College of Computer and Information Sciences
Department of Information Technology

IT 362 Course Project
Semester-2, 1446H

IT362: Principles of Data Science

< PREDINA >

Project Report: Natural Disaster Impact Prediction

Prepared by

Student name	Student ID
Nouf AlMansour	444200525
Sara AlOqiel	444203016
Shahad AlMutairi	444200935

Supervised by:
Dr. Mashael AlDayel

1 Introduction

Natural disasters cause significant damage to life and property. Leveraging historical data to predict disaster impacts can significantly improve preparedness and response strategies. This project aims to enhance understanding and forecasting of natural disaster impacts, focusing on regional patterns, financial consequences, and predictive modeling. The study incorporates data from multiple sources and addresses the following four research questions.

Research Questions:

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

Our project integrates three datasets—EM-DAT, EOSDIS, and GDACS—to analyze and predict the impacts of natural disasters using predictive models.

2 Phase 1: Data Collection

Datasets Integrated:

- **EM-DAT (CRED)**: The International Disaster Database, which tracks natural and technological disasters globally, provides data on fatalities, affected populations, and economic damages.

- **EOSDIS (NASA)**: A dataset of natural disasters from 1900 to 2021, sourced from NASA's Earth Observing System Data and Information System (EOSDIS). This dataset includes disaster type, location, dates, and impacts like fatalities and economic damages.

- **GDACS**: The Global Disaster Alert and Coordination System, which provides real-time data on natural disasters like earthquakes, tsunamis, and storms.

Data Preprocessing:

- **Data Transformation:** Raw data from EM-DAT and EOSDIS was cleaned and converted into a format suitable for analysis (e.g., CSV).

- **Feature Extraction**: Important features like **Disaster Type, Magnitude, Country, ISO Code**, and **Geographical Coordinates** were extracted for the analysis. Missing or incomplete data (e.g., Losses in GDACS data) was handled by exclusion or imputation.

3 Phase 2: Data Processing, Cleaning, and Exploratory Data Analysis (EDA)

Key Objectives:

- **Data Inspection**: The dataset was examined for completeness, missing values, and inconsistencies.
- **Exploratory Data Analysis:** Trends, patterns, and anomalies were identified through summary statistics and visualizations.
- **Data Cleaning:** Missing values were handled by removing rows with missing critical data (e.g., Losses column in GDACS), and data types were standardized for consistency.

Findings from EDA:

- **Disaster Types**: Earthquakes and wildfires were found to be the most frequent disaster types, with significant geographical patterns.
    - **Earthquakes** were most prevalent in **Asia**.
    - **Wildfires** occurred frequently in **Oceania** and **South America**.
- **Countries Most Affected:** Greece, Australia, and Argentina were identified as the top countries affected by disasters.
- **Missing Data**: Columns like Event Name and Losses in the primary dataset had high missingness, requiring imputation or removal for analysis.

Visualizations:

- **Disaster Frequency by Country**: A bar chart showed the top 10 countries with the most reported disasters.
- **Disaster Type Distribution:** A count plot displayed the distribution of different disaster types across regions, highlighting the dominance of earthquakes and wildfires.

4 Phase 3: Modeling and Communication

Approach:

This phase focused on applying appropriate models to predict disaster impacts, specifically:

- **RQ1:** Predicting Total Affected and Total Damages based on earthquake magnitude.
- **RQ2:** Assessing the predictive power of geographical factors (e.g., region) and economic development level.

Model Selection:

- **Linear Regression:** A baseline model to predict disaster impacts.
- **Random Forest Regressor:** A tree-based model to capture non-linear relationships.
- **Gradient Boosting Regressor:** An ensemble model to improve prediction accuracy.

Feature Engineering:

- **Magnitude** was the primary feature used for prediction, with additional features like Development Level and Region tested for improvement.

Data Splitting:

The data was split into training and testing sets (80% for training and 20% for testing) to evaluate the model performance.

Model Evaluation:

- **Reaearch Question 1**: Predicting Total Affected and Total Damages:
    - **Linear Regression** had limited predictive power with low R^2 scores for both targets.
    - **Random Forest** and **Gradient Boosting** showed better results but were still limited in performance, especially for Total Damages.

    Key Results:

    - **Total Affected**:
        - Linear Regression R^2: 0.1668
        - Random Forest R^2: 0.2683
        - Gradient Boosting R^2: 0.2444
    - **Total Damages**:
        - Linear Regression R^2: 0.0278
        - Random Forest R^2: -0.0430
        - Gradient Boosting R^2: -0.0548

- **Reaearch Question 2**: Adding **Region** and **Development Level** improved the prediction for Total Affected:
    - **Random Forest** and **Gradient Boosting** models performed better with the addition of Region and Development Level.

    Key Results for Total Affected (Magnitude + Region):

    - **Linear Regression R^2**: 0.1112
    - **Random Forest R^2**: 0.4856
    - **Gradient Boosting R^2**: 0.4794

Feature Importance:

- For **Random Forest**, both **Magnitude** and **Region** were identified as the most important features for predicting Total Affected.

5 Conclusions

This project integrated multiple datasets to predict the impacts of natural disasters, with a specific focus on earthquakes. The models demonstrated that:

- **Magnitude** is a significant predictor of disaster impact, but additional features like **Region** and **Development Level** improve model accuracy.
- **Gradient Boosting** was the most effective model for predicting the Total Affected population, while the prediction of Total Damages was less accurate.

Key Insights:

- **Magnitude** alone has limited predictive power, and incorporating **geographical features** (e.g., Region) significantly enhances predictions.
- The integration of **Development Level** also adds value but requires further refinement for better performance.

Future Work:

- **Feature Expansion**: Adding features such as earthquake depth and socioeconomic factors could improve model accuracy.
- **Real-time Prediction**: Implementing the models in a real-time prediction system could provide valuable insights for disaster management.