**EDA**

The dataset csv file was loaded, after which the fir five rows of data were previewed to have a sense of the data. Afterward, a copy of the data was made on which the pre-processing would take place, in order to preserve the original dataset.

On the new copy of the dataset, the number of missing values on each column of he dataset was checked, and since the ratio of missing data to valid data was very small, the missing data points were dropped for simplicity. The same was done for duplicate data points.

From the dataset description, it can be observed that the different years basically serve as the time series index of the emissions for each category. Thus, the bulk of the processing logic lies in how to effectively parse through the categorical features in order to extract the best level of analysis which will make the data easier to grasp and process. To achieve this, the categorical features were parsed to get the total number of unique values in each feature, as well as to examine the relationship between the different categorical features and the emissions foreach time period.

The data had three four categorical features: 'Area', 'Item', 'Element', 'Unit'. Of these four, only the first three were useful as the unit column basically had the same value all through.

‘Area’, which represented the country or region for which data was gathered had 268 unique values. ‘Item’ which represented the activity which generated the emissions had 42 unique values, and ‘Element’ which represented the type of emission had 9 unique values.

After this exploration, it became clear that for each country, it would make more logical sense to classify the emissions by Element as it makes the number of categories manageable.

**PREPROCESSING**

*Data Restructuring*

The data was then transformed to stack all the years into one column as opposed to each year having a separate column. After stacking the data by year, the first 500 entries were visually inspected and a pattern was observed; it was observed that some emission categories were subcategories of others, as they consistently added up to make up the super- category across the different Item column values. In addition, the CO2 equivalents from different Element added up to make up CO2 equivalent for each Item.

The was parsed to solve for these observations. After summing up the sub-categories and removing duplicate data, the ‘Element” column then had 4 unique items as opposed to the initial 9: 'Emissions (N2O)', 'Emissions (CO2eq) (AR5)', 'Emissions (CH4)', 'Emissions (CO2)'

To make further pre-processing lighter, only data for one country was used to simplify the structure of the data. Using a pivot table, the data was restructured such that the unique values in the Element column were split into separate columns and the year column was made the index of this new DataFrame structure. With this setup, for any country selected, the emission data for all the years on the dataset is filtered out and organised in columns according to the element emitted.

*Statistical Tests*

The Augmented Dickey-Fuller test which measures to ascertain that the mean and standard deviation of the data remains fairly constant with time was carried out on each Element of the data and it was found that the data was non-stationary with time, however, applying one differencing to the dataset solved the problem to a large extent.

In addition, the Granger Causality test was carried out on the data and it was observed that N2O emissions had causality effect on CH4 and CO2 emissions.

*Train-Test Split*

The data was split into train and test sets with the first 18 years of data for training and the last 3 years for training

*Model Building and Evaluation*

Since the Granger Causality test showed a relationship between the Elements of the dataset, it made sense to use a predictive model which could take into account what was happening with other features in the dataset when making forecasts for a particular feature. Thus, the VARMAX (Vector Autoregressive Moving Average model with eXogenous variables) was used since it satisfies this condition and is also explainable, compared to deep learning models which take a black-box approach to model building.

The model was trained on he training set and tested for accuracy using the test dataset. To assess the accuracy of the model, the model was used to predict the last three years of data stored in the test dataset and the root mean squared error (RMSE) between the test set and the predictions was calculated. The smaller the RMSE, the more accurate the model was determined to be.

The VARMAX model was trained using VAR and VMA order values of ‘p’ and ‘d’. To ensure the best model was in use, hyperparameter tunning was carried out by training different VARMAX models using a combination of ‘p’ and ‘d’ values to give the model with the best RMSE score for all three Elements being forecasted for. It was taken that if the RMSE of the model is <10% of the mean emission value for that Element, then the model is acceptable.

**Side Note: To build a table of results and conclude on the effectiveness of the model, you can choose a set of countries which is diverse enough to cover the range of the data, possibly according to how the UN groups countries, then run the pipeline for each country and then compare and contrast the RMSE of the model for each of those countries**