Task A: Feasibility Prediction Model

1. Workflow

The notebook trains a binary classifier to predict whether a climate-resilience intervention will be delivered on time, given district characteristics, hazard data, and intervention parameters.

Step-by-step:

- Load data: We are using features from district profiles, hazard timeseries, project history and interventions catalog to predict the outcome variable
- Data exploration: Perform analysis of distributions, on-time rates by region/country, summary statistics and missing values to understand the input data better and pick features
- Feature engineering: Merge project history data with district profiles on district code and with interventions data on intervention_id to get district and intervention related features for each project. Also calculate prior-year hazard features by aggregating hazard timeseries data at district and year level and merging it with the combined features data on district coder and year. While joining project history data with hazards data, year from project history is matched with hazard data from previous year to avoid leakage. Also, encode categorical data(partner, region, hazard focus, country) to convert them to numeric values.
- Time-aware train/test split: Split the data for training and testing: 2021-2023 for training, 2024 for testing (prevents leakage)
- Train the model: We chose 3 models to train that are suitable for a classification problem– Gradient Boosting, Random Forest, Logistic Regression. Performed hyperparameter tuning via GridSearchCV (3-fold CV, ROC-AUC scoring) for three models.
- Model evaluation: Calculated ROC-AUC, PR-AUC, Brier score, accuracy for each model to find the best performing model. Best model by ROC-AUC is **Logistic Regression** (0.8331)

Model	ROC-AUC	PR-AUC	Brier	Accuracy
Random Forest	0.7983	0.6921	0.1034	0.8788
Gradient Boosting	0.8113	0.6807	0.1088	0.8636
Logistic Regression	0.8331	0.6795	0.1095	0.8636

- Error analysis: Performed error analysis by using accuracy by region, poverty quintile.
- Prediction: Generated delivery probability for every district x intervention pair using the best model, save to dist_int_delivery.csv to be used for task b.

2. Output

dist_int_delivery.csv

One row per district x intervention pair with columns:
district_code, district_name, region, country intervention_id, intervention_name, hazard_focus, cost, risk_reduction, impl_months, delivery_prob - predicted probability of on-time delivery (0 to 1), predicted_on_time - binary label (1 if delivery_prob >= 0.5)
This file is consumed by Task B for portfolio allocation.

3. Instructions

- Open task_a_feasibility_model.ipynb in Jupyter or Google Colab
- Ensure the dalberg_climate_case_data_clear_trends is saved as “data” under “dalberg_case_study” folder

- Run all cells sequentially (Cell > Run All)
- The output file dist_int_delivery.csv will be saved in the output folder
- Run this notebook BEFORE Task B, as Task B depends on dist_int_delivery.csv

4. Dependencies

- Python >= 3.8
- pandas, numpy
- scikit-learn (GradientBoostingClassifier, RandomForestClassifier, LogisticRegression, GridSearchCV, StandardScaler, LabelEncoder, metrics)
- matplotlib (for visualizations)

5. Assumptions

- The delivered_on_time label in project_history.csv is the ground truth for training.
- Prior-year hazard data is used (start_year - 1) to avoid using concurrent hazard information.
- For prediction on hypothetical district x intervention pairs, project-level features
- (baseline_risk_score, community_engagement_score) use district-level historical averages from past projects, with global mean as fallback for districts with no history.
- Missing feature values are filled with column medians (training set).
- The best model is selected by ROC-AUC on the 2024 test set.
- Categorical encodings (LabelEncoder) are fit on training data and applied consistently.