Predictive Analytics & Imagery for Environmental Monitoring

-Shambhavi Adhikari, Yeji Kim, Dilrose Karakattil

Abstract

Climate change poses multifaceted challenges, impacting health, food security, biodiversity, and the economy. This study explores predictive analytics and satellite imagery to address climate change effects, focusing on deforestation monitoring, carbon emission analysis, and flood prediction. Using machine learning models, including a Random Forest for emissions and a Custom U-Net for deforestation, we developed predictive tools that provide actionable insights. The findings show high accuracy in predicting carbon emissions and flood risks and successful monitoring of deforestation areas, highlighting the potential for advanced monitoring systems to mitigate environmental threats.

Acknowledgement

We would like to express our heartfelt gratitude to the AI Policy Hackathon team at Johns Hopkins University for organizing this impactful and engaging event. Their dedication to fostering collaboration and addressing real-world AI policy challenges has provided us with an invaluable platform to innovate and learn.

Special thanks to the organizing committee, including Archana Vaidheeswaran, Abe Hou, Amy Wang, Idris Sunmola, Angela Tracy, Joy Yu, Andreas Jaramillo, Jace Lafita, and all other team members, whose efforts have been instrumental in making this hackathon a success. The inclusion of distinguished speakers and mentors from OpenAI, Microsoft, Apart Research, and academia has enriched our experience, inspiring us to push the boundaries of what AI and policy can achieve.

We are also thankful for the judges who took the time to review our work and offer valuable feedback. Your insights have encouraged us to think deeply about the intersection of AI technology and governance. This experience has been an extraordinary opportunity to engage with diverse perspectives and explore the future of responsible AI.

Introduction

Climate change, exacerbated by deforestation and carbon emissions, accelerates environmental degradation with severe consequences on biodiversity, health, and economic stability. Effective monitoring of environmental factors like carbon emissions, deforestation, and rainfall trends is essential for proactive responses. This project applies machine learning to predict climate-related impacts, enabling targeted environmental strategies through data-driven insights.

Problem Statement

The project addresses the growing environmental risks due to climate change, such as:

- **Health Impacts**: Adverse effects from pollution and extreme weather.
- **Food Security**: Impact on agricultural yield due to changing climate conditions.
- **Economic Impacts**: Costs related to disaster response and damage.
- **Biodiversity Loss**: Habitat destruction from deforestation and climate shifts.

Climate change accelerates due to deforestation and rising carbon emissions. A robust monitoring solution is needed to manage and mitigate these impacts effectively.

Objectives

- Analyze key contributors to carbon emissions using a Random Forest algorithm.
- Use satellite imagery to monitor global deforestation.
- Develop a model to predict flood risks based on meteorological and geographical data.

Scope

The study aims to enhance environmental monitoring capabilities by:

- Identifying factors contributing to carbon emissions.
- Enabling real-time deforestation detection.
- Providing predictive insights for flood risks, aiding early warning systems.

Literature Review

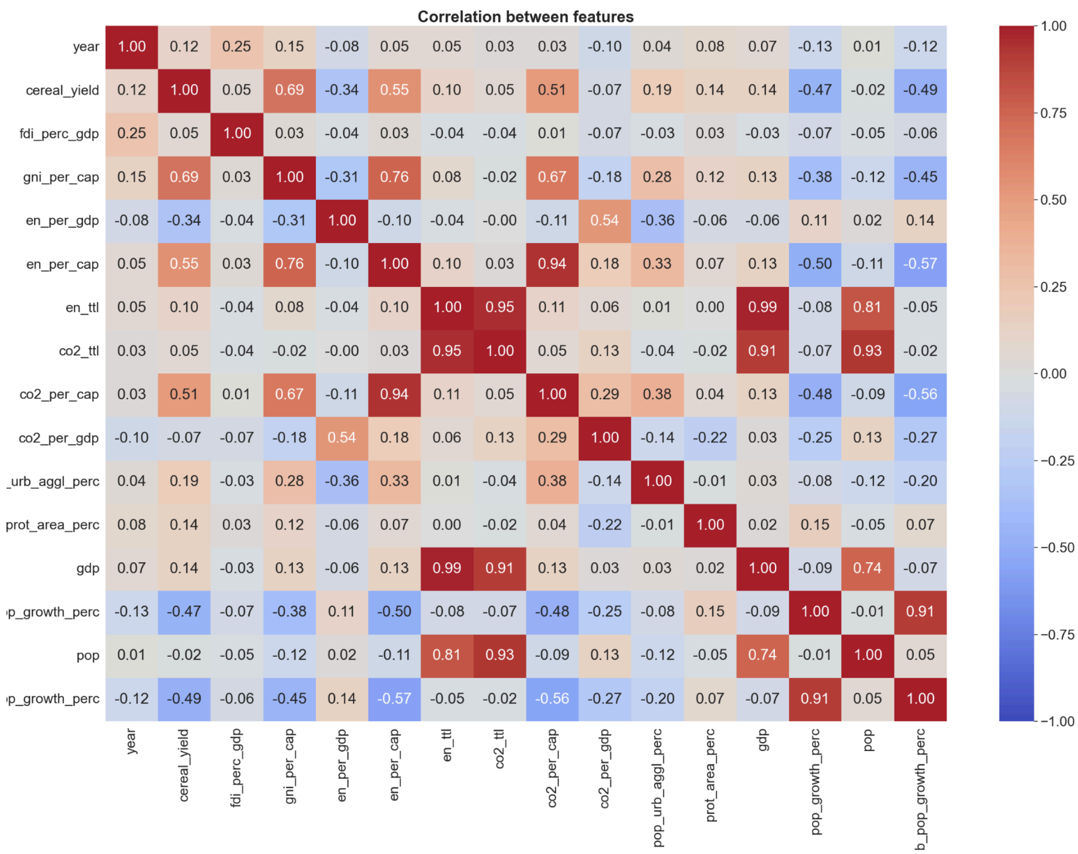
Existing research highlights the importance of machine learning in environmental monitoring. Random Forest algorithms are commonly used for emissions analysis, providing interpretable results on influential factors. Similarly, U-Net architectures have proven effective for satellite image segmentation, aiding deforestation monitoring. Lastly, decision tree models show potential in hydrological predictions by analyzing feature proximities related to rainfall and topography. This study builds on these foundations, integrating them into a comprehensive framework for climate impact analysis.

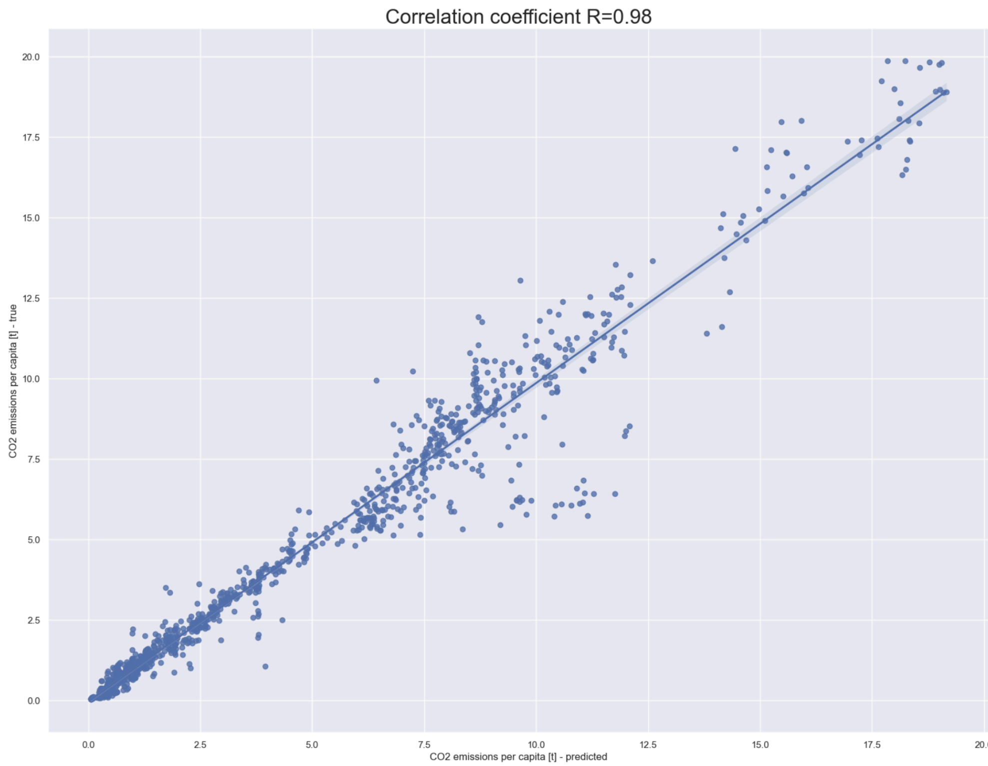
Methodology

Model Design and Implementation

Carbon emissions analysis:

- **Model**: Random Forest.
- **Features**: energy use per capita, cereal yield, urban population, and population growth.
- **Data Handling**: Normalization and dataset splitting (30% for training) to prevent overfitting.
- **Metrics**: Achieved an R2 score of 0.96.





The provided code leverages a machine learning pipeline for predicting CO2 emissions per capita based on a variety of socio-economic and environmental indicators. Below is a breakdown of its **algorithmic novelty**, **design novelty**, and **engineering novelty**:

Algorithmic Novelty:

1. **Recursive Feature Elimination (RFE)** with Cross-Validation (RFECV) is applied to identify the most relevant features. This recursive feature selection technique iteratively removes less important features, thereby refining the model and improving predictive accuracy. Using cross-validation within RFE ensures that the selected features generalize well across different subsets of the data.
2. **Random Forest with Hyperparameter Tuning**: The code performs hyperparameter tuning for the Random Forest model using a randomized search over a predefined parameter grid. By tuning hyperparameters like `n_estimators`, `max_features`, `max_depth`, `min_samples_split`, and `min_samples_leaf`, the algorithm can be optimized for better prediction accuracy and performance.

Design Novelty:

1. **Cross-Validation Strategy**: The code uses a two-level cross-validation strategy—one within RFECV for feature selection and another outside it for final model evaluation. This nested cross-validation approach ensures both the features and model are selected in a robust way, reducing the risk of overfitting.

2. **Pipeline Design for Reproducibility**: By setting a fixed random state (`random_state_num`), the code ensures reproducibility of results across all random operations. This is particularly useful in research or production environments where consistency is key.
3. **Outlier Removal**: The code removes outliers specific to a country ('ARE'), which suggests that the design allows for customized data preprocessing based on domain knowledge.

Engineering Novelty:

1. **Efficient Hyperparameter Search**: RandomizedSearchCV is used over GridSearchCV, which allows for efficient sampling of the parameter space and reduces computational costs by limiting the search to a subset of possible configurations.
2. **Parallel Processing**: The code uses `n_jobs=-1` in `RFECV` and `RandomizedSearchCV` to parallelize computations, which is an efficient way to speed up training and selection on systems with multiple cores.
3. **Custom Visualization**: The `seaborn.regplot` is used to visualize the correlation between predicted and true values, providing insights into the model's performance. The graph includes an overlayed correlation coefficient `R`, which provides immediate feedback on the strength of the relationship between predictions and true values.

In summary, the code combines feature selection, hyperparameter tuning, and cross-validation within an organized structure, demonstrating robustness, efficiency, and reproducibility in predicting CO_2 emissions per capita.

Deforestation Detection:

- **Model**: Custom U-Net for satellite imagery segmentation.
- **Data Handling**: ImageDataGenerator for augmentation, with checkpoint-based model saving.
- **Expected Outcome**: Detection of deforestation with accurate spatial segmentation.

Algorithmic Novelty

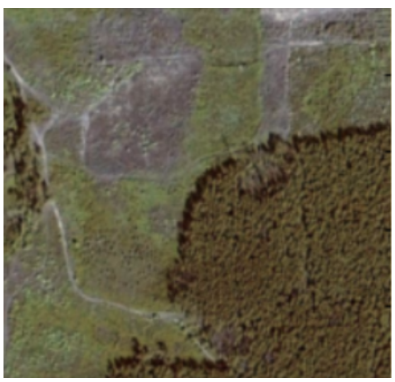
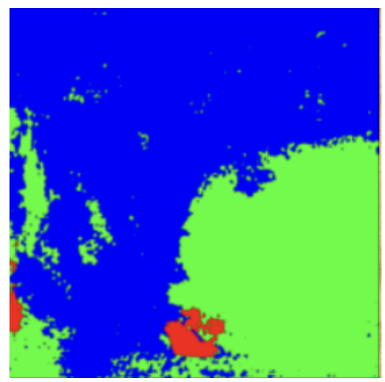
- **Customizable Architecture**: The U-Net variations enable flexible architecture adjustments (filter sizes, layer depth, batch normalization), allowing fine-tuning for segmentation tasks that vary in complexity.
- **Hybrid Blocks**: Using both convolutional and transposed convolutional layers enhances feature extraction and refinement during upsampling.

Design Novelty

- **Layered Skip Connections**: Enhanced skip connections between the encoder and decoder sections improve spatial and semantic consistency in segmented output, especially beneficial in high-resolution segmentation.
- **Adjustable Dropout**: Dropout layers (optional) help control overfitting in deeper U-Net layers, which is advantageous in smaller datasets or highly variable image segments.

Engineering Novelty

- **Resource Management**: The use of GPU-configured settings ensures efficient memory handling during training.
- **Flexible Training Pipeline**: Integration with TensorFlow's `ImageDataGenerator` allows for real-time data augmentation and balanced class distribution, key for segmentation models requiring diverse data points.

		
Satellite Image	Ground Truth	Predicted Image

Flood Prediction:

- **Model**: Decision Tree.
- **Features**: Rainfall, topographical data, and proximity to flood-prone regions.
- **Performance**: Achieved an accuracy score of 83% with reduced false positives.

Algorithmic Novelty:

- **Feature Labeling**: Efficient preprocessing method for converting categorical flood data into binary format, which simplifies further analysis or machine learning tasks.

Design Novelty:

- **Modular Preprocessing**: The design is modular, with preprocessing steps separated for clarity, allowing modifications and making it easier to scale to larger or more complex datasets.

Engineering Novelty:

- **Efficient Missing Data Handling**: Using lambda functions for checking and filling missing values streamlines preprocessing, improving processing speed and data consistency.

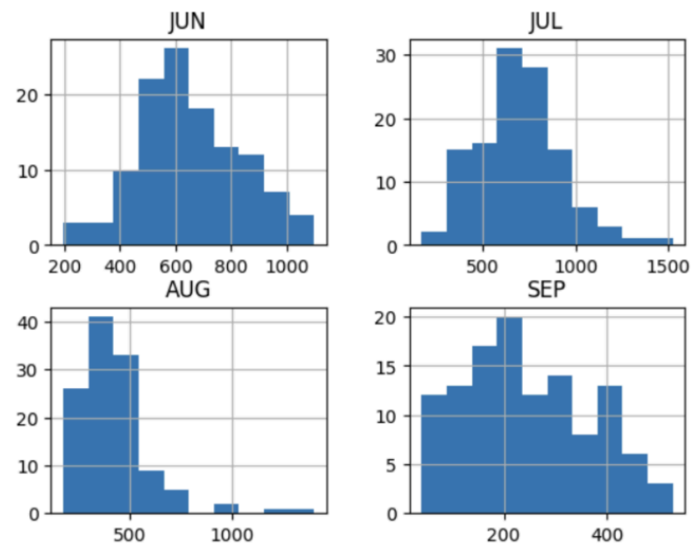


Fig: Shows the monthly data

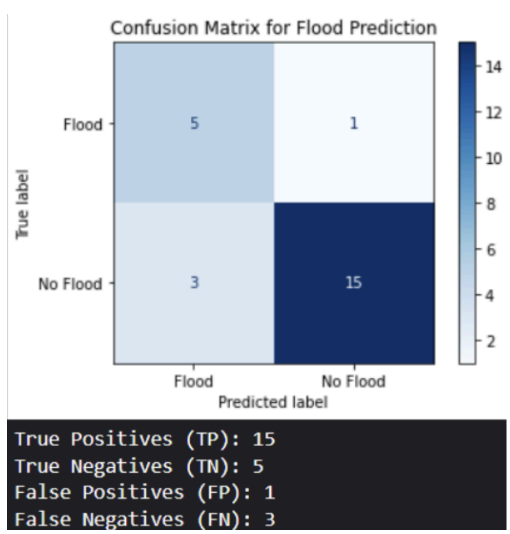


Fig: Model showing fewer false positives with an accuracy of 83%

Expected Outcome

The project is expected to provide:

- **Emission Predictions**: insight into factors affecting carbon emissions and potential future trends.
- **Deforestation Maps**: Visual and quantitative assessment of global deforestation areas.
- **Flood Prediction**: Early warning predictions with accuracy suitable for preliminary risk assessment.

References

1. https://www.flickr.com/photos/nasahubble/with/53867394329

2. P. Kadam and S. Vijayumar, "Prediction Model: CO2 Emission Using Machine Learning," 2018 3rd International Conference for Convergence in Technology (I2CT), Pune, India, 2018, pp. 1-3, doi: 10.1109/I2CT.2018.8529498. keywords: {Predictive models;Machine learning;Carbon dioxide;Data models;Training;Testing;Dairy products;GHG;CO2;prediction;RMSE;carbon emission;machine learning},

3. M. M. A. Syeed, M. Farzana, I. Namir, I. Ishrar, M. H. Nushra and T. Rahman, "Flood Prediction Using Machine Learning Models," 2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), Ankara, Turkey, 2022, pp. 1-6, doi: 10.1109/HORA55278.2022.9800023. keywords: {Temperature;Computational modeling;Support vector machine classification;Static VAr compensators;Machine learning;Predictive models;Floods;Binary Logistic Regression;Support Vector Classifier(SVC);K-Nearest Neighbor(KNN);Decision Tree Classifier(DTC);Flood Prediction;Rainfall},