities. Despite these issues, many individuals continue to flock to cities for work and social support.

The country’s population growth and economic expansion have caused a rise, in energy consumption with significant portion coming from fossil fuels like coal, oil and natural gas. The use of these fossil fuel based energy sources has sparked concerns regarding sustainability, air quality and greenhouse gas emissions as they release pollutants into the air. Despite efforts to diversify the energy sources by promoting options such as solar, wind and hydroelectric power; fossil fuels continue to dominate India’s energy sector due to their cost effectiveness and widespread availability. Openness of India i.e. imports and exports of goods and services have been increased significantly in the past few years. The country has resorted to globalization and liberalization of its economy. India's trade policy architecture follows a principle of grow-export orientation which seeks to drive economic expansion, generate employment, and improve the global competitive position. The country has been part of diverse trade agreements and partnerships involving the economies of developed as well as developing economies in order to widen market access and privileges on cross border trade. Although the impact of development looks positive for the country in broader terms, it is definitely leading to an increase in the level of $CO_2$ emissions in the country’s environment.

The problem with emitting CO2 into the atmosphere is that it is a greenhouse gas that captures the heat in the earth's atmosphere. When carbon emissions go up due to human activities like burning fossil fuels (coal, oil, and gas), deforestation, and industrial processes, the carbon dioxide level in the atmosphere also increases. And this increase results in disruptions of weather patterns and extreme weather conditions. It includes increased frequency and severity of heat waves, droughts, hurricanes, floods, and wildfires. Climate change affects ecosystems, agriculture, water supply, and human health, which creates widespread interruptions and vulnerabilities. On the other hand, a considerable amount of the emitted carbon dioxide is also absorbed by the world's oceans, contributing to ocean acidification. Elevated acidity is highly detrimental to many marine animals, especially those with calcium carbonate shells or skeletons, like coral reefs, shellfish, and certain plankton species. Ocean acidification can upset marine ecosystems and endanger the livelihoods of people who depend on the sea as their resource. Along with CO2, carbon emissions usually contain other pollutants like particulate matter, nitrogen oxides, and sulfur dioxide, which lead to air pollution. The health of people breathing air pollution deteriorates. Climate change and environmental defilement caused by carbon emissions also challenge biodiversity by changing habitats, disrupting ecosystems, and escalating animal livelihood problems. This results in many species being unable to adapt to the rapid fluctuations in temperature, precipitation, and habitat availability, which results in a decline in the population and the threat of extinction.

India has implemented several policies and frameworks to address the issue of carbon emissions and reduce the impact of climate change. These efforts include launching the National Action Plan on Climate Change (NAPCC) in 2008, which outlines a comprehensive strategy to tackle climate change and its impacts in India. The NAPCC comprises eight national missions that cover various areas such as solar energy, energy efficiency, sustainable agriculture, water conservation, and afforestation. Launched in 2010, the National Solar Mission aims to promote developing and deploying solar energy technologies across India. It includes targets for solar power generation capacity expansion, incentives for solar power projects, and initiatives to reduce the cost of solar energy production. The country has also established emissions standards and regulations for industries, vehicles, power plants, and other sources of pollution to limit air pollution and greenhouse gas emissions. India is also a signatory to international agreements and frameworks to combat climate change, including the Paris Agreement. Under the Paris Agreement, India has committed to reducing its greenhouse gas emissions intensity by 33-35% from 2005 levels by 2030 and achieving 40% of its cumulative electric power capacity from non-fossil fuel sources by 2030. India has launched many afforestation and reforestation programs to increase forest cover and sequester atmospheric carbon dioxide. Initiatives such as the Green India Mission and the National Afforestation Programme aim to expand forested areas, restore degraded lands, and enhance carbon sinks. The country has also established emissions standards and regulations for industries, vehicles, power plants, and other sources of pollution to limit air pollution and greenhouse gas emissions.

There are studies that examined the dynamic impact of urbanization, energy consumption and trade openness on $CO_2$ emissions in the existing literature but no single study attempted to investigate this relationship in India; therefore, our paper contributes to the existing literature by examining the causal linkage between urbanization, energy consumption, and trade openness on $CO_2$ emissions using ARDL model in India over the period of 1998 – 2022. Along with determining the linkage between the nexus, this paper also employs Decision Tree and Support Vector Machine models to forecast $CO_2$ emissions.

# CHAPTER 2

# REVIEW OF LITERATURE

Wang F et.al, (2020) studied the impact of urbanization of 166 Chinese cities on carbon emissions by employing a Generalized Method of Moments of dynamic panel data model to explore the nonlinear relationship between urbanization and $CO_2$ emissions along with impact of spatial agglomeration using the Gini coefficient to analyze whether the spatial agglomeration of cities contributed to environmental protection. The results revealed that there is an inverted U-shaped curve between urbanization and carbon emissions; and U-shaped relationship between spatial agglomeration and carbon emissions.

Li et.al, (2018) employed data envelopment analysis and a spatial lag panel model to investigate the effect of urbanization on $CO_2$ emissions at Yangtze River Delta, China. The study found a U curve relation between $CO_2$ emissions efficiency and urbanization.

Saidi and Mbarek (2016) studied the impact of financial development, income, trade openness, and urbanization on carbon dioxide emissions for the panel of emerging economies using the time series data over the period 1990–2013 by employing unit root test, co-integration test, and GMM-SYS of panel data. The results showed that GDP per capita had a significant positive impact, urbanization had a statistically significant negative impact, trade openness was not significant and financial development had a negative impact on $CO_2$ emissions.

Shahbaz M et.al, (2016) investigated the relationship between urbanization and $CO_2$ emissions in Malaysia by using the STIRPAT model. In this study after unit root test Bayer–Hanck combined cointegration approach was used to examine the cointegration relationship between the variables; ARDL bounds testing approach and VECM Granger causality test were applied. The findings expose that economic growth is a major contributor to $CO_2$ emissions, energy consumption raises emissions intensity; trade openness also increases $CO_2$ emissions and the relationship between urbanization and $CO_2$ emissions was found to be U shaped i.e. urbanization initially reduces $CO_2$ emissions, but after a threshold level, it increases $CO_2$ emissions. The causality analysis also suggested that the urbanization causes $CO_2$ emissions.

Poumanyvong and Kaneko (2010) analysed the urbanization effect on $CO_2$ emissions and energy consumption using a STIRPAT model and a balanced panel data analysis for a sample of 99 nations spanning from 1975 to 2005. The study verified that the impact of urbanization on $CO_2$ emissions and energy consumption depends on the levels of the economies development. It further subscribed that, urbanization reduces energy consumption in low-income class, and causes energy consumption for middle and high-income classes to rise. Besides, the study noted that urbanization in all the income groups positively impacted carbon emissions, but more was reflected in the middle and high-income classes.

Ali H S et.al (2016) examined impact of urbanization on carbon dioxide emissions in Singapore from 1970 to 2015 using autoregressive distributed lags approach. The main finding reveals a negative and significant impact of urbanization on carbon emissions in Singapore. The result also highlighted that economic growth had a positive and significant impact on carbon emissions; and variable of trade openness remained insignificant on carbon emissions in the country.

Shehu (2019) studied the urbanization and $CO_2$ emissions nexus in Nigeria using the ARDL method to analyze the annual time series data spanning from 1974 to 2015. Findings suggested that urbanization, GDP, energy use, and carbon emissions are strongly and positively correlated, while trade and carbon emissions exhibit a weak and negative correlation. The study concluded from the findings that urbanization is not a significant factor in contributing to an increase in carbon emissions, but rather energy use.

Pata (2017) examines the relationship between urbanization, industrialization, and carbon emission in Turkey between 1974 and 2013 using the ARDL model. The study concludes that in Turkey, urbanization and industrialization decrease the level of environmental quality captured by an increase in carbon emissions per head.

Poku (2016) examined the relationship between urbanization, population and $CO_2$ emissions in 45 SSA countries by using panel data from 1990 – 2010 and establishing pooled mean group estimator for dynamic heterogeneous panels. The major findings is that population and urbanization are two of the major driving forces behind increasing $CO_2$ emissions in SSA over the past two decades.

Shahbaz M et.al, (2014) used the ARDL approach and investigated the nexus among economic growth, electricity consumption, urbanization, and environmental condition during 1975–2011 in the UAE. The findings reveal the existence of an inverted U-shaped relationship between economic growth and $CO_2$ emissions. Electricity consumption reduces carbon emissions and urbanization enhances it, while export improves environmental quality as it reduces carbon emissions.

Ali R et.al, (2019) analysed the impact of urbanization on carbon dioxide emissions in Pakistan using time series data from 1972 to 2014. Their analysis, employing ARDL bound testing and VECM models, revealed a positive co-integrating relationship between urbanization and emissions, indicating a long-term influence. Notably, a 1% increase in urbanization was found to lead to a 0.84% increase in carbon emissions in the long run.

Martinez-Zarzoso and Maruotti (2011) investigate the relationship between urbanization and $CO_2$ emissions in developing countries. The result reveals an inverted U shaped relationship between urbanization and carbon emissions. Grouping the countries based on threshold analysis indicates that at a given point, emission urbanization became negative and when urbanization reaches beyond such a point, emissions remain stagnant. The results of other groups show that urbanization does not promote carbon emissions but rather wealth and population.

Ouyang and Lin (2016) conducted a comparative study between China and Japan at the urbanization stages to analyze the similarities as well as differences of influencing factors of $CO_2$ emissions. A cointegration model is constructed to examine the long-run equilibrium relationship between $CO_2$ emissions and factors including GDP, urbanization level, energy intensity and cement manufacture. Results indicated that although $CO_2$ emissions in Japan and China showed similar characteristics of rigid growth during the urbanization processes, significant differences existed in factors such as $CO_2$ emissions per capita, energy structure and energy intensity between the two countries, which are the determinants for $CO_2$ emissions growth.

However, some studies found no significant impact of urbanization on $CO_2$ emissions for instance; a finding by Ali H S et.al (2016) in Nigeria based on ARDL approach for the period of 1971–2011 stated that urbanization did not have any significant impact on $CO_2$ emissions. The same finding was also documented by the study of Hossain (2012) in the case of Japan, who examined the causal relationship among $CO_2$ emissions, energy consumption, economic growth, foreign trade, and urbanization for the period of 1960−2009;  in which economic growth, trade openness, and urbanization does not affect $CO_2$ emissions in the long-run.

# CHAPTER 3

# RESEARCH DESIGN AND METHODOLOGY

### 3.1 Problem Statement

The rapid development of India in recent years has led to a number of challenges, starting off with traffic congestion, reduction of green spaces to accommodate burgeoning migration and traffic, increase in immense infrastructure projects and so on which ultimately paves path to increasing carbon emissions in the environment. India is ranked 8th as per Climate Change Performance Index (CCPI, 2023) in the emission of $CO_2$. This is not a pleasant upward movement of rank.

Although there are quite few studies conducted in this area of interest, we have don't have any literature showing the India's relationship and impact of the nexus on carbon emissions. Hence, I chose to address the lack of research in India. Most of the existing research has been conducted in developed and developing countries, and it is not clear whether the same findings would apply to a country like India.

### 3.2 Objective of Study

The research sets out to mainly to address the following:

- To understand various factors that is driving carbon emissions in India.
- To analyze the impact of fossil fuel energy consumption, trade openness, and urban population on $CO_2$ emissions.
- To predict $CO_2$ emission levels using supervised machine learning techniques.
- To compare prediction accuracy of the  machine learning models.
- To select the best model based on the highest prediction accuracy.

### 3.3 Scope of Study

This study encompasses a comprehensive analysis of the factors driving carbon emissions in India, aiming to contribute to the existing body of knowledge on this subject along with using machine learning models to predict the emissions. The study will focus on investigating within the Indian context, considering the unique socio-economic, environmental, and developmental dynamics of the country.

### 3.4 Hypotheses

***Null Hypothesis*** (H0): There is no significant impact of all 3 independent variables namely fossil fuel energy consumption, trade openness, and urban population on dependent variable that is $CO_2$ emissions.

***Alternative Hypothesis*** (H1): There is significant impact of all 3 independent variables namely fossil fuel energy consumption, trade openness, and urban population on dependent variable that is $CO_2$ emissions.

### 3.5 Data

$CO_2$ emissions = $CO_2$ emissions in kilo terms ($CO_2$kt)

Urbanization = percentage of total population (UB)

Energy Consumption = fossil fuel consumption (EC)

Trade Openness = total export and import as a percentage of GDP (TO)

### 3.6 Source of Data

The data used for this study was obtained from World Development Indicators (WDI, 2024)

### 3.7 Period of Study

The study collected annual data of $CO_2$ emissions, Urbanization, Energy Consumption and Trade Openness for the period of 25 years from 1998-2022.

### 3.8 Analytical Tools

#### 3.8.1 Econometric Model

The proposed research will follow a methodology outlined by Ali H. S. et.al, (2016) in their studies conducted in Singapore and Nigeria, respectively, in examining the Stochastic Impacts by Regression on Population, Affluence, and Technology (STIRPAT) model. Dietz and Rosa (1994, 1997) originally introduced the idea of formulating a stochastic version of the IPAT equation, which integrates quantitative variables such as population size (P), per-capita affluence (A), and the weight of the industry in economic dealings as a measure of environmentally damaging technology (T). STIRPAT is mainly applied to study the factors that affect environment (for example, check Dietz and Rosa 1994, 1997; York et al. 2003; Cole and Neumayer 2004). The STIRPAT model provides a theoretical framework for understanding the relationships between human activities and environmental impacts, particularly in terms of resource consumption and pollution generation. Hence, we choose to apply the ARDL model to empirically test the relationships postulated by the STIRPAT model. We initiate the project by employing unit root test to examine the stationarity of the variables because even if one of the series is integrated at second difference i.e. I(2), ARDL approach cannot be applied. Stationarity test is a statistical method for determining whether a time series variable is stationary or not. A stationary variable is one whose statistical features (mean, variance, and autocorrelation) stays constant across time, indicating that it does not show trends, seasonality, or other non-stationary patterns. The Augmented Dickey-Fuller (ADF) Test which is a traditional unit root test developed by Dickey and Fuller (1979) is used in this project to assess stationarity of variables.

Following Fosu O. A. E. & Magnus F. J. (2006), we apply the ARDL Bounds Testing approach to cointegration to empirically analyse the long-run cointegration relationships among the variables; which was developed by Pesaran et al. (2001). The ARDL approach to co-integration does not require all variables to be in identical order i.e. this technique can be used with a mixture of I(0) and I(1) order of integration in the series. However, this approach cannot be used when one of the variables is I(2) (Ali R, 2019). Then the null hypothesis of no cointegration ($H0 = \alpha1 = \alpha2 = \alpha3 = \alpha4 = \alpha5 = 0$) is tested against the alternate hypothesis of cointegration ($Ha \neq \alpha1 \neq \alpha2 \neq \alpha3 \neq \alpha4 \neq \alpha5 \neq 0$). The decision rule as recommended by Pesaran et al. (2001) is that, if the calculated F statistics are higher than the upper critical bound I(0), the null hypothesis will be rejected which means there is a cointegration among the variables signifying long run relationship, while if the calculated F-statistics are less than the critical bound I(0), the null hypothesis cannot be rejected, which means there is no cointegration among variables signifying no long run relationship.

Once the long run relationship is established we deploy ARDL model to check the dependency of carbon emissions on the elements. As conclusion of the model some diagnostic tests are then applied for examining appropriateness of the model. In this regards, heteroscedasticity tests are applied.

### 3.8.2 Machine Learning Models

The second part of this paper is to deploy machine learning models to predict the $CO_2$ emission in the country. The models chosen for this project are Decision Tree Classifier and Support Vector Machine.

#### 3.8.2.1 Decision Tree Classifier

It is a type of supervised machine learning algorithm used for classification tasks. According to Song YY, Lu Y (2015) decision tree methodology is a commonly used data mining method for establishing classification systems based on multiple covariates or for developing prediction algorithms for a target variable. The main components of a decision tree model are internal nodes which denotes an attribute, branches representing an outcome of the test and leaf nodes (terminal node) holding class label. A decision tree supports non-linearity, making it adept at capturing complex relationships within data. In general, decision trees tend to exhibit better average accuracy across various scenarios. Additionally, decision trees are known to handle colinearity more effectively compared to linear regression models. We have deployed this machine learning model to predict the $CO_2$ emission values and used the natural logarithmic $CO_2$ values which show an upward trend keeping in mind that this transformation does not necessarily change the underlying relationship between the variables.

### 3.8.2.2 Support Vector Machine

Support Vector Machines (SVMs) emerged within the framework of statistical learning theory by Vapnik (1998) and Cortes and Vapnik (1995). Like the Decision Tree Classifier, SVMs are supervised machine learning model that can handle both linearly separable and non-linearly separable datasets through the use of kernel functions. SVMs are adaptable and efficient in a variety of applications that been successfully applied across diverse domains such as image classification, text categorization, and bioinformatics etc. For this project we will be using Linear SVM since the relationship between the inputs features (years) and the target variable ($Ln\_CO_2$) appears to be linear.

### 3.8.2.3 Decision Tree vs Support Vector Machine

| Basis | Decision Tree | Support Vector Machine |
|---|---|---|
| Model Structure | Decision Trees break down feature space into regions based on feature values, forming a tree-like structure where each internal node represents a decision based on a feature, and each leaf node represents a class label or a regression value. | It constructs a hyperplane or set of hyperplanes in a high-dimensional space, which can be used for classification, regression, or outlier detection tasks. They mainly aim to find the hyperplane that maximizes the margin between classes or fits the data points most effectively. |
| Decision Boundaries | Decision boundaries in decision trees are orthogonal to feature axes, resulting in axis-aligned splits in the feature space. They | SVMs aim to find the hyperplane that best separates different classes in the feature space. Depending on the kernel function used, SVMs can |

| | | |
|---|---|---|
| | can create complex decision boundaries that are non-linear and irregular. | create linear or non-linear decision boundaries, providing flexibility in handling complex relationships between features. |
| Handling Non - Linearity | Decision trees inherently support non-linearity in the data and can capture complex relationships between features and the target variable without the need for explicit feature transformation. | SVMs can handle non-linearity by using kernel functions such as polynomial, radial basis function (RBF), or sigmoid kernels, which implicitly map the input features into a higher-dimensional space where the data may be linearly separable. |
| Interpretability | Decision trees are highly interpretable and intuitive, as the decision-making process can be easily understood by following the tree structure and examining the splits. | SVMs are less interpretable compared to decision trees, especially when using non-linear kernel functions, as the decision boundary may exist in a higher-dimensional space that is not directly interpretable. |

**Table - 1**

### 3.8.2.4 Prediction Accuracy Evaluation

- Mean Squared Error (MSE) represents the average of the squared difference between the original and predicted values in the data set. It measures the variance of the residuals.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2$$

- Root Mean Squared Error (RMSE) is the square root of Mean Squared error. It measures the standard deviation of residuals.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2}$$

- Mean Absolute Error represents the average of the absolute difference between the actual and predicted values in the dataset. It measures the average of the residuals in the dataset.

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}|$$

### 3.9 Software Used

***Eviews*** is an econometric, statistics, and forecasting package used mainly for time - series oriented econometric analysis. We in this project have used this software to conduct unit root test.

***Python*** is a programming language designed to assist programmers in writing simple, logical code for both small and large projects. In this project the language was used to code and execute the ARDL model, cointegration test, long run & short run coefficient estimation and diagnostic tests of the econometric model and; Decision Tree Classifier and Support Vector Machine of machine learning models.

# CHAPTER 4

# ANALYSIS AND INTERPRETATION

### 4.1 Econometric Model

| | Ln_CO2 | Ln Ec | Ln Trade | Ln UP |
|---|---|---|---|---|
| **Mean** | 14.2483 | 4.2168 | 3.6924 | 3.4344 |
| **Median** | 14.3223 | 4.2342 | 3.7377 | 3.4317 |
| **Maximum** | 14.7149 | 4.3382 | 4.0217 | 3.5800 |
| **Minimum** | 13.6367 | 3.7759 | 3.1655 | 3.3047 |
| **Std. Dev.** | 0.3692 | 0.1158 | 0.2599 | 0.0852 |
| **Skewness** | -0.2714 | -2.1981 | -0.7261 | 0.0983 |
| **Kurtosis** | 1.5182 | 9.3998 | 2.3763 | 1.8076 |
| **Jarque-Bera** | 2.5942 | 62.7960 | 2.6020 | 1.5213 |
| **Probability** | 0.2733 | 0.0000 | 0.2723 | 0.4674 |
| **Observations** | 25 | 25 | 25 | 25 |

**descriptive statistics of variables (Table – 2)**

The above figures shows basic descriptive statistics of variables taken in this project, where Ln_CO2 which is natural logarithmic value of carbon emission is dependent variable while the rest Ln Ec which is natural logarithmic value of fossil fuel energy consumption, Ln Trade which is natural logarithmic value of trade openness and Ln UP which is natural logarithmic value of urban population are independent variables.

| | Ln_CO2 | Ln Ec | Ln Trade | Ln UP |
|---|---|---|---|---|
| **Ln_CO2** | 1 | 0.334486 | 0.696157 | 0.962399 |
| **Ln Ec** | 0.334486 | 1 | 0.423203 | 0.192346 |
| **Ln Trade** | 0.696157 | 0.423203 | 1 | 0.617947 |
| **Ln UP** | 0.962399 | 0.192346 | 0.617947 | 1 |

**cross correlation of variables (Table – 3)**

Ln_CO2 being the dependent variable has a moderate positive correlation with Ln Trade and has a strong positive correlation with Ln UP, while it has weak positive correlation with Ln Ec. These positive correlations suggest that as Ln_CO2 increases, Ln Trade and Ln UP tend to increase as well.

**Graphs of Variables Over Years**

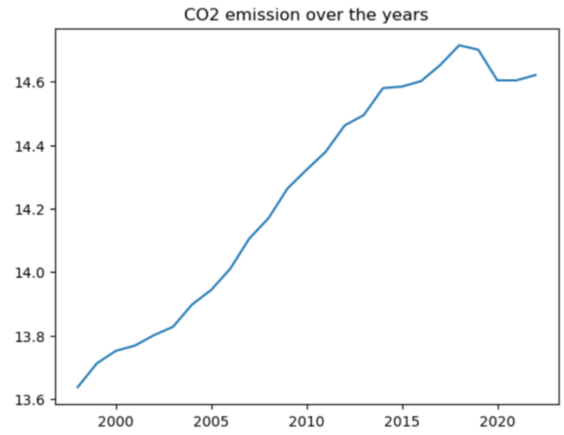


**Figure – 1**

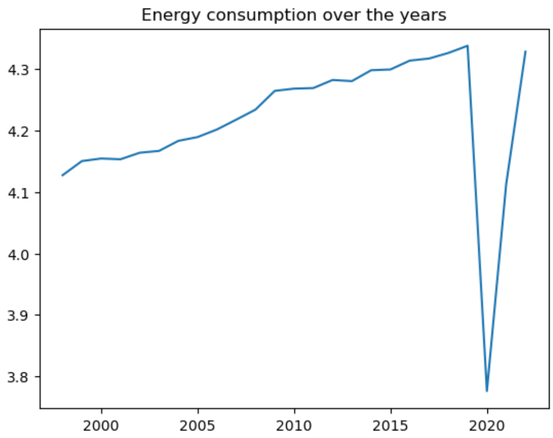


**Figure – 2**

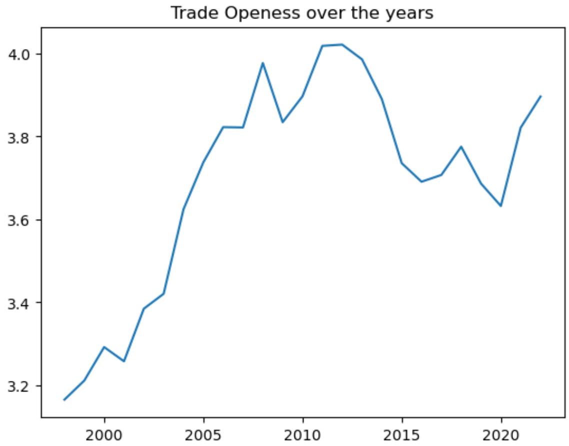


**Figure – 3**

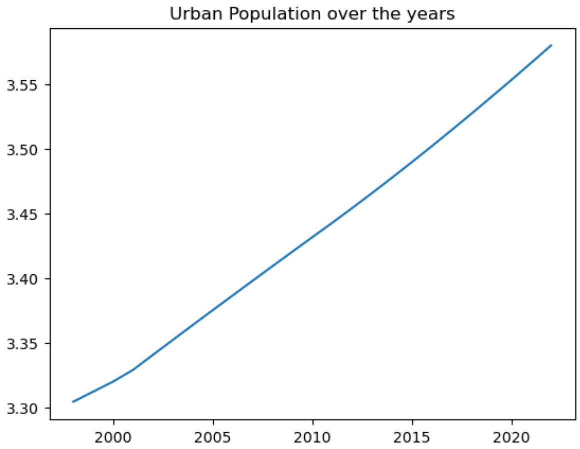


**Figure – 4**

|  | Level |  | I(1) |  |
|---|---|---|---|---|
| **Variable** | **t - statistic** | **Prob.** | **t - statistic** | **Prob.** |
| Ln_CO2 | -2.0453 | 0.2669 | -2.9077 | 0.0598 |
| Ln Ec | -3.6696 | 0.0117 | -5.4651 | 0.0002 |
| Ln Trade | -2.0181 | 0.2776 | -4.0458 | 0.0052 |
| Ln UP | 4.06608 | 1 | -3.7009 | 0.0116 |

**unit root test of variables (Table – 3)**

All the variables lie in first difference order i.e., I(1).

| **ARDL Bound Testing Results** | |
|---|---|
| F-statistic | 1.333333333 |
| Upper critical bound at 0.05 significance level | 2.796375489 |
| Variables are not likely cointegrated (at the specified significance level). | |

**ARDL Bound Test (Table – 4)**

We have used the ARDL Bound Test to help us understand cointegration amongst the variables its absence suggests that they may not move together in the long run. As per the result F-statistic is lower than the upper critical bound at the 0.05 significance level. Therefore, the variables are not likely cointegrated at the specified significance level.

| **ARDL Model Results** | | | | | | |
|---|---|---|---|---|---|---|
| **Dep. Variable :** | | Ln_CO2 | | **No. Observations :** | | 25 |
| **Model :** | | ARDL(2, 2, 2, 2) | | **Log Likelihood :** | | 64.716 |
| **Method :** | | Conditional MLE | | **S.D. of innovations :** | | 0.015 |
| | **coef** | **std err** | **z** | **P>\|z\|** | **[0.025** | **0.975]** |
| **const** | -1.9455 | 1.107 | -1.758 | 0.106 | -4.381 | 0.49 |
| **Ln_CO2.L1** | 0.0425 | 0.324 | 0.131 | 0.898 | -0.671 | 0.756 |
| **Ln_CO2.L2** | 0.4274 | 0.244 | 1.749 | 0.108 | -0.11 | 0.965 |
| **Ln Ec.L0** | 0.3073 | 0.062 | 4.974 | 0 | 0.171 | 0.443 |
| **Ln Ec.L1** | 0.2156 | 0.101 | 2.142 | 0.055 | -0.006 | 0.437 |
| **Ln Ec.L2** | 0.1971 | 0.098 | 2.009 | 0.07 | -0.019 | 0.413 |
| **Ln Trade.L0** | 0.0701 | 0.065 | 1.073 | 0.306 | -0.074 | 0.214 |
| **Ln Trade.L1** | 0.0372 | 0.077 | 0.486 | 0.636 | -0.131 | 0.206 |

| | | | | | | |
|---|---|---|---|---|---|---|
| **Ln Trade.L2** | 0.0567 | 0.073 | 0.777 | 0.453 | -0.104 | 0.217 |
| **Ln UP.L0** | -11.8994 | 10.772 | -1.105 | 0.293 | -35.609 | 11.81 |
| **Ln UP.L1** | 9.5326 | 19.659 | 0.485 | 0.637 | -33.737 | 52.802 |
| **Ln UP.L2** | 4.1434 | 11.021 | 0.376 | 0.714 | -20.113 | 28.399 |

**ARDL regression model (Table – 5)**

The coefficient values represent the impact of each independent variable on the dependent variable (Ln_CO2). The standard error (std err) measures the precision of the coefficient estimates. The z-score and p-value (P>|z|) indicate the statistical significance of each coefficient. The confidence interval [0.025, 0.975] provides a range in which the true population parameter is likely to fall. We look for smaller p-values which suggest coefficient to be statistically significant. From the result above only fossil fuel energy consumption seems to show significant impact on $CO_2$ emissions.

| **Tests** | |
|---|---|
| **Auto Serial correlation :** | 0.570702911 |
| **Heteroscedasticity:** | |
| **LM Statistic** | 6.436008917 |
| **LM-Test p-value** | 0.092220605 |
| **F-Statistic** | 2.42685219 |
| **F-Test p-value** | 0.093981087 |

**ARDL diagnostic tests (Table – 6)**

Auto serial correlation, also known as autocorrelation, measures the correlation of a variable with itself at different lags. A value closer to 1 indicates strong autocorrelation, while a value closer to 0 suggests weaker autocorrelation. And our data lies in the middle showing moderate autocorrelation. Now, heteroscedasticity refers to the situation where the variance of errors or residuals in a regression model is not constant across observations. A smaller p-value indicates stronger evidence against the null hypothesis of no heteroscedasticity. In both cases, while the p-values are not extremely low, they are close to conventional significance levels (e.g., 0.05). Since the p-value is greater than 0.05, we suggest that there isn't enough evidence to conclude that there is heteroscedasticity in data.

### 4.2 Machine Learning Models

In this part of the project we take ‘Ln_$CO_2$’values which are the natural logarithmic form of the $CO_2$ emission. First, the descriptive statistics of the variable is calculated;

| | |
|---|---|
| **count** | 25 |
| **mean** | 14.24831 |
| **std** | 0.36922 |
| **min** | 13.63671 |
| **25%** | 13.89771 |
| **50%** | 14.32232 |
| **75%** | 14.60181 |
| **max** | 14.71493 |

**descriptive statistics of $CO_2$ emissions (Table – 7)**

From the table above, we can infer that there are 25 observations in the dataset of variable ‘Ln_$CO_2$'. On an average, the natural logarithm of $CO_2$ emissions across the dataset is around 14.25 and its average deviation from the mean is 0.37 from the mean. The range of values covered by the dataset is minimum of 13.64 and maximum of 14.71. The spread of the data in 1st quartile is around 13.90, the median is around 14.32, and finally the 3rd quartile is approximately 14.60.

### 4.2.1 Decision Tree Classifier

We predicted the values on test set, after which we have evaluated the metrics used to check the performance of the model.

| | |
|---|---|
| **Mean Squared Error (MSE)** | 0.005229915 |
| **Root Mean Squared Error (RMSE)** | 0.072318148 |
| **Mean Absolute Error (MAE)** | 0.064177366 |
| **Accuracy (%)** | 96.05610218 |

**performance & evaluation of model (Table – 8)**

By calculating various errors, it tells how well the data fits the model. Lower the value of error, the better the model. In this case the Mean Absolute Error is 0.0052, Mean Square Error is 0.0723 and the Root Mean Square Error is 0.0641. The high accuracy percentage (96.06%) is indicating that the decision tree classifier is effective in correctly classifying instances. So this can be deemed as a better model keeping in mind various other aspects of model performance and evaluation.

| Year | Actual_CO2 | Predicted_CO2 |
|---|---|---|
| 2006 | 14.010424 | 13.94 |
| 2014 | 14.579632 | 14.49 |
| 1998 | 13.636707 | 13.71 |
| 2021 | 14.604292 | 14.6 |
| 2009 | 14.26332 | 14.17 |

**predicted values using decision tree (Table – 9)**

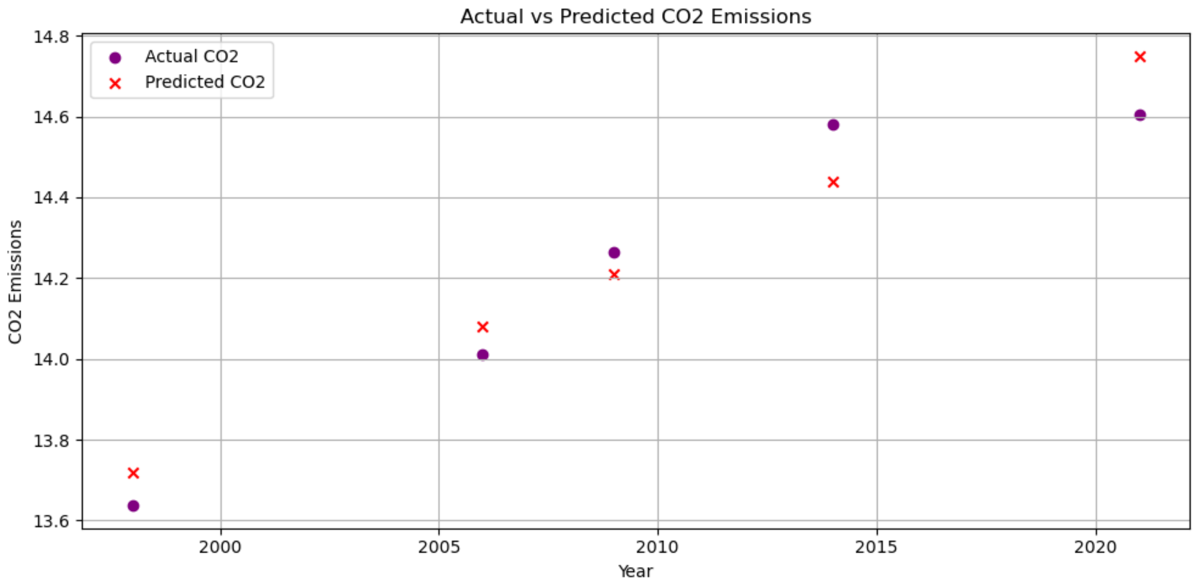


**Figure - 5**

The figures above are the outputs arrived from deploying the 'Decision Tree Classifier' model; where the former (figure 13) showcases the actual $CO_2$ emission values and their predicted values for 5 different years while the latter (figure 14) shows the graph plot of same data.

### 4.2.2 Support Vector Machine

We similarly predicted the values on test set, after which we have followed the same procedure as the first model.

| | |
|---|---|
| **Mean Squared Error (MSE)** | 0.01120076 |
| **Root Mean Squared Error (RMSE)** | 0.10583363 |
| **Mean Absolute Error (MAE)** | 0.09847753 |
| **r2 score** | 0.91553468 |

**performance and evaluation of model (Table – 10)**

By calculating errors, the Mean Absolute Error is 0.0984, Mean Square Error is 0.0112 and the Root Mean Square Error is 0.1058. Since the values of the above 3 errors are low the

SVM model's predictions are close to the actual values, indicating good performance. And higher r2 scores indicate better fit of the model to the data, 91.55% significant proportion of the variance in the dependent variable is explained by the independent variable.

| Year | Actual_CO2 | Predicted_CO2 |
|---|---|---|
| 2006 | 14.010424 | 14.08 |
| 2014 | 14.579632 | 14.44 |
| 1998 | 13.636707 | 13.72 |
| 2021 | 14.604292 | 14.75 |
| 2009 | 14.26332 | 14.21 |

**predicted values using SVM (Table – 10)**

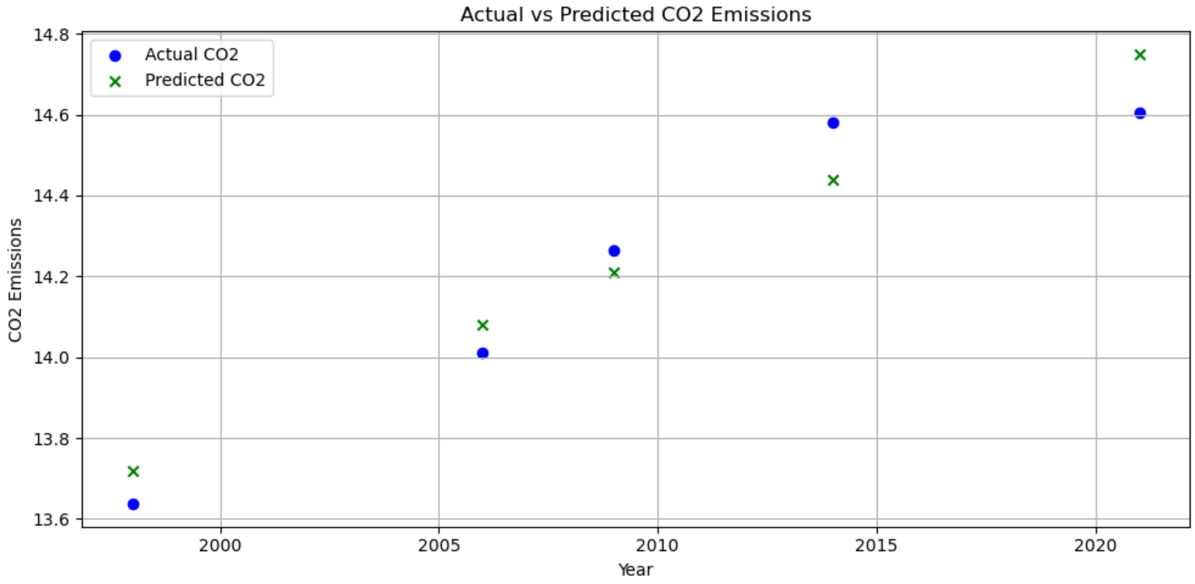


**Figure – 6**

The figures above are the outputs arrived from deploying the ‘Support Vector Machine’ model; where the former (figure 16) showcases the actual $CO_2$ emission values and their predicted values for 5 different years and; the latter (figure 17) shows the graph plot of same data.

### 4.2.3 Decision Tree vs Support Vector Machine

| Year | Actual_CO2 | Predicted_CO2 (DT) | Predicted_CO2 (SVM) |
|---|---|---|---|
| 2006 | 14.0104 | 13.94 | 14.08 |
| 2014 | 14.5796 | 14.49 | 14.44 |
| 1998 | 13.6367 | 13.71 | 13.72 |
| 2021 | 14.6043 | 14.6 | 14.75 |
| 2009 | 14.2633 | 14.17 | 14.21 |

**comparison of both the models (Table – 11)**

The figure above shows the comparison of actual $CO_2$ emission values with both predicted $CO_2$ emission values from ‘Decision Tree Classifier’ and ‘Support Vector Machine’. Whilst individually looking at the each model, both of them respectively fitted well but when looking at them in comparison decision tree predicted values seems to be more closer than to the actual values of $CO_2$ emissions than SVM predicted values.

# CHAPTER 5

# SUMMARY OF FINDINGS

### 5.1 ARDL Model

In this part we examined the relationship between carbon emissions (Ln_CO2) and several independent variables, namely fossil fuel energy consumption (Ln Ec), trade openness (Ln Trade), and urban population (Ln UP).

- The descriptive statistics provided a basic overview of the variables under consideration.
- From cross correlation we understand that Ln_CO2 displayed a moderate positive correlation with Ln Trade and a strong positive correlation with Ln UP. Conversely, it exhibited a weak positive correlation with Ln Ec.
- Next we conducted a unit root test in Eviews, all variables were found to lie in the first difference order I(1), indicating stationarity after differencing.
- Then the ARDL Bound Test was employed to assess cointegration among the variables. The results suggested absence of cointegration among the variables, indicating they may not move together in the long run.
- Finally after the ARDL model was deployed among the independent variables, only fossil fuel energy consumption (Ln Ec) demonstrated a statistically significant impact on CO2 emissions, as indicated by the smaller p-value associated with its coefficient.
- Succeeding this we ended by conducting ARDL diagnostic tests such as auto-serial correlation and heteroscedasticity.

### 5.2 Machine Learning Models

In this part of this project, we predicted the values of $CO_2$ emissions using “Decision Tree Classifier” and “Support Vector Machine”.

- The descriptive statistic has provided a basic overview of the Ln_CO2 variable since it is the target variable.
- Following each model’s training, we evaluated the performances of the models using various metrics. These metrics provide insights into the model's ability to fit the data.
- The Decision Tree Classifier model effectively predicted $CO_2$ emission values for various years, with a high accuracy of 96.06% and the comparison between actual and predicted $CO_2$ emission values suggested that the model fits the data well.
- While, Support Vector Machine model performed well, with low error metrics and a high r2 score of 91.55%, indicating a significant proportion of the variance.

- But however, the SVM model's predictions showed slightly higher errors (10.58%) compared to the Decision Tree Classifier’s predictions ( 7.23%).
- Based on the comparison, the Decision Tree Classifier appeared to be slightly superior in terms of accuracy and closeness to actual values.

# CHAPTER 6

# CONCLUSION

Based on the analysis of the econometric analysis model, we can conclude that fossil fuel energy has a significant impact on carbon emissions in the country under study .i.e. India. The ARDL model found that among the independent variables considered, only fossil fuel energy use has a statistically significant effect on $CO_2$ emissions. As for the prediction models the decision tree classification model seemed to be slightly better in terms of accuracy and closeness to actual values. Overall, the findings suggest that reducing the energy consumption of fossil fuels can contribute to the reduction of carbon dioxide emissions in the study country. In conclusion, both the Decision Tree Classifier and the Support Vector Machine models demonstrated effective performance in predicting $CO_2$ emissions. But however, based on the comparison, the Decision Tree Classifier appeared to be slightly superior in terms of accuracy and closeness to actual values.

It is important to keep in mind that these models are based on historical data and may not reflect unanticipated future events or changes. These models should therefore be used as tools to inform decisions rather than as indicators of future $CO_2$ emissions.

Overall, the findings of this work highlight the importance of considering multiple factors when assessing carbon emissions and potential impacts on reducing fossil fuel energy consumption. These insights can be useful in informing policy decisions aimed at reducing carbon emissions and climate change impacts.

# REFERENCES

Ali H S, Abdul-Rahim AS, Ribadu M B (2016) Urbanization and carbon dioxide emissions in Singapore: evidence from the ARDL approach. Environmental Science Pollution Research 24, 1967–1974 (2017). https://doi.org/10.1007/s11356-016-7935-z

Ali H S, Law S H, Zannah T I (2016) Dynamic impact of urbanization, economic growth, energy consumption, and trade openness on CO2 emissions in Nigeria. Environmental Science Pollution 23, 12435–12443 (2016). https://doi.org/10.1007/s11356-016-6437-3

Ali R, Bakhsh K, Yasin M A (2019) Impact of urbanization on $CO_2$ emissions in emerging economy: Evidence from Pakistan. Sustainable Cities and Society, Volume 48, July 2019, 101553. https://doi.org/10.1016/j.scs.2019.101553

Bloom DE, Canning D, Fink G (2008) Urbanization and the wealth of nations. Science 319(5864):772–775. https://doi.org/10.1126/science.1153057

Cole MA, Neumayer E (2004) Examining the impact of demographic factors on air pollution. Population and Environment 2(1):5–21.

Cortes C, Vapnik V (1995) Support-vector networks. Machine Learning 20, 273–297. https://doi.org/10.1007/BF00994018

Dietz T, Rosa EA (1994) Rethinking the environmental impact of population, affluence and technology. Human Ecology Review 1:277– 300

Dietz T, Rosa EA (1997) Effects of population and affluence on CO2 emissions. Proceedings of the National Academy of Sciences USA 94(1):175–179

Fosu O. A. E. & Magnus F. J. (2006) Bounds Testing Approach to Cointegration: An Examination of Foreign Direct Investment Trade and Growth Relationships. American Journal of Applied Sciences, 3(11), 2079-2085. https://doi.org/10.3844/ajassp.2006.2079.2085

Glaeser E (2011) Cities, Productivity, and Quality of Life. Science 333(6042):592–594. https://doi.org/10.1126/science.1209264.

Hossain S (2012) An econometric analysis for CO2 emissions, energy consumption, economic growth, foreign trade and urbanization of Japan. Low Carbon Econ 2012(3):92–105. http://dx.doi.org/10.4236/lce.2012.323013

Li J, Huang X, Kwan M, Yang H, Chuai X (2018) The effect of urbanization on carbon dioxide emissions efficiency in the Yangtze River Delta, China. Journal of Cleaner Production 188(2018) 38 – 48. https://doi.org/10.1016/j.jclepro.2018.03.198

Martinez-Zarzoso I, Maruotti A (2011) The impact of urbanization on $CO_2$ emissions: evidence from developing countries. Ecological Economics 70:1344–1353. https://doi.org/10.1016/j.ecolecon.2011.02.009

Ouyang X and Lin B (2016) Carbon dioxide (CO2) emissions during urbanization: A comparative study between China and Japan. Journal of Cleaner Production 143:356-368. https://doi.org/10.1016/j.jclepro.2016.12.102

Pata U K (2017) The effect of urbanization and industrialization on carbon emissions in Turkey: evidence from ARDL bounds testing procedure. Environmental Science and Pollution Research 25, 7740–7747 (2018). https://doi.org/10.1007/s11356-017-1088-6

Pesaran HM, Shin Y (1999) Autoregressive distributed lag modelling approach to cointegration analysis. Cambridge University Press.

Pesaran MH (1997) The role of economic theory in modelling the long run. The Economic Journal 178–191.

Pesaran MH, Shin Y, Smith RJ (2001) Bounds testing approaches to the analysis of level relationships. Journal of Applied Econometrics 16:289–326

Poku F A (2016) Carbon Dioxide Emissions, Urbanization and Population: Empirical Evidence in Sub Saharan Africa. Energy Economics Letters, 2016, 3(1): 1-16. http://dx.doi.org/10.18488/journal.82/2016.3.1/82.1.1.16

Poumanyvong P, Kaneko S (2010) Does urbanization lead to less energy use and lower CO2 emissions? A cross-country analysis. Ecological Economics 70 (2010) 434–444. https://doi.org/10.1016/j.ecolecon.2010.09.029

Saidi K, Mbarek MB (2016) The impact of income, trade, urbanization, and financial development on CO2 emissions in 19 emerging economies. Environmental Science Pollution Research 24, 12748–12757 (2017). https://doi.org/10.1007/s11356-016-6303-3

Shahbaz M, Loganathan N, Muzaffar A T, Ahmed K, Jabran M A (2015) How urbanization affects CO2 emissions in Malaysia? The application of STIRPAT model. Elsevier, Renewable and Sustainable Energy Reviews 57 (2016) 83–93. https://doi.org/10.1016/j.rser.2015.12.096.

Shahbaz M, Sbia R, Hamdi H, Ozturk I (2014) Economic growth, electricity consumption, urbanization and environmental degradation relationship in United Arab Emirates. Ecological Indicators 45:622–631. https://doi.org/10.1016/j.ecolind.2014.05.022

Shehu M (2019) Does urbanization intensify carbon emissions in Nigeria? The European Journal of Applied Economics 17(2):161-177. https://doi.org/10.5937/ejae17-19472

Song YY, Lu Y (2015) Decision tree methods: applications for classification and prediction. Shanghai Arch Psychiatry. 27(2):130-5. https://oi.org/10.11919/j.issn.1002-0829.215044.

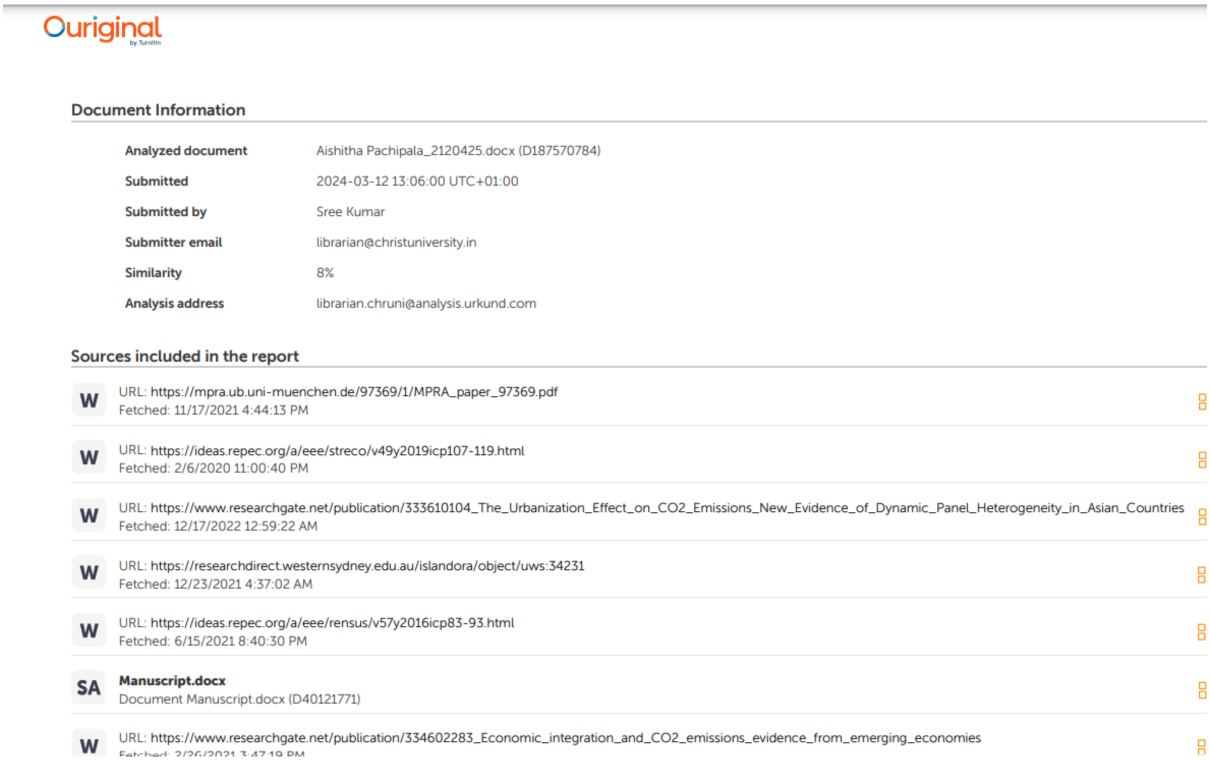
Stewart Jr, CT, & Lee, JH (1986) Urban concentration and sectoral income distribution. The Journal of Developing Areas, 3:357–368.

Vapnik V (1998) The Support Vector Method of Function Estimation. J. A. K. Suykens et al. (eds.), Nonlinear Modelling. https://doi.org/10.1007/978-1-4615-5703-6_3

Wang F, Fan W, Liu J, Wang G, Chai W (2020) The effect of urbanization and spatial agglomeration on carbon emissions in urban agglomeration. Environmental Science and Pollution Research (2020) 27:24329–24341. https://doi.org/10.1016/j.jclepro.2017.09.273

York R, Rosa EA, Dietz T (2003) STIRPAT, IPAT and IMPACT: analytic tools for unpacking the driving forces of environmental impacts. Ecological Economics 46(3):351–365 https://doi.org/10.1016/S0921-8009(03)00188-5

# ANNEXURE



Ouriginal

## Document Information

| | |
|---|---|
| **Analyzed document** | Aishitha Pachipala_2120425.docx (D187570784) |
| **Submitted** | 2024-03-12 13:06:00 UTC+01:00 |
| **Submitted by** | Sree Kumar |
| **Submitter email** | librarian@christuniversity.in |
| **Similarity** | 8% |
| **Analysis address** | librarian.chruni@analysis.urkund.com |

## Sources included in the report

W URL: https://mpra.ub.uni-muenchen.de/97369/1/MPRA_paper_97369.pdf
Fetched: 11/17/2021 4:44:13 PM

W URL: https://ideas.repec.org/a/eee/streco/v49y2019icp107-119.html
Fetched: 2/6/2020 11:00:40 PM

W URL: https://www.researchgate.net/publication/333610104_The_Urbanization_Effect_on_CO2_Emissions_New_Evidence_of_Dynamic_Panel_Heterogeneity_in_Asian_Countries
Fetched: 12/17/2022 12:59:22 AM

W URL: https://researchdirect.westernsydney.edu.au/islandora/object/uws:34231
Fetched: 12/23/2021 4:37:02 AM

W URL: https://ideas.repec.org/a/eee/rensus/v57y2016icp83-93.html
Fetched: 6/15/2021 8:40:30 PM

SA **Manuscript.docx**
Document Manuscript.docx (D40121771)

W URL: https://www.researchgate.net/publication/334602283_Economic_integration_and_CO2_emissions_evidence_from_emerging_economies
Fetched: 2/26/2021 3:47:19 PM