**I-GUIDE MODEL CARD**

The I-GUIDE Model Card is an easy-to-use tool that will allow you to create documentation for each model that you create or use in a project.

Using this tool will help facilitate transparency and reproducibility about your project. It will also help you comply with relevant policies of journals, funding agencies, and universities.

The Model Card applies to:

* Pre-existing models acquired from other sources, e.g., produced by other researchers;
* Models you and your collaborators produced yourselves;
* Models you and your collaborators produced by integrating two or more other models (e.g., coupling).

**Model Card Attribution**

This Model Card template is an adapted version of the I-GUIDE Data Card template, which itself is based on Google’s *Data Cards Playbook* (https://pair-code.github.io/datacardsplaybook/).  
It has been restructured to address key considerations for geospatial model transparency, performance evaluation, and ethical deployment, in alignment with the I-GUIDE research lifecycle.

This work is shared under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).  
License details: <https://creativecommons.org/licenses/by-sa/4.0/>

A logo with text on it

AI-generated content may be incorrect.

* + 1. **BASIC INFORMATION**

|  |  |
| --- | --- |
| Model Card ID Number | MC-01 |
| Model Name | Machine learning models – Random Forest, XGBoost, Logistic Regression |
| Model Version | v.1 |
| Persistent Identifier |  |
| Outputs Supported |  |
| Model Card Author | IGUIDE- Team 1: Jennifer Marlon, Deepika Pingali, Surabhi Upadhyay, Emine Senkardesler, Pratyush Tripathy, Okikiola Michael Alegbeleye |

* + 1. **MODEL OVERVIEW**

|  |  |  |
| --- | --- | --- |
| Model Type | *(Select all that apply)*  ☑ AI model: Random Forest, XGBoost  ☑ Statistical model: Logistic Regression  ☐ Other: *(Specify)* | |
| Purposes | *(Select all that apply)*  ☑ Classification  ☑ Decision support  ☐ Forecasting  ☐ Regression  ☐ Simulation  ☐ Spatial analysis  ☐ Other: *(Specify)* | |
| Domains of Application | *(Select all that apply)*  ☑ Climate science  ☐ Economics  ☐ Environmental impact modeling  ☑ Geospatial analysis  ☐ Hydrology  ☐ Population modeling  ☑ Other social systems modeling: *(Specify)*  ☐ Other: *(Specify)* | |
| Model Authors and Developers | I-GUIDE Team 1 | |
| Source and Acquisition Method | ☐ Acquired (from external source)  ☑ Developed internally  ☐ Integrated from multiple models (e.g., coupled) | |
| User Licensing | ☑ Open source: *(Specify license type)*  ☐ Proprietary: *(Specify owner)*  ☐ Other restrictions on use: *(Specify restrictions)* | |
| Storage Location | ☑ Repository: *Github repo – RAP\_India*  ☐ Project-specific storage: *(Describe location)* |
| Access Control Policies | ☑ Open  ☐ Embargoed: *(Describe release timeline)*  ☐ Restricted: *(Describe access criteria)* |
| Use Case | The model estimates survey-based flood vulnerability at the district level by integrating demographic, environmental, and geospatial covariates. | |

* + 1. **MODEL INPUTS AND TRAINING DATA**

|  |  |
| --- | --- |
| Model Inputs | Individual-level survey responses (Age, Gender, Education, Caste)  District-level covariates:   * VIIRS Nighttime Radiance * ERA5 monthly precipitation and temperature * Web-scraped news media frequency (flood-related mentions) * Alpha Earth Embeddings |
| Input Data Types | *(Select all that apply)*  ☐ Raster  ☑ Tabular  ☐ Time Series  ☑ Vector  ☐ Other: *(Specify)* |
| Training Data Used | * Survey dataset on drought vulnerability (internal) * ERA5 Climate Reanalysis Data: <https://doi.org/10.24381/cds.adbb2d47> * VIIRS Nighttime Lights: <https://developers.google.com/earth-engine/datasets/catalog/NOAA_VIIRS_DNB_MONTHLY_V1_VCMCFG> * Alpha Earth Embeddings: <https://developers.google.com/earth-engine/datasets/catalog/GOOGLE_SATELLITE_EMBEDDING_V1_ANNUAL> |
| Training Dataset Representativeness | ERA5 and VIIRS data are global but may have biases at local scales. |

* + 1. **MODEL STRUCTURES**

|  |  |
| --- | --- |
| Feature Selection | Multiple models with different combinations of training data:   * Model 1: Demographics only * Model 2: Demographics + Physical features (climate, lights, news) * Model 3: Demographics + Alpha Earth embeddings * Model 4: Integrated model (all features) |
| Hyperparameters and Tuning | * **Random Forest:** n\_estimators=200, tuned max\_depth via grid search * **Gradient Boosting:** Tuned n\_estimators, learning\_rate, max\_depth via grid search * **Logistic Regression:** L1/L2 penalty, C tuned via grid search |
| Software and Dependencies | * Python * Libraries: scikit-learn, xgboost, pandas, numpy, geopandas |

* + 1. **MODEL PERFORMANCE AND EVALUATION**

|  |  |
| --- | --- |
| Validation Approach | ☒ Cross-validation  ☒Holdout set (20%)  ☐ Time series split  ☐ Other: *(Specify)* |
| Evaluation Results | * **Random Forest (Best: Model4):** Accuracy = 0.583, AUC = 0.61, F1 = 0.61 * **Gradient Boosting (Best: Model4):** Accuracy = 0.581, AUC = 0.61, F1 = 0.59 * **Logistic Regression (Best: Model2):** Accuracy = **0.583,** AUC = **0.62,** F1 = 0.57 |
| Testing or Validation Data Used | 20% holdout of survey dataset (stratified by district). |
| *(If model is integrated from multiple other models)*  Contribution of Constituent Models | *(Describe the contribution of each constituent model to the integrated model’s performance)* |

* + 1. **MODEL ADAPTATION AND CUSTOMIZATION (for acquired or integrated models only)**

|  |  |
| --- | --- |
| Source Models | *(List original models integrated into this one, with proper citations and links, where available)* |
| Availability of Source Model Code | *(Complete for each original model)*  ☐ Openly available: *(Include link)*  ☐ Restricted availability: *(Describe restrictions)*  ☐ Unavailable: *(Explain reason for unavailability)* |
| Modifications | *(Describe any changes made to an acquired model, and why these changes were made, or how multiple models were integrated)* |
| Training Data Adjustments | *(Describe new datasets used for fine-tuning or retraining, and provide DOI or URL where the dataset can be accessed. If dataset is not accessible, or only has restricted availability, state that here and describe restrictions)* |

* + 1. **MODEL DEPLOYMENT AND USAGE**

|  |  |
| --- | --- |
| Computational Requirements | *(List hardware/software requirements)* |
| Geospatial Considerations | *(Specify if the model includes spatial constraints, e.g., region-specific calibration)* |

* + 1. **TRANSPARENCY, EXPLAINABILITY, AND INTERPRETABILITY**

|  |  |
| --- | --- |
| Model Transparency | ☒ Fully transparent (rule-based, interpretable ML)  ☐ Partially transparent (some explainability features)  ☐ Black box (deep learning, complex ML models) |
| Explainability Features | ☒ Feature importance analysis  ☐ LIME  ☐ Sensitivity analysis  ☒ SHAP values  ☐ Other: *(Specify)* |
| Interpretability Challenges | Alpha Earth embeddings are latent features, making spatial contributions less interpretable. |
| Communication of Model Limitations | *(Describe how model limitations and uncertainty is communicated to prospective users of model impacts, or those potentially impacted by model impacts)* |

* + 1. **OTHER ETHICAL CONSIDERATIONS**

|  |  |
| --- | --- |
| Ethical Risks (Other Than Transparency, Explainability and Interpretability) | *(Select all that apply)*  ☐ Bias in training data: *(Specify)*  ☐ Intentional misuse risks: *(Specify)*  ☐ Privacy risks and surveillance: *(Specify)*  ☐ Security risks: *(Specify)*  ☐ Stigmatization of individuals or communities: *(Specify)*  ☐ Other: *(Specify)* |
| Measures Taken to Address Ethical Risks | *(Describe steps taken to mitigate ethical risks)* |
| Suitable Uses | Research, climate risk mapping, disaster preparedness. |
| Unsuitable Uses | Individual-level decision-making or resource allocation; application to non-Indian contexts without revalidation. |