sai_kumar_dataset_and_model_justification[1] (1).docx

Turnitin

Document Details

Submission ID

trn:oid:::31188:99495946

Submission Date

Jun 5, 2025, 4:53 PM GMT+5

Download Date

Jun 5, 2025, 4:54 PM GMT+5

File Name

sai_kumar_dataset_and_model_justification[1] (1).docx

File Size

445.3 KB

7 Pages

1,493 Words

9,346 Characters

0% detected as AI

The percentage indicates the combined amount of likely AI-generated text as well as likely AI-generated text that was also likely AI-paraphrased.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Detection Groups

0 AI-generated only 0%
Likely AI-generated text from a large-language model.

0 AI-generated text that was AI-paraphrased 0%
Likely AI-generated text that was likely revised using an AI-paraphrase tool or word spinner.

Disclaimer
Our AI writing assessment is designed to help educators identify text that might be prepared by a generative AI tool. Our AI writing assessment may not always be accurate (it may misidentify writing that is likely AI generated as AI generated and AI paraphrased or likely AI generated and AI paraphrased writing as only AI generated) so it should not be used as the sole basis for adverse actions against a student. It takes further scrutiny and human judgment in conjunction with an organization's application of its specific academic policies to determine whether any academic misconduct has occurred.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?
The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI-paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?
Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.

AI-Driven Climate Change Model: Dataset, Feature Selection, Modeling Methods, and Justification

1. Dataset Overview and Source

Initially, ClimateSet (CMIP6) was considered for this project. However, due to its complex structure (NetCDF format), large size, and significant preprocessing requirements, I decided to switch to the Kaggle Climate Change Dataset, which is machine learning-ready, clean, and better suited to this project's timeline and modeling goals. This change also aligns with the guidance provided during project discussions. The data set used in this project is the Climate Change Dataset on Kaggle by Bhadra Mohit. It contains 1000 rows and 10 columns of climate and population of various countries for certain years. It contains climate columns such as carbon dioxide emissions, sea level rise, rainfall, and temperature and socio-economy columns such as population and consumption of renewable power. One row is the snapshot of one year's climate of the country and hence can be used both for classification (temperature prediction from the risk category) as well as regression (prediction of mean temperature).

It's from reputable sources like NASA, NOAA, and the United Nations, and it's in machine learning-easy-to-consume format too. Its tabular format, tidiness, and having almost no missing values make it a great one to experiment with and test various AI models without necessarily needing to worry about fancy preprocessing.

2. Column-wise Description

The dataset includes the following key columns:

- The data set consists of the following primary columns:
- Year – Year in which data were gathered.
- Country – The country involved.
- Avg Temperature (°C) – Average temperature per year, our dependent variable.
- CO2 Emissions (Tons/Capita) refers to the CO_2 per capita emissions, which is one of the most powerful drivers of climate change.
- Sea Level Rise (mm) – Sea level rise due to glaciemelt and thermal expansion.
- Rainfall (mm) – Total annual precipitation, a measure of hydrological cycle variation.

- Population – Total National Population.
- Renewable Energy (%) – Percentage of renewable energy in the country's energy market.
- Severe Meteorological Events – Event frequency of occurrences like heatwaves or storms.
- Forest Area (%) – Forest area percentage, a carbon sink.

3. Why This Dataset Was Selected

This data set has been selected because it has been well balanced between exhaustiveness and utility. Data sets like ClimateSet (CMIP6) and ERA5 Reanalysis are highly exhaustive and need to be interpreted in the domain knowledge context and preprocessed in a way so that they are useful in a meaningful sense. They are typically offered in high-level formats like NetCDF and are meant for scientific simulation, not machine learning.

The Kaggle data is pre-formatted, pre-cleaned, and pre-tagged with informative variables ready for model construction. It is also ready for rapid deployment of artificial intelligence models like Random Forest, LSTM, and Gradient Boosting and explainable artificial intelligence package compliant like SHAP and LIME. It is therefore ready for instant scholarly testing, light forecasting, and scenario planning.

4. Feature Selection and Standardization

The dataset consists of aggregated, publicly available data sourced from NASA, NOAA, and United Nations repositories. No personal or sensitive data is involved, and the dataset fully complies with UH’s ethical guidelines for data usage in academic projects.

Seven features were selected based on exploratory data analysis (EDA) and domain relevance. These are:

1. CO2 Emissions (Tons/Capita)
2. Sea Level Rise (mm)
3. Rainfall (mm)
4. Population
5. Renewable Energy (%)
6. Extreme Weather Events

7. Forest Area (%)

Feature selection was guided by correlation analysis and domain knowledge. For example, **CO_2 emissions and population** are known contributors to temperature change, while **renewable energy usage and forest area** act as mitigating factors. After selection, all features were **standardized using StandardScaler** to bring them to a common scale, improving model training stability and speed.

Feature selection was conducted using both domain knowledge and statistical correlation analysis. An initial Exploratory Data Analysis (EDA) was performed, including correlation heatmaps to identify relationships between features and the target variable (Avg Temperature). Features showing strong positive or negative correlation (such as CO_2 emissions, population, and renewable energy) were selected. Additionally, feature importance scores from the Random Forest model were examined to validate these selections.

5. Modeling Methods

Regression Task

The goal was to predict the continuous value of **average temperature** using selected features. Models used:

- **Random Forest Regressor**
- **LSTM (Long Short-Term Memory) Neural Network**
- **Transformer-based Regressor**

Performance was measured using **Root Mean Squared Error (RMSE)**. Among these, **Random Forest** delivered the best balance of accuracy and interpretability.

Classification Task

For classification, the target temperature variable was binned into four categories: **Low**, **Moderate**, **High**, and **Extreme**. Models used:

- **Gradient Boosting Classifier**
- **XGBoost Classifier**
- **MLP Classifier**

The classifiers were trained and evaluated using accuracy and F1-scores. **Gradient Boosting** performed the best and was selected as the final classification model.

6. Model Selection and Justification

Random Forest Regressor

This model was chosen for its high accuracy, robustness to overfitting, and ease of interpretation. It was implemented and tested on this project's dataset and performed well, achieving the best regression results among the models evaluated. Additionally, Random Forest supports feature importance ranking and can be explained using SHAP, making it ideal for real-world climate applications.

LSTM and Transformer

These deep learning models were evaluated for their ability to capture time-series trends. Both models were implemented and tested on this project's dataset. LSTM effectively modeled sequential patterns in the data, though it required significant computational time. Transformer-based models showed promising results but were more complex and resource-intensive. These models were retained for comparative insights but are not the primary models selected for deployment.

Gradient Boosting for Classification

Gradient Boosting was implemented and tested within this project for the classification task (temperature categories). It produced stable learning curves and high accuracy during evaluation on this project's data. Given its interpretability and suitability for scenario-based classification, Gradient Boosting was selected as the final classification model for this project.

7. Literature Survey

This section critically examines recent studies that inform the AI techniques, model choices, and explainability methods used in this project. Artificial Intelligence (AI) has emerged as a powerful tool for modeling climate change. Recent literature explores the effectiveness of various models and interpretability frameworks. This section highlights five key studies that inform the methodologies used in this project.

Agrawal (2023) explores generative AI in sustainability, emphasizing interpretable models like **Random Forest** and **XGBoost** for system optimization under constraints. His work supports using these models for both prediction and resource-aware forecasting. Amiri et al. (2024) offer a comprehensive review of AI techniques for climate change mitigation. They highlight **tree-based ensembles** and **LSTM networks** as effective for capturing complex, nonlinear climate patterns—aligning closely with the model choices in this project.

Cavus et al. (2025), although focused on electric vehicles, demonstrate the strength of **LSTM models** in handling time-series data, reinforcing their use in climate forecasting tasks involving temporal dependencies. Chakraborty et al. (2021) advocate for **explainable AI (XAI)** in environmental prediction, using SHAP for model transparency. Their scenario-based modeling approach mirrors this project's use of SHAP and LIME. Chang and Kidman (2023) stress the ethical importance of using interpretable models in climate education. They argue against black-box models, validating the use of transparent techniques like Gradient Boosting in public tools.

Together, these studies support the integration of ensemble and deep learning models, explainability, and scenario planning. However, most fail to combine all elements cohesively—this project addresses that gap by unifying forecasting, classification, and interpretability into one AI pipeline.

8. Final Model Selection and Explainability

Based on comparative results, the final models selected are:

- **Random Forest Regressor** – For continuous temperature prediction.
- **Gradient Boosting Classifier** – For temperature classification and risk scenario generation.

These models deliver high accuracy and allow integration with **SHAP (global interpretability)** and **LIME (local interpretability)** frameworks, ensuring **transparency** and **trust** in predictions.

This project combines a well-curated climate dataset with proven machine learning techniques to develop a predictive system for climate change modeling. The selected features capture a wide array of environmental dynamics, and the use of interpretable models ensures responsible AI deployment. Through model comparison, literature grounding, and scenario simulations, the system provides a meaningful step toward data-driven climate forecasting.

References

[1] Agrawal, K.P. (2023) ; Organizational sustainability of generative AI-Driven optimization intelligence,; *Journal of Computer Information Systems*, pp. 1–15. https://doi.org/10.1080/08874417.2023.2286540.

[2] Amiri, Z., Heidari, A. and Navimipour, N.J. (2024) ;Comprehensive survey of Artificial intelligence Techniques and Strategies for climate change mitigation,; *Energy*, 308, p. 132827. https://doi.org/10.1016/j.energy.2024.132827.

[3] Cavus, M., Dissanayake, D. and Bell, M. (2025) ;Next generation of electric vehicles: AI-Driven approaches for predictive maintenance and battery management,; *Energies*, 18(5), p. 1041. https://doi.org/10.3390/en18051041.

[4] Chakraborty, D. *et al.* (2021) ;Scenario-based prediction of climate change impacts on building cooling energy consumption with explainable artificial intelligence,; *Applied Energy*, 291, p. 116807. https://doi.org/10.1016/j.apenergy.2021.116807.

[5] Chang, C.-H. and Kidman, G. (2023) ;The rise of generative artificial intelligence (AI) language models - challenges and opportunities for geographical and environmental education,; *International Research in Geographical and Environmental Education*, 32(2), pp. 85–89. https://doi.org/10.1080/10382046.2023.2194036.