ASSESSING INTERCONNECTED FACTORS IN CO$_2$ EMISSIONS: A CASE STUDY OF INDIA USING PRINCIPAL COMPONENT ANALYSIS

Yiteng Zhang$^{1,2}$, Arjun Pakrashi$^{1,3}$, Soumyabrata Dev$^{1,2,3}$

$^1$University College Dublin, Dublin, Ireland
$^2$The ADAPT SFI Research Centre, Dublin, Ireland
$^3$Beijing-Dublin International College, Beijing, China
Background

- Air pollution is one of the major problems facing humanity today. Advances in computer science, including machine learning, offer new possibilities for monitoring and predicting air pollution [1].
- Research currently focuses on specific greenhouse gas sectors like agriculture and industry [2]. However, there's a gap in understanding emissions on a national scale.

What is our paper about?

- Focuses on Carbon dioxide (CO\(_2\)) emissions in India and assesses the relationship between CO\(_2\) and other emission indicators and national development.
- Aims to shed light on the dependencies between different variables and their impact on estimating CO\(_2\) emissions.
- Utilizes PCA to analyze the variance in the data and reduce the feature space for predictive modeling.
Objectives

- Assessing the strength of the relationship between CO$_2$ emissions and associated indicators.
- Evaluating the predictive capacity of indicators for model development.
- Constructing machine learning models for precise greenhouse gas emission predictions.
Contribution

- Research on CO₂ emissions not only contributes to a better understanding of the impacts of climate change, but also provides a basis for action on sustainable development.
- Provides a theoretical basis for further development of machine learning models for accurate prediction of CO₂ emissions.
Data collection

- We collected 78 different data related to CO$_2$ emissions from India.
- The dataset used for this study was selected as the 10 most complete and relevant features [3].

What is Principal Component Analysis?

- Principal Component Analysis (PCA), is a dimensionality reduction technique commonly used in machine learning and statistics.
- The main goal of PCA is to transform a high-dimensional dataset into a new coordinate system, where the axes are the principal components.
- These principal components are linear combinations of the original features and are chosen in such a way that they capture the maximum variance in the data.
The figure demonstrates that more than 98% of the variance in the dataset is explained by the first three principal components. The high percentage of variance explained by the first three principal components indicates a strong interdependence or correlation between the input variables.
Biplot representation of CO$_2$ emission variables

- Biplot is a graphical representation that allows for the simultaneous visualization of samples and variables in a data matrix.
- The correlation between variables along the y-axis is less uniform compared to the x-axis.
- The distributions along the x-axis direction are all positive. This implies that these variables contribute positively to the first principal component.
- The distribution along the y-axis direction is described as relatively dispersed.
Conclusion

- The results of the PCA process show that these metrics in the selected dataset are highly correlated and can be used as a basis for further building machine learning models.
- Validated that the processing of data related to CO$_2$ emissions through PCA techniques is effective.
Future works

- Based on previous research, we will train dataset using popular deep learning models like CNN, LSTM, GRU, etc. Then compare the prediction accuracy of different models and select the best one.
- On the basis of which further optimisation is done to propose our new models. Make it capable of efficiently predicting greenhouse gas emissions.
Thanks

Contact us:

yiteng.zhang@ucdconnect.ie
arjun.pakrashi@ucd.ie
soumyabrata.dev@ucd.ie