Project Report for Group 19

Daniel Tribaldos

Co-authors: Donald Gjoka, Sebastian Barrio, Julian Schumacher

Abstract

This project aims to create innovative algorithms that leverage real-world climate datasets to predict climate change, detect climate anomalies, and find meaningful correlations between various climate-related variables. Our approach included two major algorithms: a prediction/anomaly detection algorithm, and an unsupervised clustering algorithm.

Dataset Description

Clustering Dataset

For clustering, we utilized a dataset from Kaggle containing global land temperatures by major cities. Key attributes included City, Latitude/Longitude, AverageTemperature, and Country. We believed this dataset would allow our unsupervised learning algorithm to effectively discover patterns in geographic and climate-based features.

Prediction and Anomaly Detection Dataset

The prediction model used a merged dataset from two primary sources:

- **Temperature Deviations:** From the Berkeley Earth Project, which recorded global temperature deviations based on a historical average, likely centered around 1951–

1980. These values represented deviations from a global average rather than absolute temperatures.

- **CO_2 Concentrations:** From Zenodo, which provided monthly atmospheric CO_2 levels in parts per million (ppm) from 1850–2013. A value of 1 ppm indicates one molecule of CO_2 for every million molecules of air.

These datasets were merged on a monthly basis and reduced to key features: time, CO_2 ppm, and temperature deviation. Geographic coordinates were dropped to generalize predictions globally.

Machine Learning Methodology

Initial Approach: RNN with LSTM

We initially implemented a Recurrent Neural Network (RNN) with Long Short-Term Memory (LSTM) cells, chosen for their strength in handling sequential data. Despite normalization and sequence grouping, this model underperformed. It returned negative R^2 values (e.g., $R^2 = -50$), indicating that simple averaging outperformed our model. We concluded that the lack of diverse feature columns limited the RNN's learning capacity.

Feed-Forward Neural Network (FFNN)

Due to the limited feature set, we pivoted to a Feed-Forward Neural Network (FFNN), which doesn't rely on sequential input. We implemented a hyperparameter tuner using Keras Tuner to allow the model to self-optimize architecture complexity. Despite limited input features, the FFNN achieved:

- R^2 **Score:** 0.6391, indicating the model explained 63.91% of the variance in the target variable.

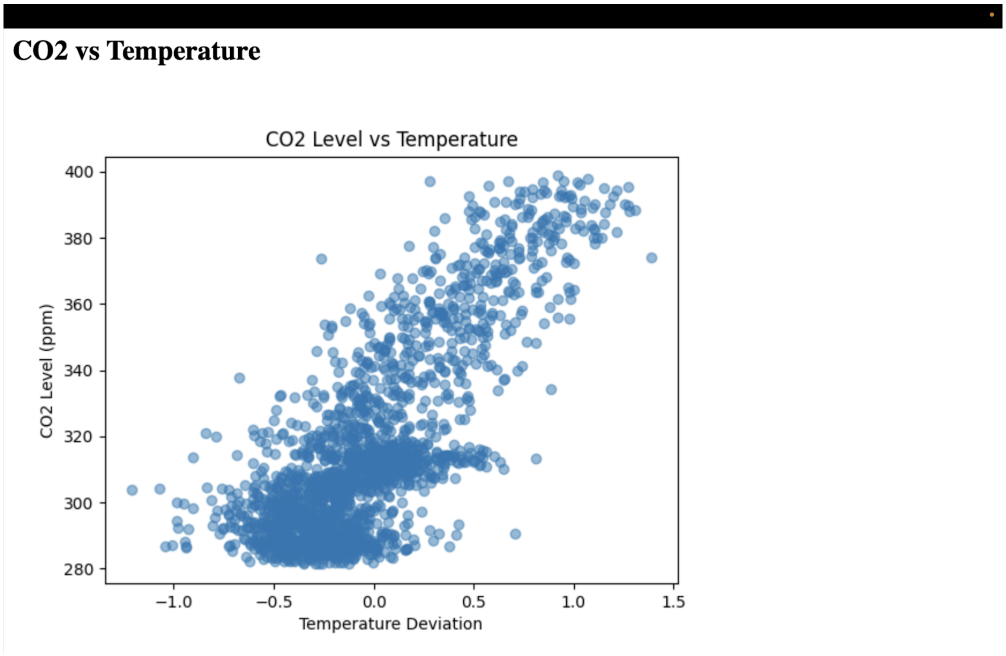
- **Mean Absolute Error (MAE):** 0.0634, meaning the model's predictions were on average 0.0634 units from the actual values.

These results demonstrate that even with limited features, neural networks can deliver moderately accurate predictions over long temporal ranges.

Clustering Results

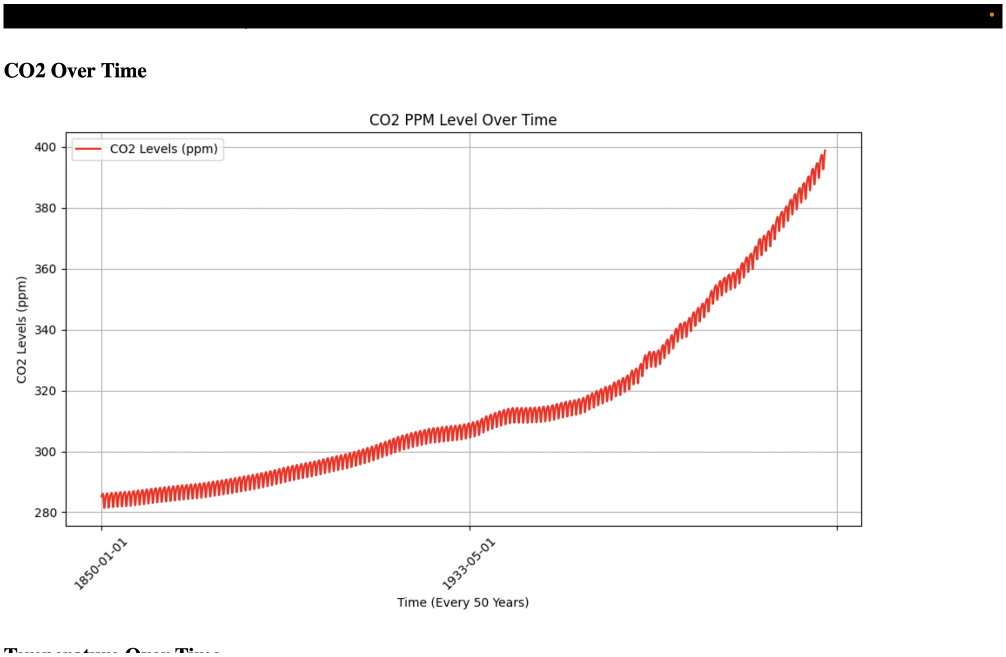
We used K-Means clustering on the city-based dataset. After manual testing, we found $k = 5$ to yield the most meaningful clusters. We confirmed this mathematically using an elbow graph. Both MinMax and Standard Scaler normalizations were tested, and Standard Scaling performed slightly better.

Correlations and Analysis



CO2 vs Temperature

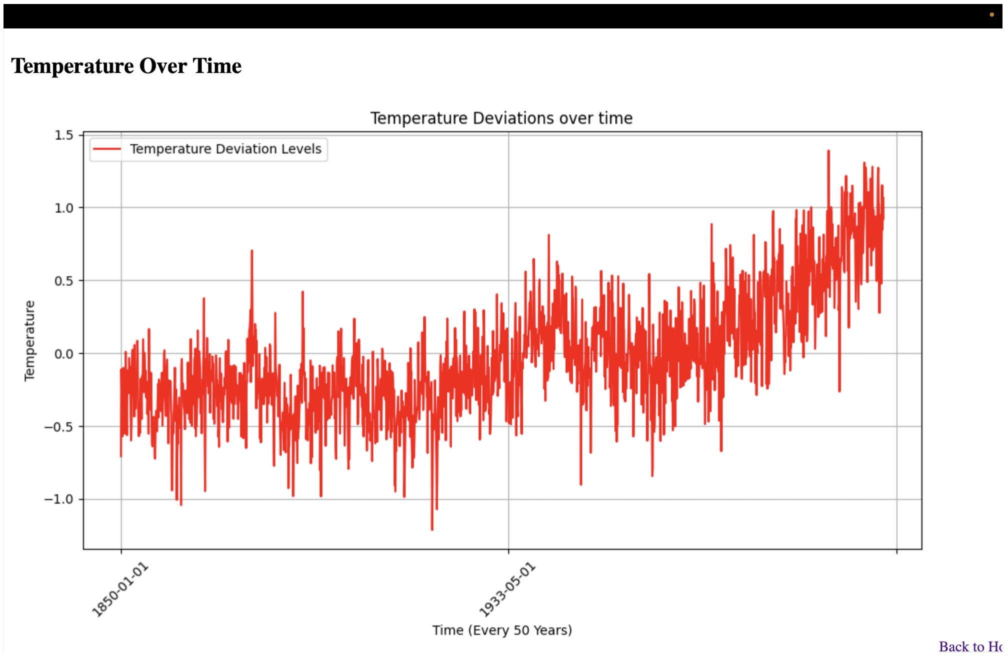
Figure 1: Scatter plot of CO_2 ppm vs Temperature Deviation



CO2 Over Time

Temperature Over Time

Figure 2: CO_2 PPM levels over time



Temperature Over Time

Figure 3: Temperature deviation trend over time

From the visualizations:

- CO_2 ppm levels correlate directly with increased temperature deviation.

- When temperature deviation is near zero or negative, CO_2 shows little impact.
- Both temperature deviation and CO_2 ppm rise steadily over time.

This supports the hypothesis that increased CO_2 emissions directly contribute to global warming.

Future Predictions

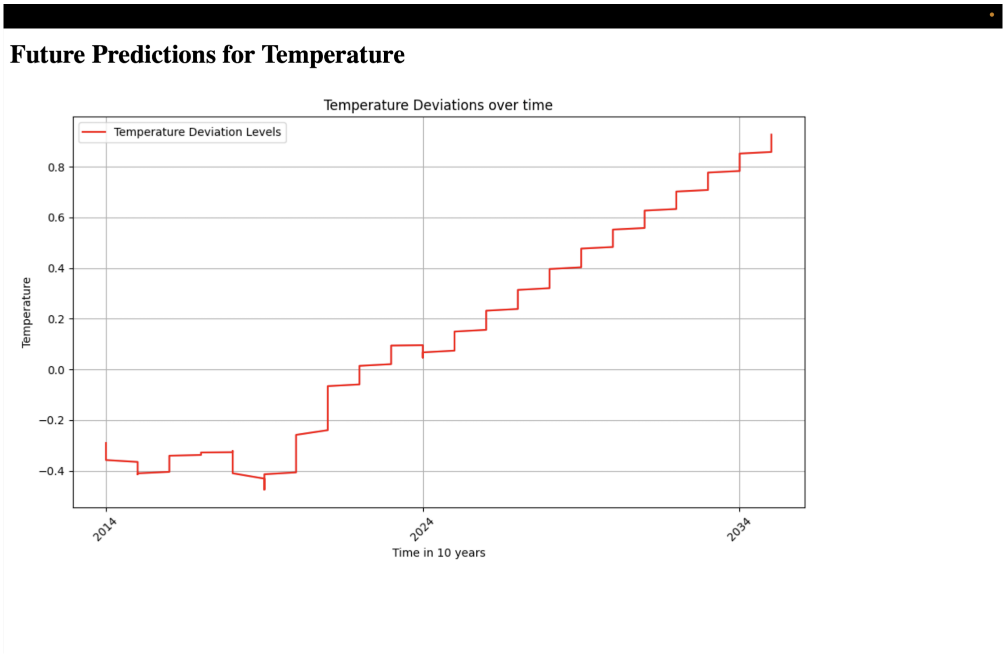

Future Predictions for Temperature

Figure 4: Predicted Temperature Deviations for Future Years (2025 and beyond)

We used our best-performing FFNN to predict future temperature deviations for approximately 20 years. Some of the predictions coincided with historical data post-2013, and the model's estimates aligned well. If actual measurements from 2014–2025 match our predictions, it would validate the model's long-term reliability.

Conclusion

Despite the project's challenges, we successfully developed a neural network capable of making reasonably accurate predictions using simplified global data. Clustering further confirmed regional similarities. This project offers a foundation for scalable and accurate climate forecasting algorithms.