THE UNIVERSITY OF CHICAGO

DATA SCIENCE INSTITUTE

# Data Science Clinic

**with University of Rwanda**

Rachelle Cho, Jenny Li, Sana Fessuh, Grace Rowan

# RESEARCH QUESTION

Predict **landslide risks** of the **Gitwe Kadhua Corridor** by…

- Utilizing available geo topical factors
- Implementing various models
- Determining best performing model

# WE CARE BECAUSE…

- …climate fluctuations will lead to **increases** in landslides
- …landslides lead to **significant fatalities** and **irreversible damage…**
  - Substantial loss of life
  - Billions of dollars in property damage
- …limited research thus far

# RELEVANT BACKGROUND — PRIOR RESEARCH

**“Landslide susceptibility and influencing factors analysis in Rwanda” by Mind’je, R., L., Nsengiyumva, J.B. et al. (2020)**
**“Landslide Susceptibility Assessment Using Spatial Multi-Criteria Evaluation Model in Rwanda” by Nsengiyumva, J. B., Luo, G., Nahayo, L. et al. (2018)**

## CLAIMS

- West, North, and South provinces show high susceptibility to landslides
- Key causal factors:
  - Steep slopes
  - High elevation
  - Heavy rainfall

# RELEVANT BACKGROUND — LOCATION

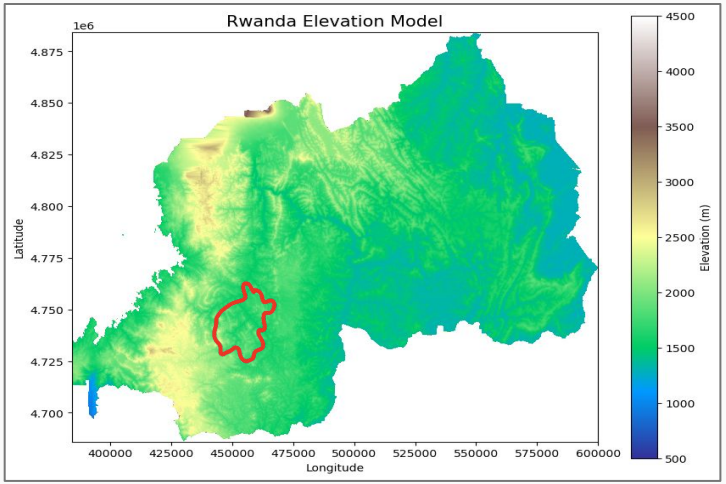
**“Landslide susceptibility and influencing factors analysis in Rwanda” by Mind’je, R., L., Nsengiyumva, J.B. et al. (2020)**

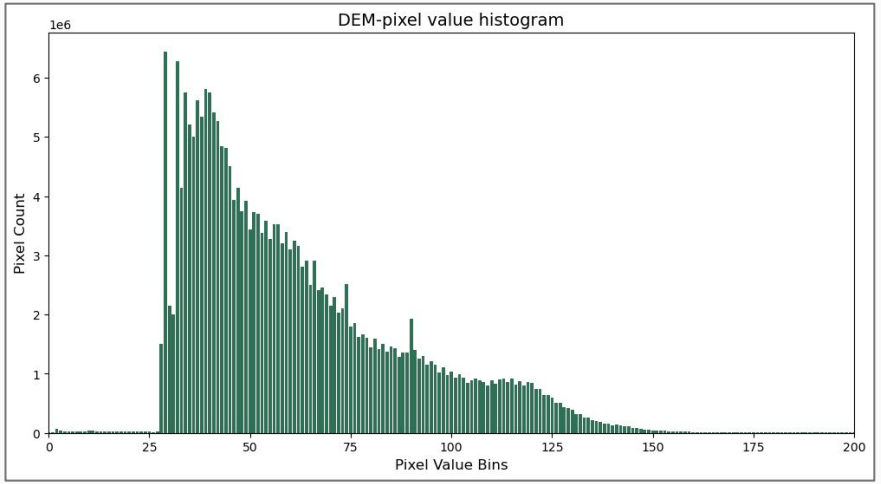
## CLAIMS

- 10,169 sq mile landlocked country in Central Africa, located in the Great Lake region
  - Region highly susceptible to landslides
- The Gitwe-Kadhua Corridor is an region of interest due to high risk levels

# OUR DATA — DEM

- 10m raster resolution
  → Pixel-based elevation data

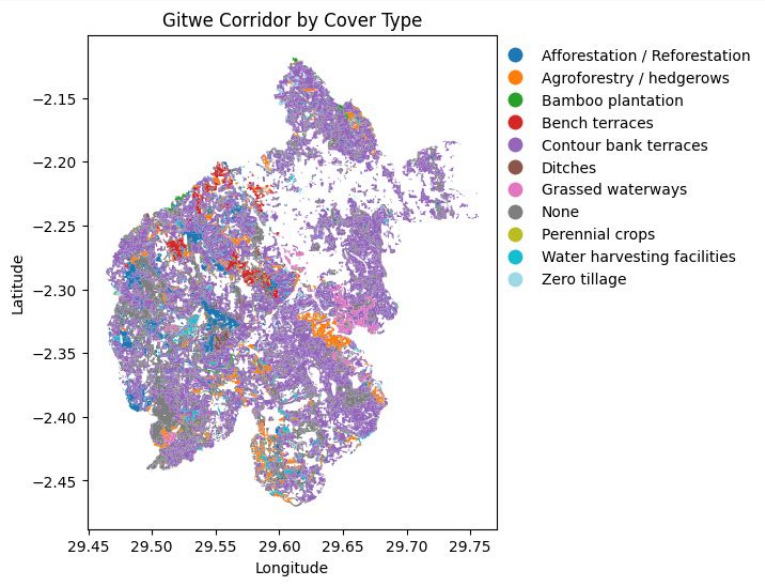


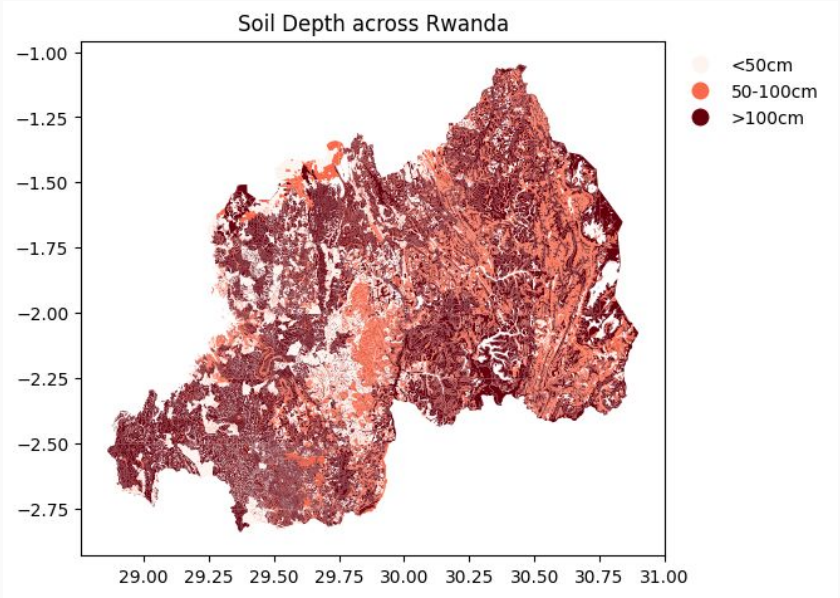




- Peaks in the histogram
  → Frequent & dominant elevation ranges
- Flat sections in the histogram
  → Less terrain variation (e.g., plateaus or flatlands)

# OUR DATA — GEOFILE EXPLORATION

- 6 GeoPackage files describing varying properties in Rwanda
- Range widely in coverage and size



# DATA CLEANING & PROCESSING

Challenge

- Direct merging is not possible
  - Each GeoPackage contains a unique set of polygons

Resolution

- Create a hexagonal grid
- Assigned attributes based on proximity to hexagon centers

**Input**

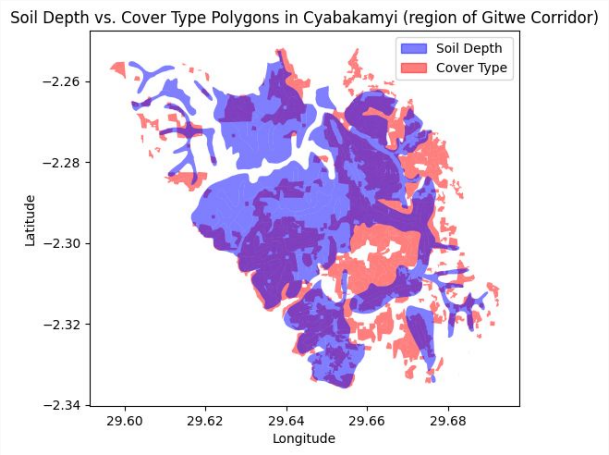
**6 GeoPackages**

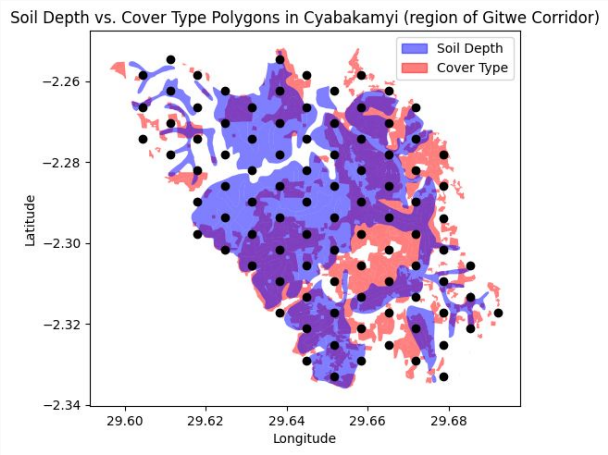
**Output**

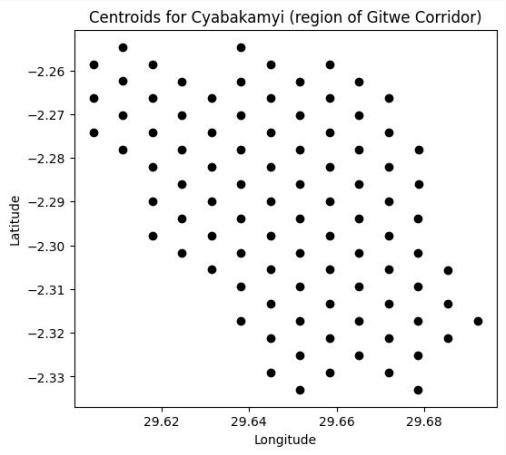
**7 predictors**: includes soil depth and type of land coverage

**1 target variable**: landslide risk

# DATA CLEANING & PROCESSING — VISUAL



# DELIVERABLES

- **White Paper**

- **Four Different Models:**
  - Ordered Linear Model
  - Random Forest Model
  - Neural Network Model - EDLT
  - Large Language Model - DistilBERT


# MODEL OVERVIEW

- **Model Characteristics**:
  - **Data split:** 60% training, 20% testing, 20% validation
  - **Features**: type of land coverage, soil class, soil depth, riverside, roadside, area of land coverage, land coverage density

| “Moderate” | “High” | “Very High” | “Extremely High” |
|---|---|---|---|

## deliverables:

# Ordered Linear Model

- **Ordered Categories:**
  - More nuanced interpretation of relationship between features and risk
- **Handling Class Imbalance:**
  - SMOTE oversampling technique

| ***method=newton*** | |
|---|---|
| Validation Accuracy | 43.3% |
| Test Accuracy | 41.4% |
| Test Overprediction Rate | 6.3% |
| Test Underprediction Rate | 52.2% |

## deliverables:

# Random Forest Model

- **Feature importance:** removed predictors with low impact
- **Handling class imbalance:**
  - SMOTE
  - Balanced Random Forest
  - Class_weights

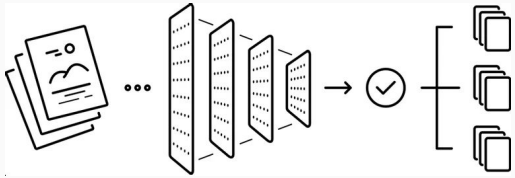
| ***Max_depth=8***<br>***Min_samples_leaf=3***<br>***Min_samples_split=10***<br>***N_estimator =50*** | |
|---|---|
| Validation Accuracy | 52.8% |
| Test Accuracy | 51.6% |
| Test Overprediction Rate | 15.5% |
| Test Underprediction Rate | 33.2% |

## deliverables:

# Neural Network Model - Convolutional Neural Network for Categorical Data (EDLT)

- **Data Processing and Learning Process**
  - Converts categorical data into numerical
  - Reorders features to maximize correlation
  - Detects relationships between features

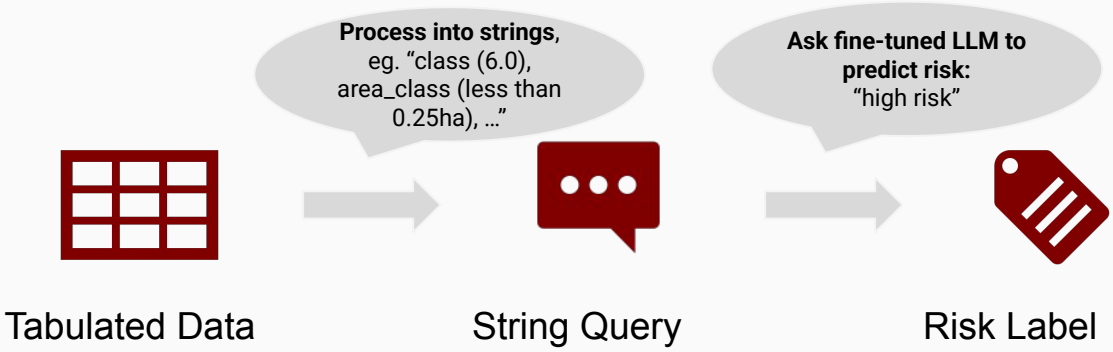
| | |
|---|---|
| Validation Accuracy | 48.1% |
| Test Accuracy | 51.0% |
| Test Overprediction Rate | 16.5% |
| Test Underprediction Rate | 32.5% |



## deliverables:

# Large Language Model

- **Model:** DistilBert
- **Data Processing**




| *Weight Decay: 2e–2*<br>*Learning rate: 5e–5* | |
|---|---|
| Validation Accuracy | 50.3% |
| Test Accuracy | 50.7% |
| Test Overprediction Rate | 17.2% |
| Test Underprediction Rate | 32.2% |

# CONCLUSIONS

- Random Forest had the highest accuracy among models
- Model Accuracy has room for improvement
  - **Feature gaps** impact performance more than model choice
- **Underprediction** > Overprediction
- Apply corridor findings across Rwanda

# Thank You!

# SOURCES


- Nsengiyumva, J. B., Luo, G., Nahayo, L., Huang, X., & Cai, P. (2018). Landslide Susceptibility Assessment Using Spatial Multi-Criteria Evaluation Model in Rwanda. International Journal of Environmental Research and Public Health, 15(2), 243. https://doi.org/10.3390/ijerph15020243