EDSA Regression Project

Erich du Plessis

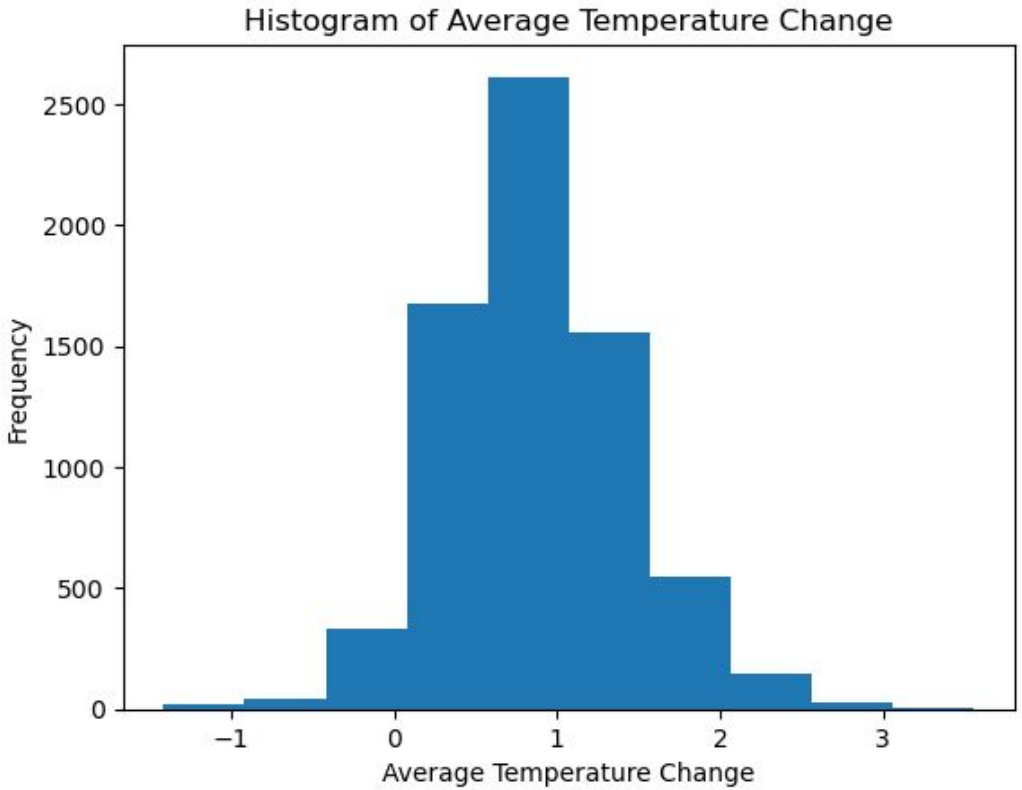
Introduction

The project involved analysing the impact of agri-food sector related emissions on the change in average temperatures between 1990 and 2020 for regions across the globe. The dataset was compiled by the Food and Agriculture Organization (FAO) and the Intergovernmental Panel on Climate Change (IPCC) and included 30 features plus the target variable.

From the dataset we created and trained four regression models to predict the change in average temperatures based on the independent variables in the dataset. The fours regression models used:

- Linear Regression Model
- Multi - Linear Regression Model
- Multi - Linear Regression Model with LASSO Regularisation
- Random Forest Regression Model

Exploratory Data Analysis

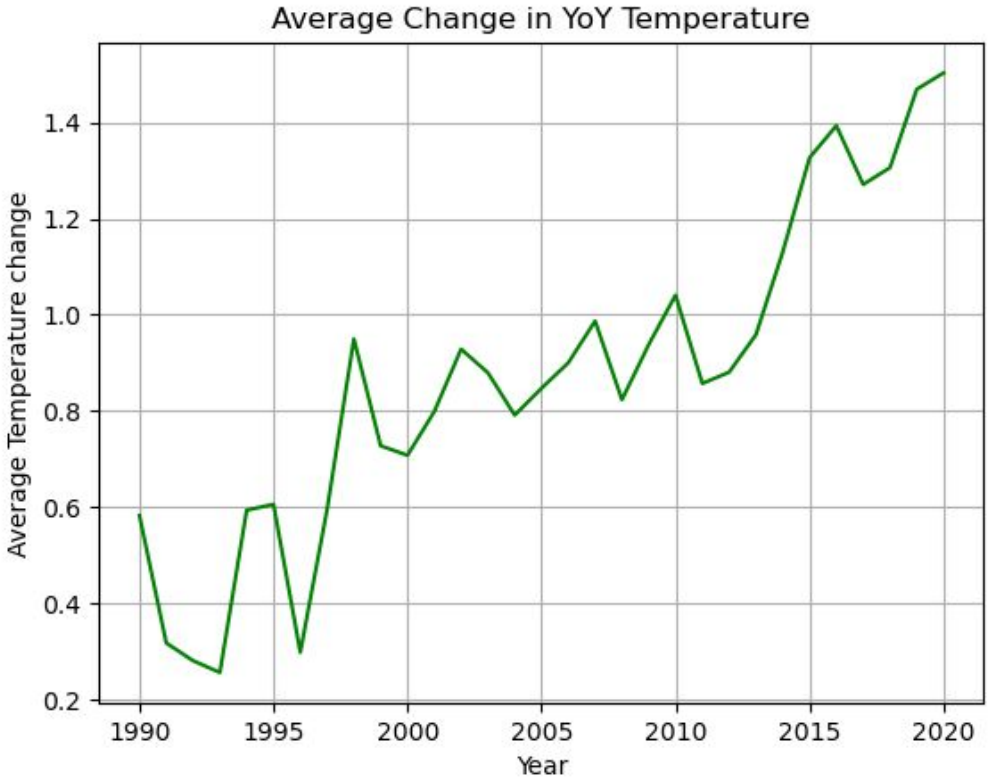


The target variable / response variable exhibits a normal distribution around 1 degree celsius.

The vast majority of observations show a positive change in average temperatures from the previous year. This implies that global temperatures likely increases over time as very few recorded a drop in average temperatures between 1990 and 2020 for regions across the globe.

Very few observations recorded a temperature change above +3 degrees celsius.

Exploratory Data Analysis



As expected from the previous histogram, the change in average temperatures has consistently increased over time with some variation along the way. It is clear that something is driving global temperatures up over time.

In the early 1990's the average change in temperature was around the 0.5 mark, but has since increased to 1.5 in 2020. This implies that not only are global temperatures increasing over time, but the speed or magnitude at which they are going up is also increasing.

The following slide shows that total emissions could potentially be linked to the trends observed for the response variable. It shows that the movement of the two variables are similar between 1990 and 2020 (Normalised to the same axis).

Exploratory Data Analysis

Exploratory Data Analysis

```
Correlation of features with Average Temperature Change:
Year                                0.545932
Food_Retail                         0.221330
IPPU                                0.210191
Food_Transport                      0.177666
Food_Household_Consumption          0.174372
On_farm_energy_use                  0.153741
Manure_applied_to_Soils             0.142669
On_farm_Electricity_Use             0.130860
Manure_Management                   0.129779
Urban_population                    0.125347
Pesticides_Manufacturing            0.114278
Agrifood_Systems_Waste_Disposal     0.104177
Female_Population                   0.095468
Male_Population                     0.094754
Food_Packaging                      0.092165
Drained_organic_soils               0.090036
Food_Processing                     0.073336
total_emission                      0.055846
Rice_Cultivation                    0.053027
Manure_left_on_Pasture              0.052080
Rural_population                    0.026955
Fertilizers_Manufacturing          -0.011104
Fires_in_organic_soils             -0.033340
Net_Forest_conversion              -0.051372
Savanna_fires                      -0.054835
Forest_fires                       -0.095693
Fires_in_humid_tropical_forests    -0.121328
Forestland                         -0.130254
```

Correlation analysis was conducted on the log transformed data to better understand the relationships between the response variables and predictors.

The data does not illustrate strong relationships with the response variable. This could imply that the true mechanism behind the temperature changes are not included in the data.

It could also be hidden by regional trends where the various factors interact with one another in different ways or where other excluded factors are more of a driving force compared to the others.

Regression Models

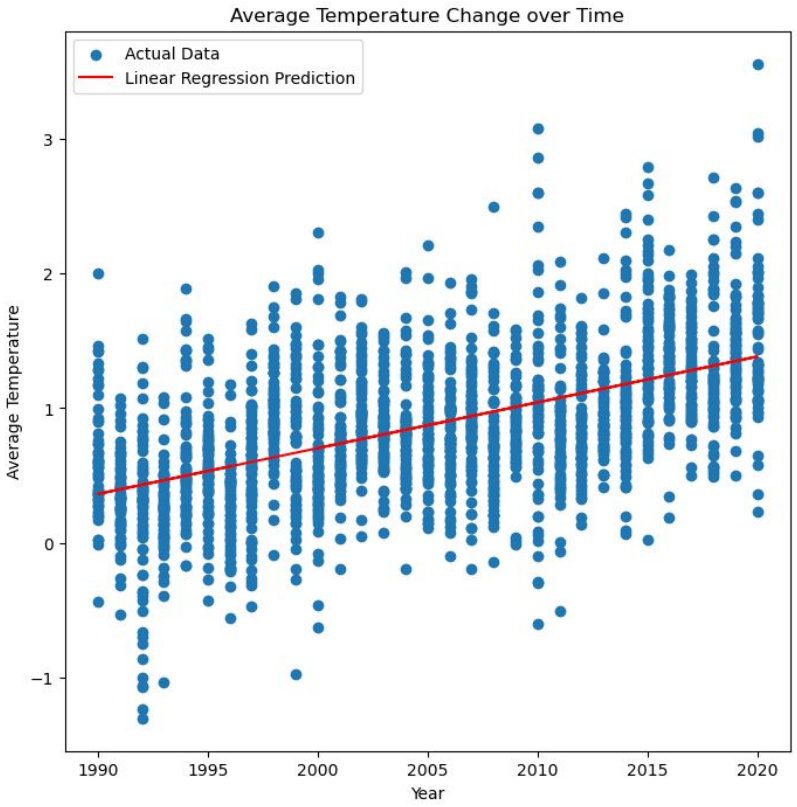
Four regression models were trained using 80% of the dataset. The models were then evaluated for performance on the remaining 20% unseen data. The models included a linear regression model, multi-linear regression model, LASSO regularisation on a multi-linear model and a Random Forest model.

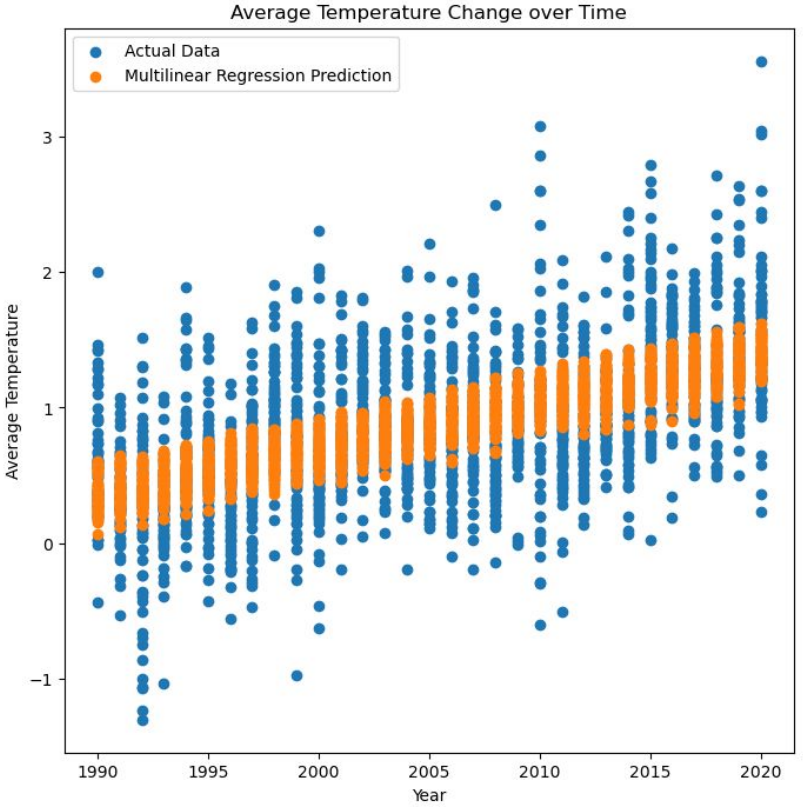
Model	MSE	RMSE	R Squared
Linear	0.2201	0.4692	0.2964
Multi-Linear	0.2061	0.4540	0.3411
LASSO	0.1822	0.4269	0.4175
Random Forest	0.1312	0.3622	0.5805

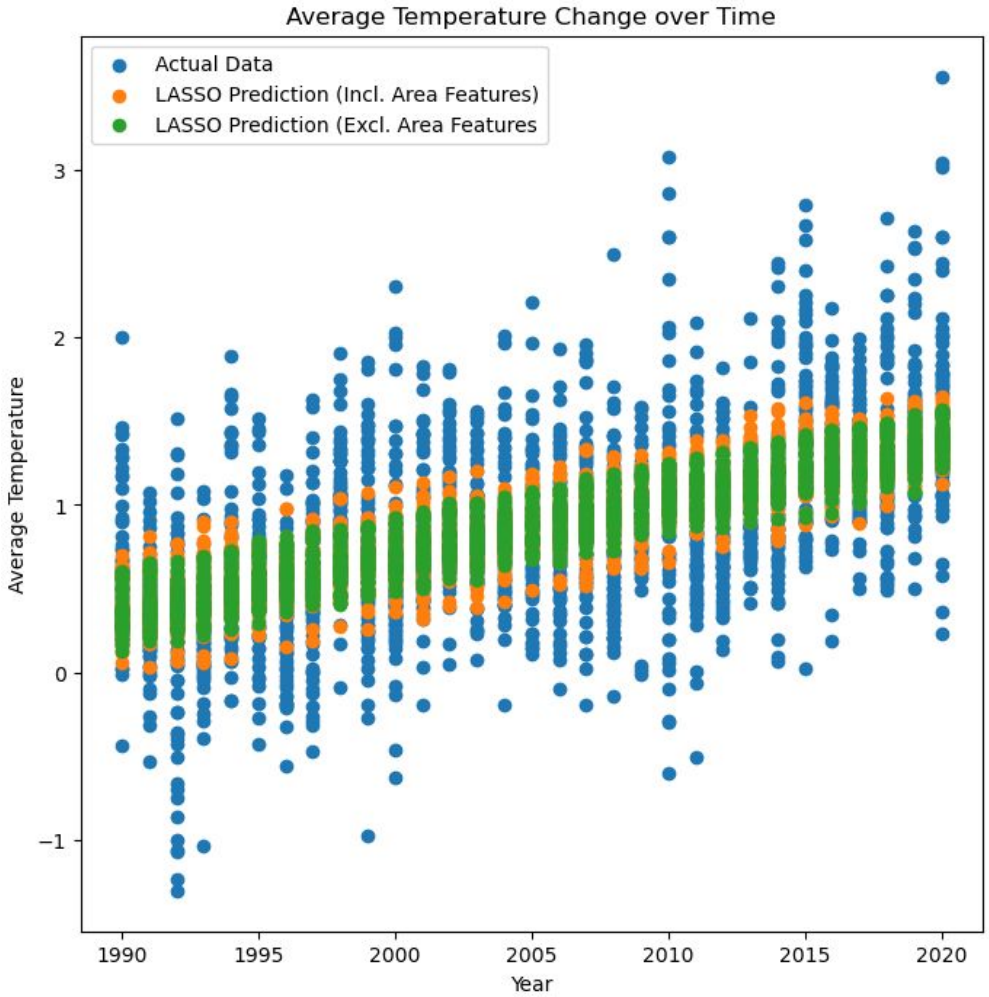
The Linear model has the worst performance with the highest MSE on unseen data. The multi-linear model improved with the inclusion of multiple features while feature selection through LASSO regularisation resulted in further improvement on model performance. The Random Forest Model however improved significantly on the linear models, likely due to its non-parametric nature allowing it to model non-linear trends in the data.

The random forest model did however exhibit signs of overfitting with a MSE of 0.0786 on training data while the linear models performed consistently on the training and test data.

The linear model only considered a single independent variable (Year) to predict the change in the average temperatures. The linear fit below is following the general trend over time but not capable of explaining the variance at each year.
The multi-linear model encapsulated the same trend over time, but managed to incorporate additional features to more accurately predict variations in each year.







LASSO Regularisation was utilise in a multi-linear regression model in an attempt to improve on the manual feature selection conducted for the previous model, as well as to improve overall model performance through modulating the linear coefficients.

The model improved marginally when trained on the data that excluded the encoded areas feature. The model did however improve with the inclusion of the hot-one encoded areas feature, implying that the model was likely able to identify some regional trends and improve model accuracy.

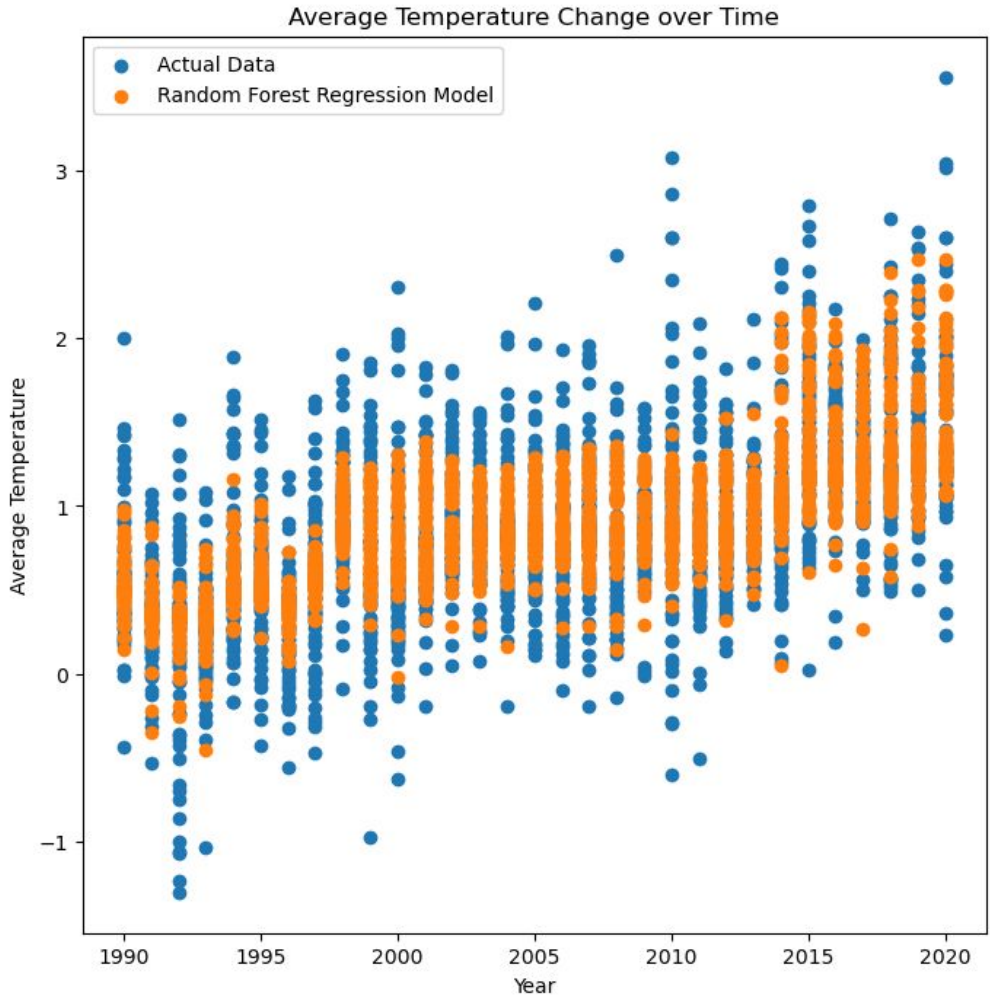
The graph shows that both LASSO models exhibits the same time trend as the standard linear model but is capable of explaining more of the variance at each point in time. We can also observe that the inclusion of the areas data has clearly increased the complexity of the model that has allowed it to predict a wider range of values.

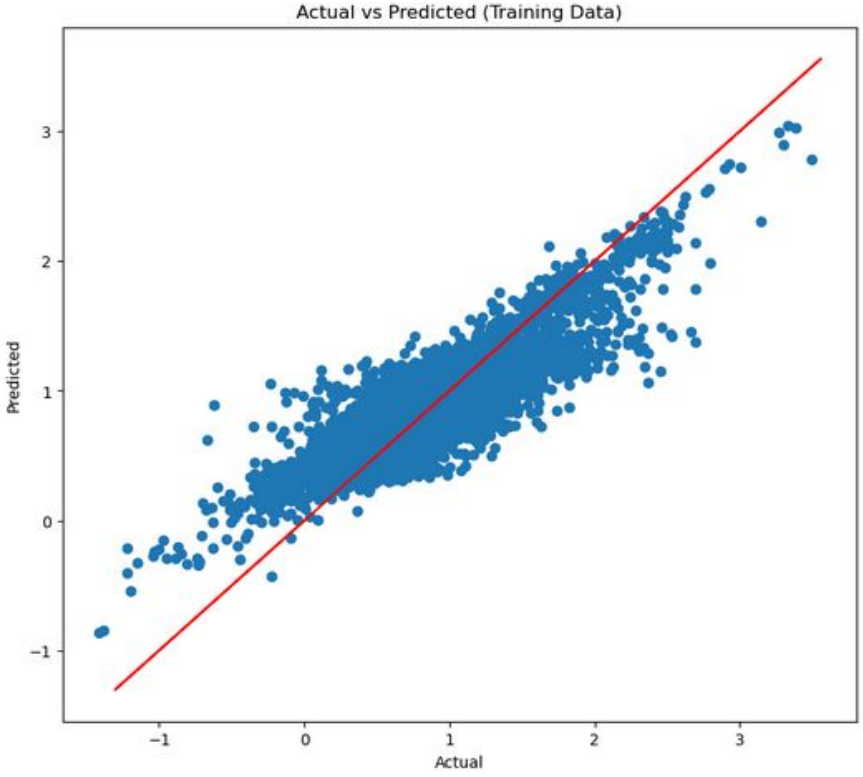
The Random forest model improved significantly on the linear models with a MSE of 0.13 on unseen data. The graph clearly shows that this model is capable of predicting a wider range of values and able to better explain the variance in the response variable.

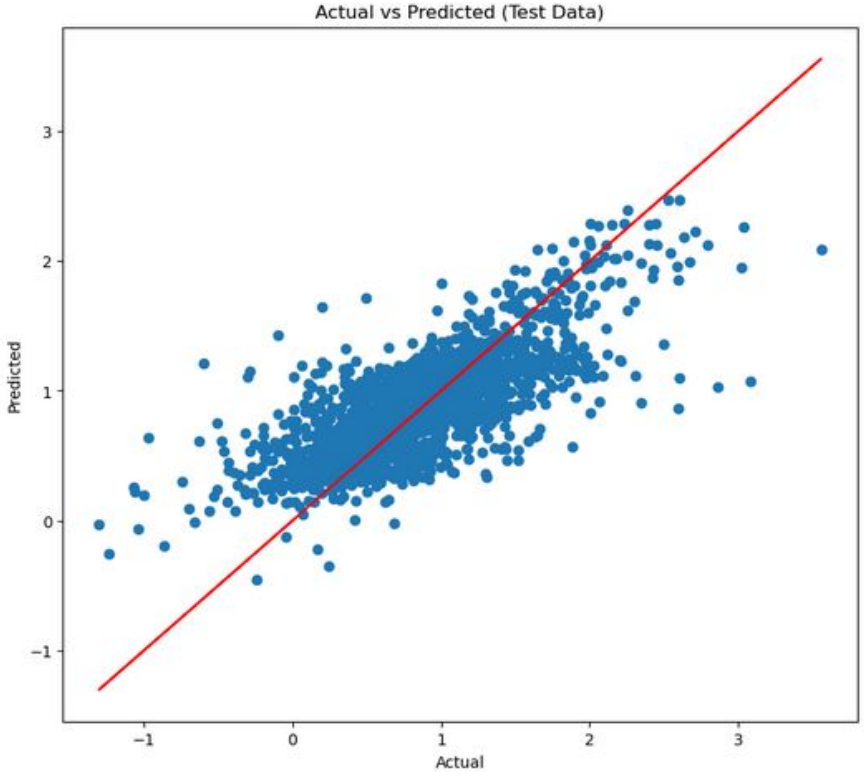
The random forest model exhibits a slightly different trend over time that is not perfectly linear in nature. The linear models exhibited a fairly linear trend over time, while this model shows a more complex trend that more accurately fits the actual data trend.

The Random forest model did undergo hyperparameter tuning using the training data to improve overall performance. The model is however exhibiting signs of overfitting with a relatively large performance difference between the training data and test data. This can also be observed on the graph on the next slide that compares the model on training data and unseen test data.

The most important features in the model was identified to be (In order of importance): Year, IPPU, Manure_left_on_pature, Fires_in_humid_tropical_forests, Drained_organic_soils, Manure_applied_to_soils, Rice_cultivation, Food_transport, Savanna_fires and Fertilizers_manufacturing.




Actual vs Predicted (Training Data)
Predicted
Actual
3
2
1
0
−1
−1
0
1
2
3


Actual vs Predicted (Test Data)
Predicted
Actual
3
2
1
0
−1
−1
0
1
2
3

Conclusion

The analysis conducted on the dataset indicated that agri-food sector related emissions are somewhat correlated with changes in average yearly temperatures across the globe. The data showed that the movement in total emissions over time matched the movement of average temperature changes between 1990 and 2020. This indicates that the two are likely connected, although the strength of the connection did vary by region and the data at each year did not necessarily show a strong correlation.

The feature with the strongest correlation with the response variable was the years in which the data was recorded. This is a clear indication that some contributing factors are not included in this data set. It is likely that other factors / mechanisms are working over time to influence the change in average temperatures. This would explain why we see relatively strong correlation between time and the response variable. This should not be surprising given that we are only looking at agri-food sector related factors in this dataset.

Although we don't observe very strong linear relationships in the dataset, the combination of various features improved the predictive power of our regression models. We constructed and trained four regression models, varying in complexity. The models were trained and then tested on unseen data to compare model performance and accuracy. The most basic model (Single linear model) had the worst performance based on the MSE. Next, the multilinear model improved on the single linear model with the inclusion of additional features and the LASSO regularisation model further improved on that model. Lastly, the Random Forest model proved to be the most accurate on both the test and train data. This is likely due to the fact that the model is a non-parametric model and can identify and utilise non-linear patterns in the data.