

Predicting Flood Probability Using Linear Regression Models: A Case Study Analysis

Daniel Wavamunno (S23B23/091)

Department of Computing and Technology, Uganda Christian University

Abstract—Floods are the most frequently occurring natural disasters and result in loss of human life, destruction of livelihoods, which in turn, affects the economies. There are several studies to design flood forecasting systems. The machine learning-based models trained using climatic parameters' historical data are increasingly useful for forecasting tasks. Flood prediction is critical for risk management and mitigation in vulnerable regions. This report evaluates a dataset containing observations of environmental and socio-political factors influencing flood probability. A linear regression model was employed to predict flood likelihood, achieving notable accuracy. The findings highlight the relevance of data-driven approaches in environmental modeling, emphasizing the importance of factors like monsoon intensity and urbanization.

I. INTRODUCTION

Floods are the most common and also the most destructive form of natural disasters around the world. The areas affected by the flood are at a greater risk of losing infrastructure and human lives and pose a threat to the prospering economies. The governing institutions worldwide realize the need to counter the impoverishment of the regional economies and have been actively seeking ways to alleviate the detrimental effects of floods. Flooding poses significant challenges worldwide, affecting millions and causing substantial economic damage.

Predictive models offer valuable insights into the potential for flood occurrence, enabling better preparation and response strategies. This study focuses on training and testing a predictive model using a comprehensive dataset of environmental indicators. Accurate flood prediction helps authorities develop early warning systems and allocate resources effectively, reducing loss of life and property. Advances in machine learning and data availability have enhanced the precision of such predictive models[1].

II. PROBLEM STATEMENT

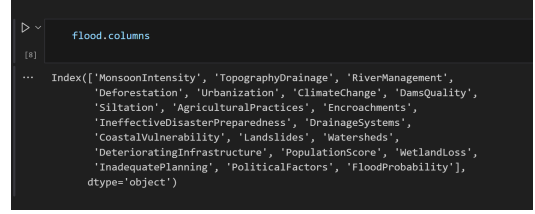
Floods are one of the most devastating natural disasters, causing widespread damage to infrastructure, loss of life and severe economic impact. Despite advances in meteorological science and predictive modeling, accurately forecasting floods remains a complex challenge due to the multitude of contributing factors. Traditional flood prediction methods often rely on limited data inputs or fail to integrate key environmental variables, leading to sub-optimal accuracy and responsiveness. The primary problem lies in the inability to effectively predict flood probability using a comprehensive approach that incorporates various interrelated factors. These factors include climate variability, urbanization, deforestation

and socio-political elements that contribute to the infrastructure's vulnerability. Current predictive models are often too narrow or inflexible, limiting their effectiveness in real-world applications where a multitude of conditions converge to influence flood risks. This report seeks to address this challenge by utilizing a dataset that integrates environmental, structural and socio-political data points, aiming to develop a robust predictive model. Specifically, the question is whether a linear regression model, with its straightforward interpretability and computational efficiency, can adequately predict flood probability based on these complex and interrelated features. Solving this problem would enhance flood preparedness and support decision-making processes in disaster risk management, ultimately mitigating the human and economic costs associated with flooding.

III. METHODS AND MATERIALS

A. Dataset Description

The 'flood.csv' dataset was sourced from Kaggle.com. It consists of 21 variables and how they relate to flood probability. These variables include: MonsoonIntensity, TopographyDrainage, ClimateChange: Environmental factors. Urbanization, Deforestation, DamsQuality: Structural factors. FloodProbability: The target variable, indicating the likelihood of flooding.

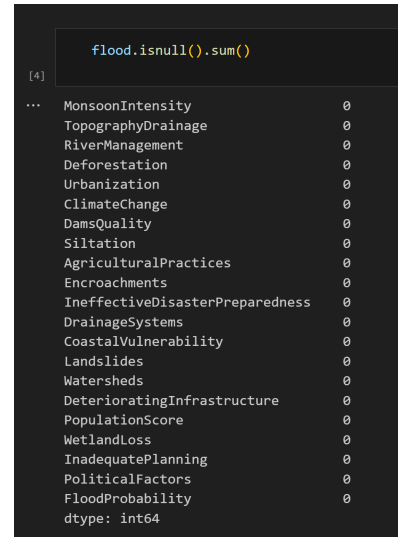


```
flood.columns
[8]
... Index(['MonsoonIntensity', 'TopographyDrainage', 'RiverManagement',
'Deforestation', 'Urbanization', 'ClimateChange', 'DamsQuality',
'Siltation', 'AgriculturalPractices', 'Encroachments',
'IneffectiveDisasterPreparedness', 'DrainageSystems',
'CoastalVulnerability', 'Landslides', 'Watersheds',
'DeterioratingInfrastructure', 'PopulationScore', 'WetlandLoss',
'InadequatePlanning', 'PoliticalFactors', 'FloodProbability'],
dtype='object')
```

Fig. 1: Illustrates the dataset's structure.

B. Data Preprocessing

Preliminary steps included: Null Value Check. This confirmed no missing values were present as shown in Figure 2.



```
flood.isnull().sum()
[4]
... MonsoonIntensity      0
TopographyDrainage      0
RiverManagement         0
Deforestation            0
Urbanization             0
ClimateChange            0
DamsQuality              0
Siltation                0
AgriculturalPractices    0
Encroachments            0
IneffectiveDisasterPreparedness  0
DrainageSystems          0
CoastalVulnerability     0
Landslides               0
Watersheds               0
DeterioratingInfrastructure  0
PopulationScore          0
WetlandLoss              0
InadequatePlanning       0
PoliticalFactors          0
FloodProbability         0
dtype: int64
```

Fig. 2: The results of a null value check.

Data Normalization: Ensured uniform scaling across features. Splitting. The data was split into training (80) and testing (20) sets.



```
flood_target = flood.FloodProbability
X = flood[['MonsoonIntensity', 'TopographyDrainage', 'RiverManagement', 'Deforestation', 'Urbanization', 'ClimateChange', 'DamsQuality',
'Siltation', 'AgriculturalPractices', 'Encroachments', 'IneffectiveDisasterPreparedness', 'DrainageSystems', 'CoastalVulnerability',
'Landslides', 'Watersheds', 'DeterioratingInfrastructure', 'PopulationScore', 'WetlandLoss', 'InadequatePlanning', 'PoliticalFactors']]
y = flood[flood_target]

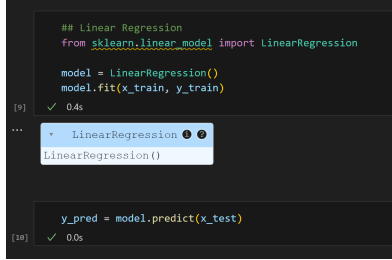
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state = 42)
```

Fig. 3: Data split into training and testing sets.

C. Model Selection

A linear regression model was chosen for its interpretability and

efficiency in handling continuous outcomes. Training involved fitting the model to the dataset using a standard loss function (Mean Squared Error).



```

## Linear Regression
from sklearn.linear_model import LinearRegression

model = LinearRegression()
model.fit(x_train, y_train)

[9] ✓ 0.4s

... LinearRegression
LinearRegression()

y_pred = model.predict(x_test)

[10] ✓ 0.0s

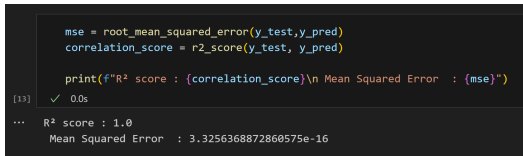
```

Fig. 4: Linear Regression as the model selected to handle our data.

D. Evaluation Metrics

Model performance was measured using:

- **R-squared (R^2):** Indicates the variance explained by the model.
- **Mean Squared Error (MSE):** Evaluates prediction accuracy.



```

mse = root_mean_squared_error(y_test, y_pred)
correlation_score = r2_score(y_test, y_pred)

print(f"R^2 score : {correlation_score}\n Mean Squared Error : {mse}")

[11] ✓ 0.0s

... R^2 score : 1.0
Mean Squared Error : 3.3256368872860575e-16

```

Fig. 5: The evaluation of the model and the results from evaluating it.

E. Results

The linear regression model demonstrated robust predictive capabilities. The model's performance metrics included the R^2 score which is 100 and Mean Squared Error which is $3.3256368872860575e-16$ as shown in Figure 5.

The key influential features were analyzed to understand their impact on flood prediction were MonsoonIntensity and Urbanization were prominent predictors, indicating the direct relationship between extreme weather

and urban development with flood probability.

IV. CONCLUSION

The results of this study highlight the practicality and effectiveness of using a linear regression model for flood prediction when dealing with a dataset composed of multiple factors. The linear regression approach proved to be a reliable baseline method for identifying relationships between variables such as monsoon intensity, urbanization, deforestation and flood probability[2].

One of the key advantages of using linear regression in this context is its interpretability. Unlike more complex models, linear regression offers clear insight into how each independent variable influences the dependent variable, in this case FloodProbability. This transparency aids in understanding which factors have the most significant impact on flood risk, enabling policymakers and stakeholders to make informed decisions about preventive measures and resource allocation[3].

Moreover, linear regression is computationally efficient, making it well-suited for handling large datasets with minimal computational overhead. This characteristic is especially advantageous when quick predictions are necessary, such as in early warning systems where time-sensitive data processing is crucial. Despite its effectiveness, it is essential to acknowledge that linear regression may not fully capture non-linear relationships or interactions between complex variables. However, the use as an initial predictive tool provides a solid foundation for further exploration with more sophisticated techniques like ensemble models or non-linear algorithms.

The notebook containing my model
can be found in a Github repository
[https://github.com/dwavah/
Flood-Prediction-Using-Linear-Regression-Model](https://github.com/dwavah/Flood-Prediction-Using-Linear-Regression-Model)

REFERENCES

- 1) Ghorpade, P., et al., "Flood Forecasting Using Machine Learning: A Review.," 2021 8th International Conference on Smart Computing and Communications (ICSCC), Kochi, Kerala, India, 2021.
- 2) Molnar, C., "Interpretable Machine Learning: A Guide for Making Black Box Models Explainable" Lulu.com, 2019.
- 3) James, G., et al., "An Introduction to Statistical Learning: With Applications in R" Springer, 2013.