Natural Disaster Prediction Model: PREDINA

Nouf AlMansour, Sara AlOqiel, Shahad AlMutairi

Supervised by: Dr. Mashael Aldayel

Introduction

Natural disasters have long posed significant challenges to communities worldwide, impacting lives, economies, and infrastructure. This project, conducted as part of the IT362 - Principles of Data Science course, explores natural disaster patterns, financial impacts, and trends from 1900 to 2021. By analyzing historical data, we aim to uncover key insights that can enhance disaster preparedness and response strategies. While disasters themselves may not always be preventable, data-driven early warning systems and proactive planning can significantly reduce their devastating effects. Through this study, we seek to deepen the understanding of natural disasters and contribute to building more resilient societies.

Objectives

1. What factors influence the intensity of natural disasters?
2. Which regions are most commonly affected by specific types of natural disasters?
3. What are the estimated losses associated with natural disasters based on current data?
4. Which regions have a greater chance in expecting earthquakes?

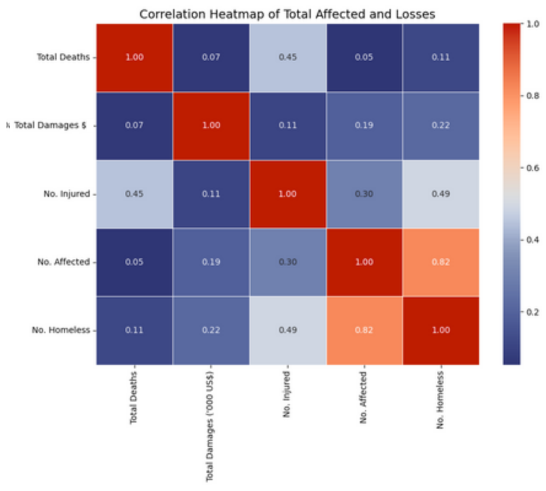
Data Collection

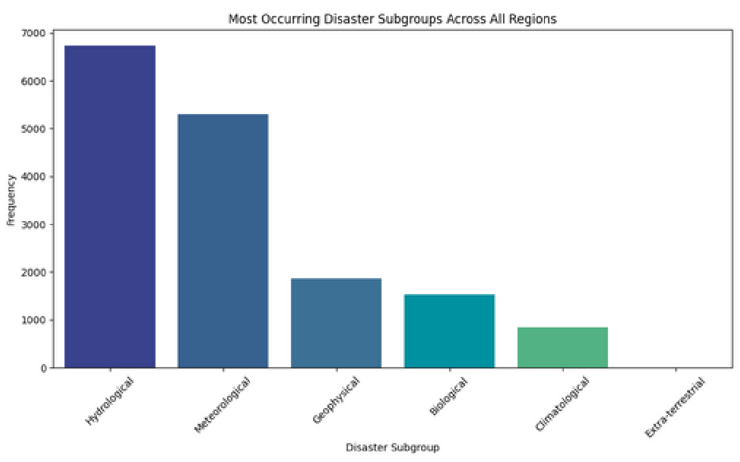
1. EM-DAT: The International Disaster Database
2. Kaggle Dataset : ALL NATURAL DISASTERS 1900-2021 / EOSDIS
3. Global Disaster Alert and Coordination System (GDACS)

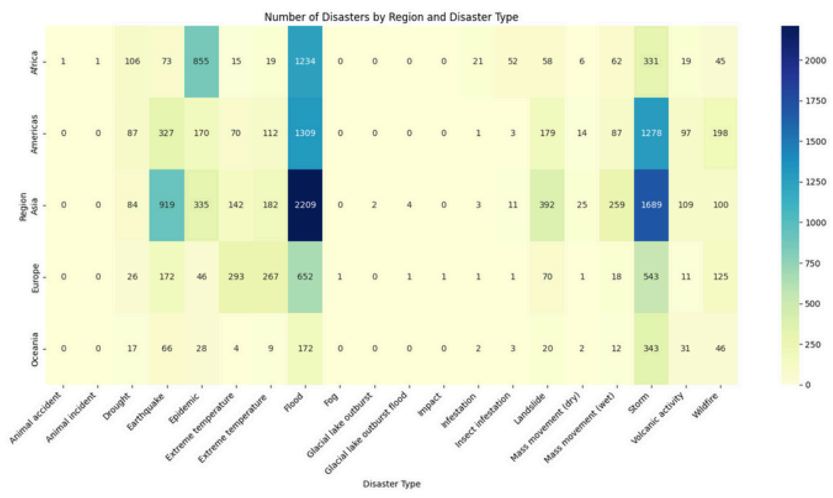
Data Analysis

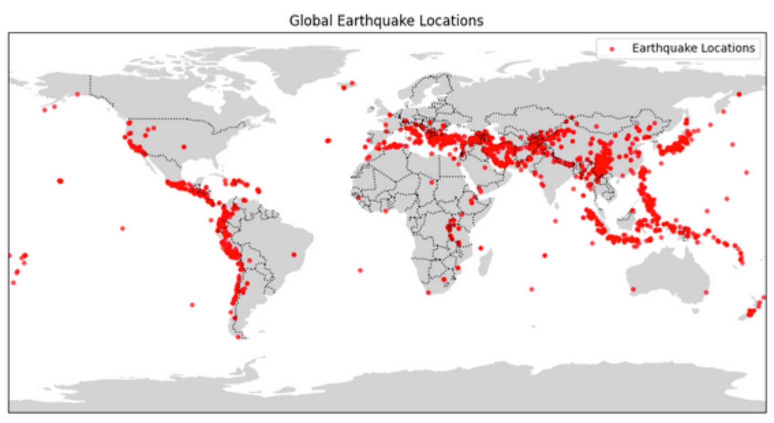
To better understand the disaster data, we explored different parts of the dataset. Here's what we found:

- Floods and storms (hydrological and meteorological disasters) happen the most around the world.
- There's a strong connection between the number of people affected by disasters and the number who end up homeless, showing that losing homes is a big problem during disasters.
- Pakistan and Iran have the highest number of disaster-related deaths, meaning these countries have been hit especially hard.
- Earthquakes happen all over the world, but most of them are near the edges of tectonic plates, especially around the Pacific Ocean and parts of Asia and the Middle East.









Models and Findings

To achieve our goal of accurately predicting the risk value, we implemented several regression models, including:

- Linear Regression
- Random Forest Regression
- Gradient Boosting Regression

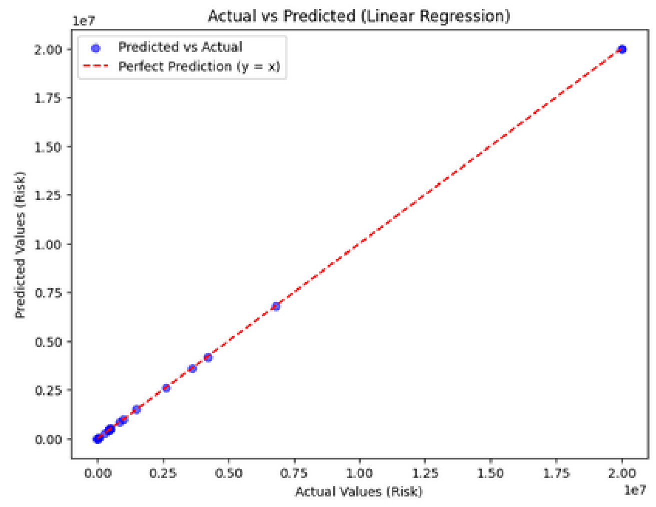
These models were selected to capture both simple and complex patterns within the data.

For evaluation, we used performance metrics such as:

- Root Mean Squared Error (RMSE)
- R-squared (R^2) Score

After training and evaluating the models, we found that Gradient Boosting Regression delivered the most accurate results.

It provided the best data fit, making it the most suitable choice for handling the complexity of our dataset.



Conclusion

In conclusion, this study focused on historical risk data, where we applied thorough data preprocessing, cleaning, and exploratory data analysis to maintain the quality and consistency of our dataset.

Through building and evaluating multiple regression models, we identified Gradient Boosting Regression as the most effective method for predicting risk values, delivering the highest accuracy among the models tested.

Moving forward, we aim to further refine our predictions by integrating additional data features and exploring alternative modeling approaches. These future enhancements are expected to strengthen the overall performance and reliability of the risk prediction system.