UNIVERSITY OF WOLVERHAMPTON

HERALD COLLEGE KATHMANDU | ing GROUP

5CS037

Concepts and Technologies of AI

Evaluation of Classification Models for categorizing landslide size

Author

Narayan Lohani

2408869, L5CG15

Supervisor

Siman Giri

Module Leader

Table of Contents

Table of Figures

Abstraction

The objective of this report is to assign landslide size. For this the dataset chosen is “Global Landslide Catalog Export” which can be found in NASA's Open Data Portal Its key features include target variable size of the landslide, distance from gazetteer, latitude, longitude, amount of population residing.

Foremost, to understand the data, Exploratory Data Analysis was performed with help of histograms and bar charts.

Decision Tree Classifier and Random Forest Classifier were used as classification models imported from scikit-learn library. Optimizing the hyper-parameters and selecting features with most impact on performance were performed to check if they improve their performance. They were evaluated using Accuracy, Recall, Precision and f1 scores on which Random Forest did better than Decision Tree with 0.6955 accuracy vs 0.6357 accuracy.

Random Forest Classifier performed better than Decision Tree Classifier due to ensemble learning and reducing overfitting, while Decision Tree struggled with generalization, leading to low accuracy. Feature selection had minimal impact, suggesting further optimization is required.

Introduction

The goal of the project is to classify landslide size-small, medium, large, and large based on their environmental factors.

The dataset used on for this project is “Global Landslide Catalogue Export” which can be found in NASA’s Open Data portal. It consists of 11,034 rows and 31 columns, providing information on key environmental factors influencing landslide size. The independent variables include landslide category (e.g., landslide, mudslide, rockfall), landslide trigger (e.g., rain, downpour), landslide setting (natural slope, above road, below road), and location accuracy (latitude and longitude). The target variable is landslide size, categorized into small, medium, and large. Preprocessing challenges involve handling missing values in columns like landslide size admin_population_division. This analysis aligns with **United Nations Sustainable Development Goals (UNSDG) 11: Sustainable Cities and Communities** by contributing to disaster preparedness and risk management. By accurately classifying landslide sizes, policymakers and urban planners can identify risk factors, mitigate hazards, and safeguard infrastructure in vulnerable areas. This promotes safer and more sustainable living environments, especially in landslide-prone regions.

The objective of this analysis is to train a predictive classification model which can estimate the landslide size based on the features mentioned above in the dataset.

Methodology

Data Preprocessing

The data set obtained was substandard quality. Substantial proportion of data was missing value. So before building the classification model, the dataset had to be preprocessed with several steps. The number of columns was also reduced as most of them were irrelevant and few had significant missing values. The location accuracy column was processed to remove the unit and keep the integer value instead. The dataset after preprocess had six features, location accuracy, admin division population, gazetteer distance, latitude, and longitude. The target variable, landslide size, was encoded for model compatibility. Since all numerical features were continuous, no additional scaling was required.

Exploratory Data Analysis (EDA)

To understand the relationship between features and target variables different plots were made, which are summarized below.

1. Bar chart of landslide sizes

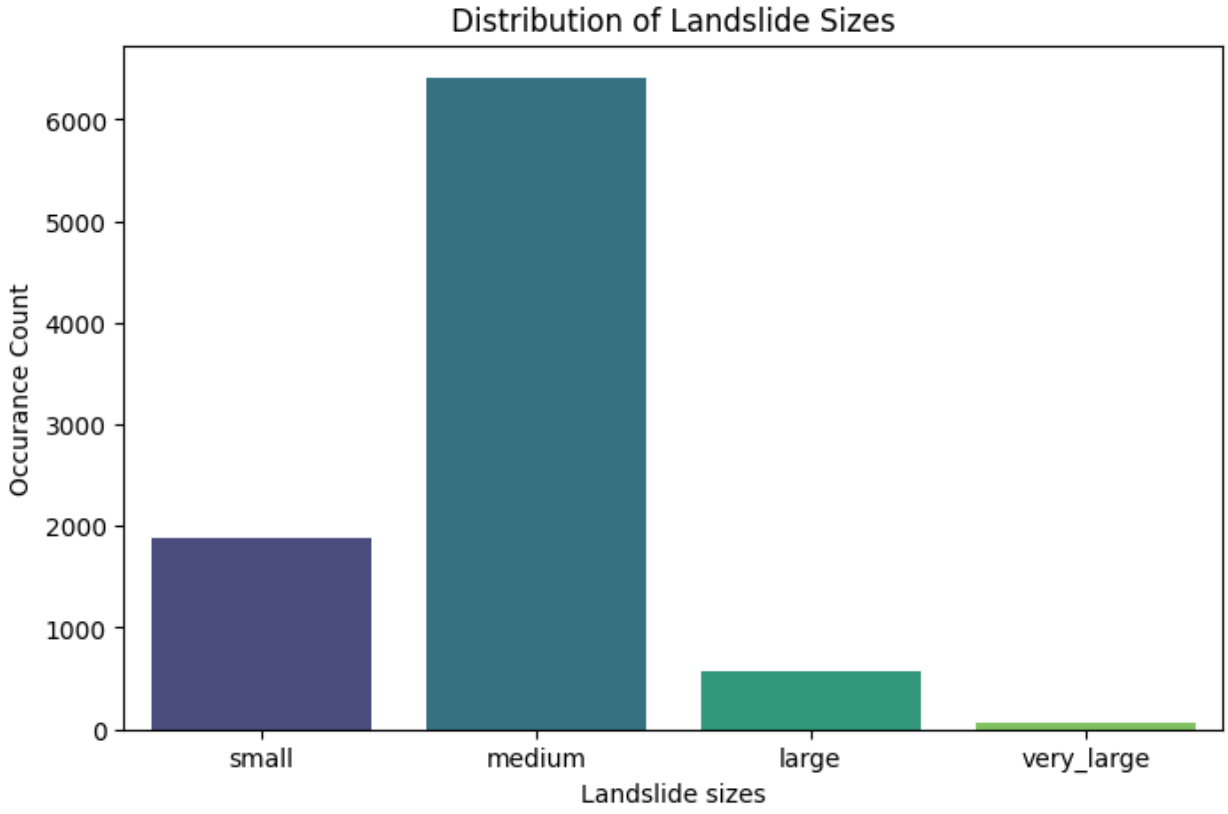

Figure 1: Bar Chart of distribution of landslide sizes

A bar chart was plotted to visualize the distribution of landslide size, which shows that the target variable landslide sizes are highly imbalanced with over 6000 values of a single class medium while small was the second largest class.

2. Histogram of location accuracy

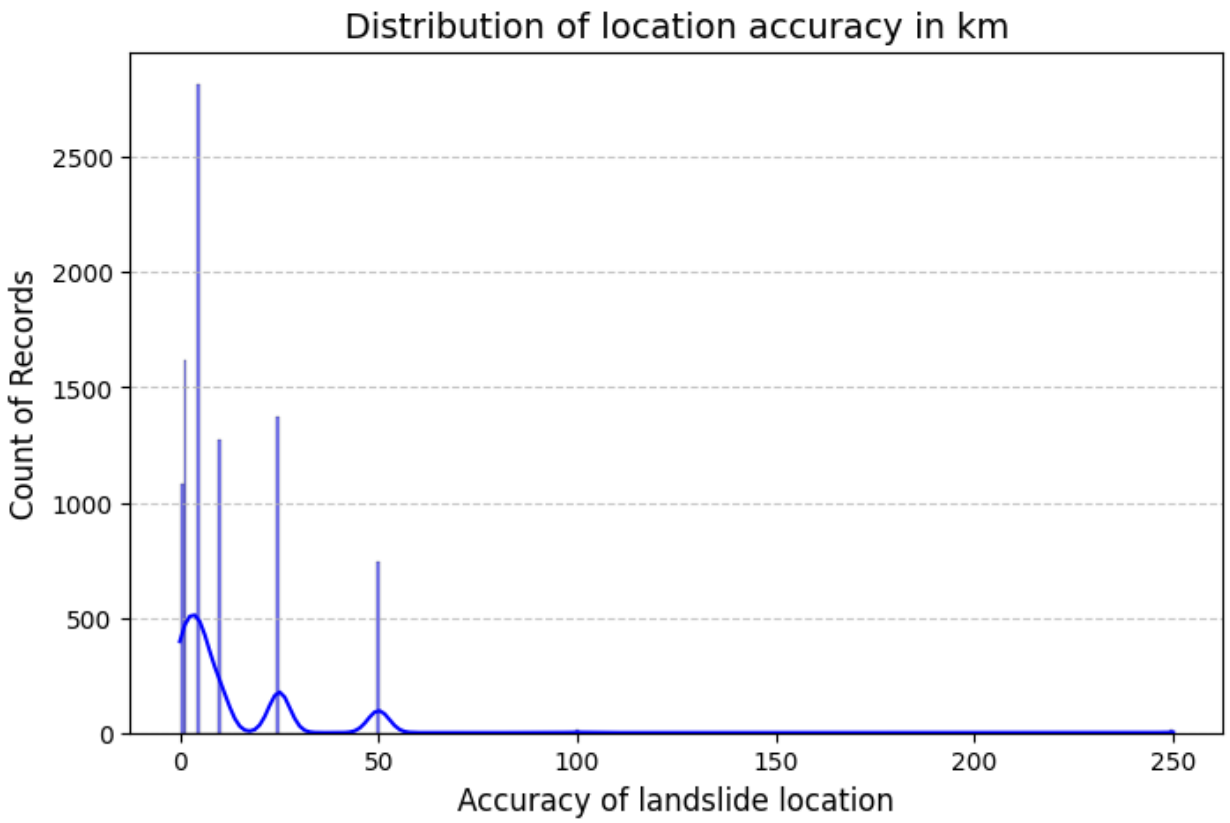

Figure 2: Histogram of Location accuracy

The histogram shows that location accuracy is highly positively skewed.

3. Average population by landslide size

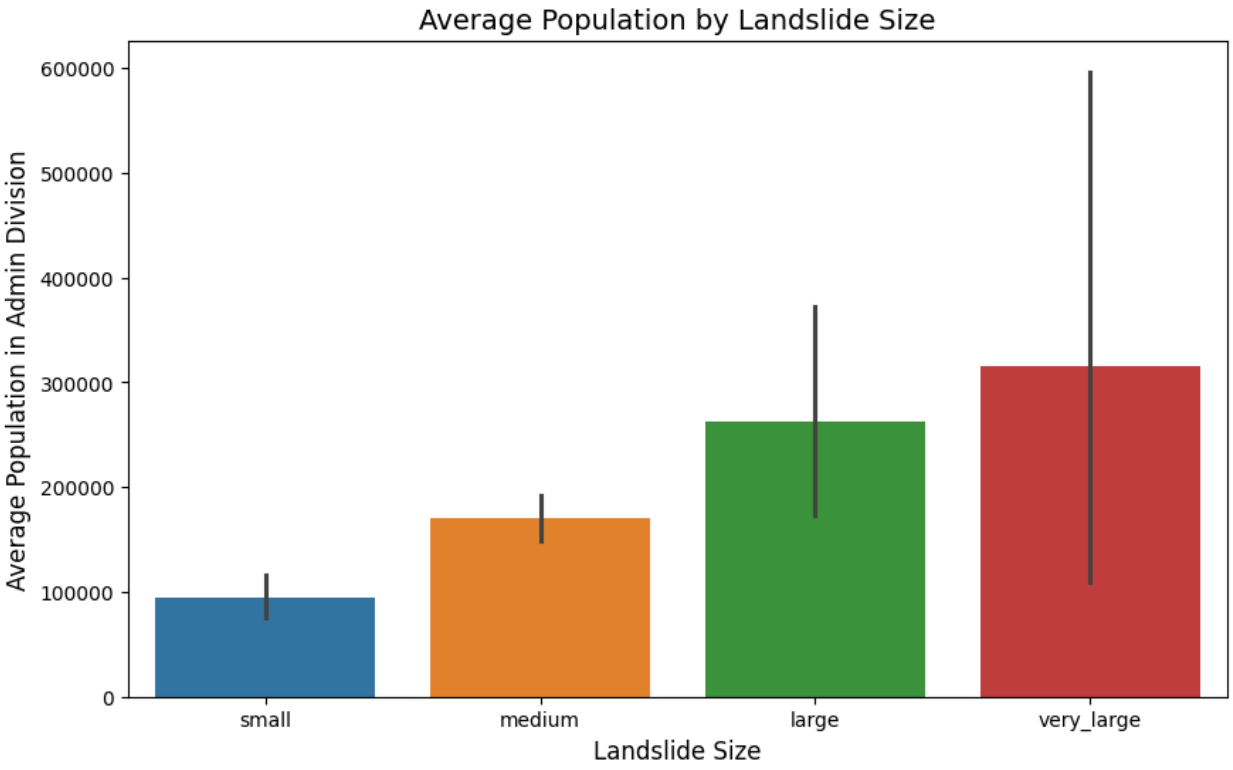

Figure 3: Bar chart of average population by landslide size

The above chart shows that the population is relative to the landslide size. Most of the people reside where large landslides occur and as the population decreases the size of the landslide also increases.

Model Building

Decision Tree Classifier and Random Forest Classifier were chosen to classify landslide sizes. The dataset was first preprocessed and cleaned. After removing landslide size column from dataset and storing it in a new series, Feature Matrix and Label Vector were created. The Feature matrix and label vector were then divided into training and test parts with test part having the 20% of whole dataset, both feature and labels. The training features and labels were then used to train both Decision Tree Classifier and Random Forest Classifier.

Learning curve of cost vs iteration:

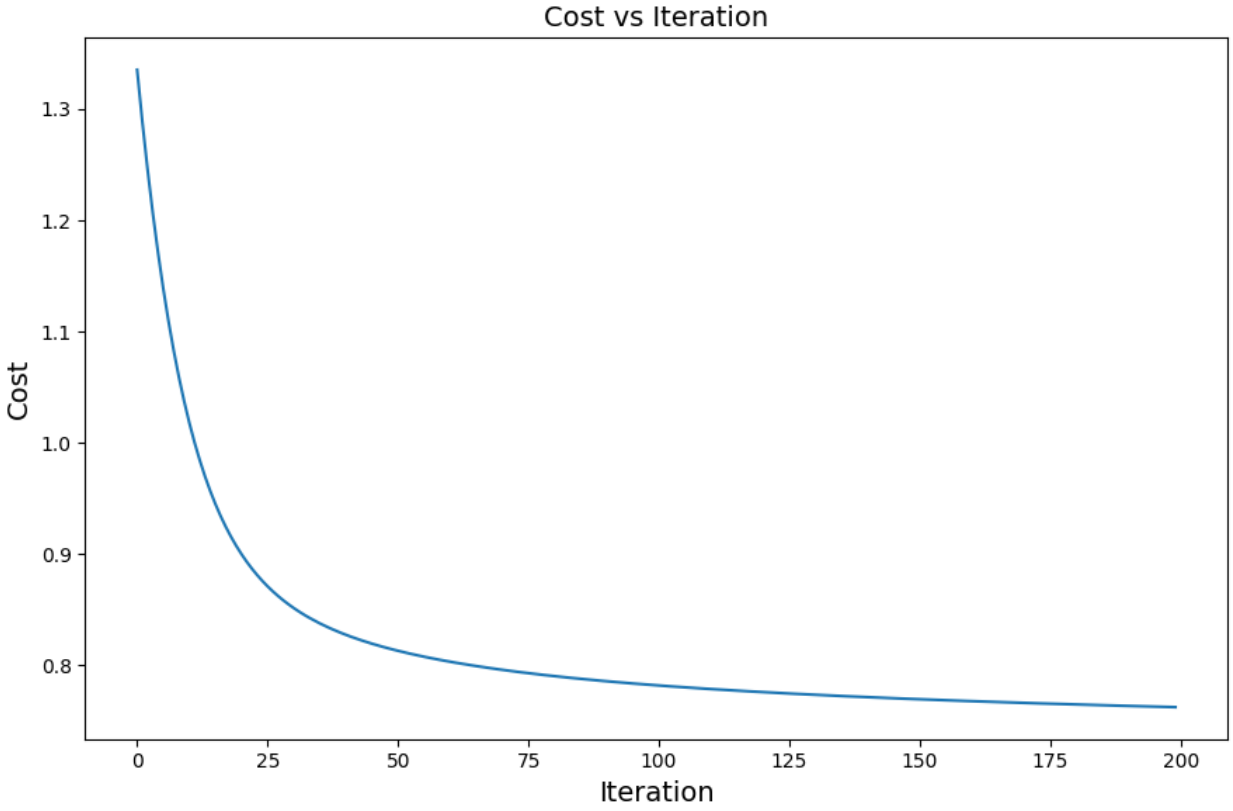

Figure 4: Learning Curve of Cost vs Iteration in training phrase

Model Evaluation

Upon accomplishment of model training, the trained model was given test features to make classification of landslide sizes. The predicted landslide sizes and actual landslide size separated label vector for test part were then evaluated on following metrices:

- **Accuracy:** The proportion of correctly classified instances.
- **Precision:** The proportion of true positive predictions out of all positive predictions made.
- **Recall:** The proportion of true positive predictions out of all actual positive cases.
- **F1-Score:** The harmonic means of precision and recall, balancing both metrics.

For Decision Tree Classifier:

Accuracy	Precision	Recall	f1 score
0.6357	0.64	0.64	0.64

Confusion Matrix :

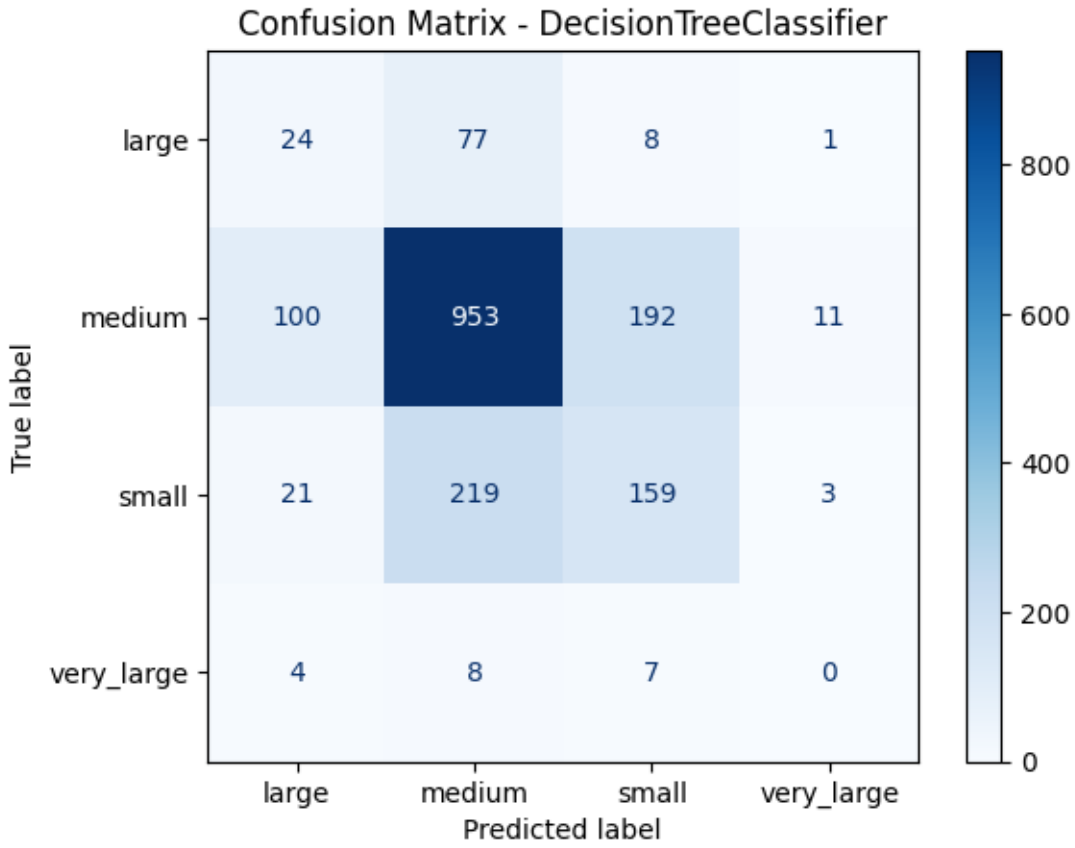

Figure 5: Confusion Matrix of Decision Tree

For Random Forest Classifier:

Accuracy	Precision	Recall	f1 score
0.6955	0.73	0.70	0.71

Confusion matrix:

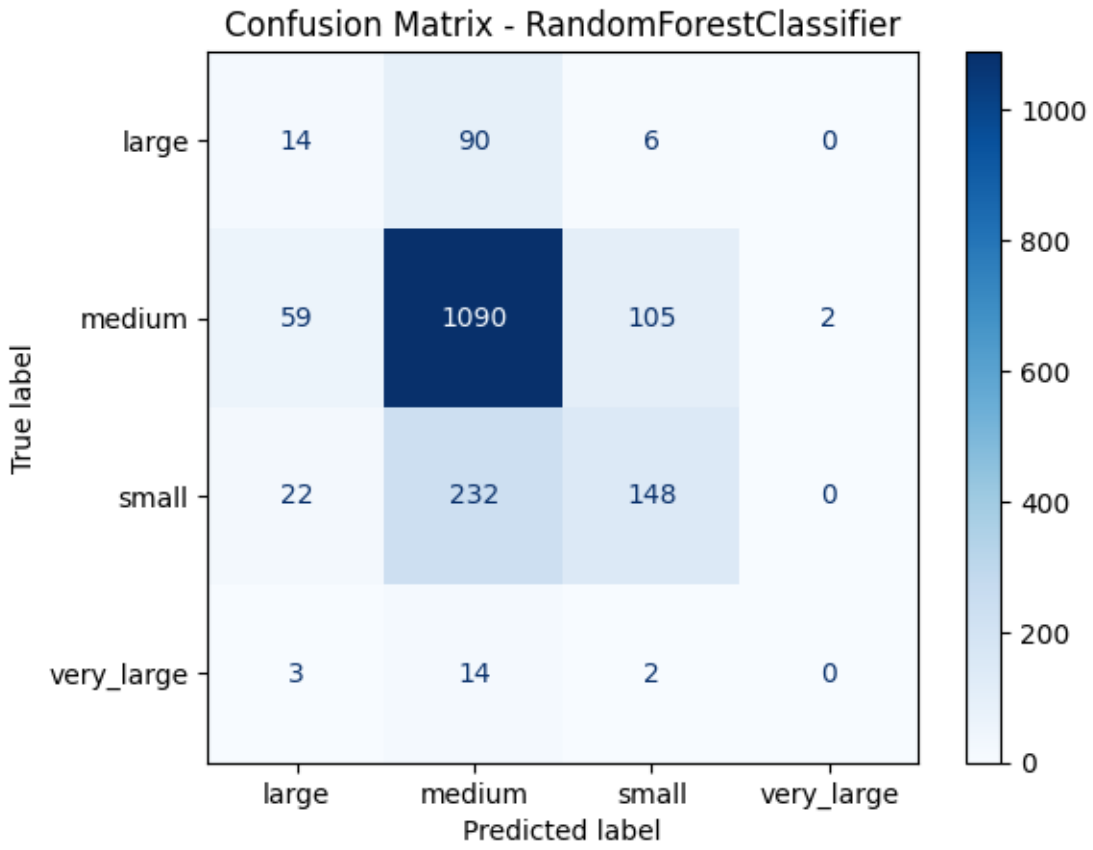

Figure 6: Confusion Matrix of Random Forest

Hyper-parameter optimization

To improve the classification of landslide sizes, hyper-parameters optimization was conducted using GridSearchCV. The optimal parameters for both modes are listed below:

Decision Tree Classifier:

- min_samples_leaf: 1
- min_samples_split: 2

Random Forest Classification:

- min_samples_leaf: 1
- min_samples_split : 2
- n_estimators : 200

Feature Selection

Another way to improve performance of models is to avoid training them on features that do not have any significant impact on models. However, since the dataset after cleaning only has five feature columns and for features selection the number of features selected should be less than total number of features choose to select four features. For this analysis, Recursive Feature Elimination was used to find the four prominent features and only train on the selected features. The selected features found for both models are listed below:

Decision Tree Classifier:	**Random Forest Regression:**
admin_division_population	admin_division_population
gazeteer distance	gazeteer distance
latitude	latitude
longitude	longitude

Same features were chosen for both models by Recursive Feature Selection.

Conclusion

Key Findings

The model's performance on the test dataset was evaluated using Accuracy, Precision, Recall, and F1-Score. The results showed that the medium landslide category performed well with high recall (~86%), small and large showed inferior performance, especially in recall and precision. The large class had moderate performance. The overall accuracy of the final model was 69.22%, with a weighted average F1-Score of 0.67.

Final Model

- The final model Decision Tree Classifier is then trained on admin_division_population, gazeteer_distance, latitude, longitude with min_samples_leaf set to 1,min_samples_split set to 2 and n_estimators set to 200.

, the metrics obtained from this model are as follows:

Accuracy	Precision	Recall	f1 score
0.6922	0.67	0.69	0.67

Confusion Matrix:

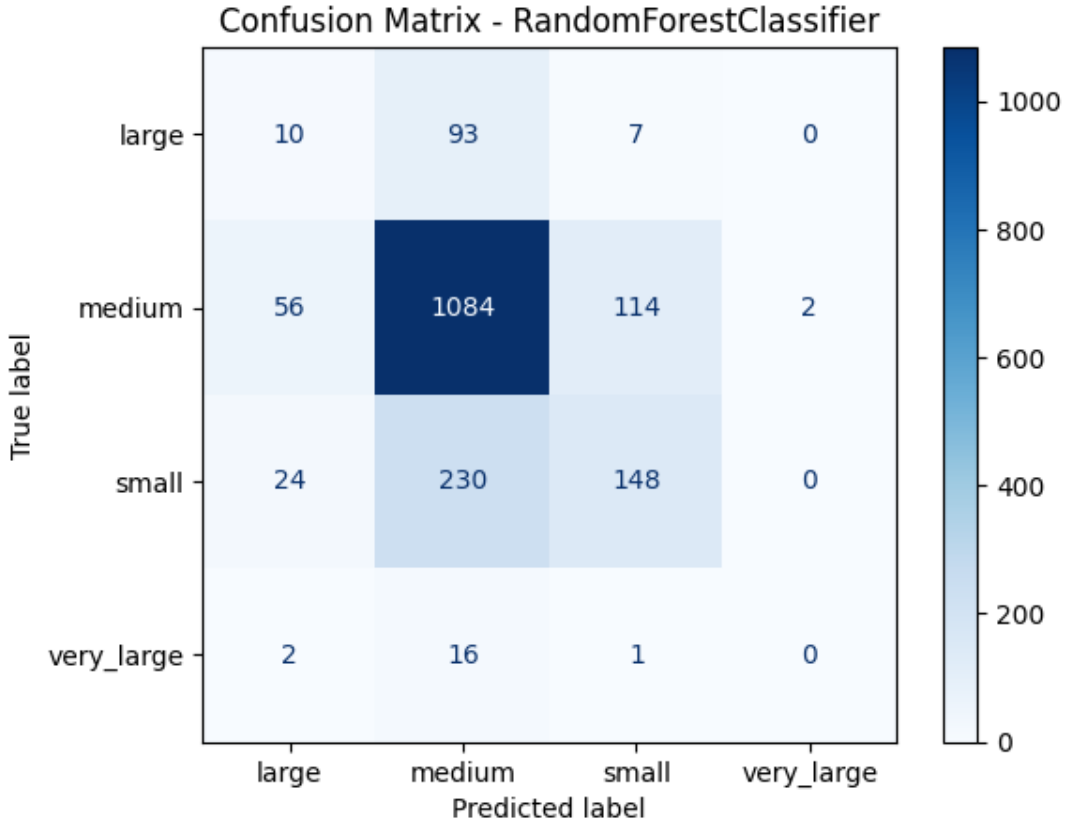

Figure 7: Confusion Matrix of Final Model

Challenges

Working on this project, many challenges were faced. First challenge was cleaning the dataset. More than 20 columns of dataset were irrelevant. Despite being relevant injury count, location setting and fatality count had considerable number of missing values. Also, the dataset is significantly positively skewed. Another challenge faced was lack of computation power. Considerable time was required to optimize the parameters.

Future Task

To classify target more accurately feature transformation could be applied to create new derived features. To address imbalance techniques such as SMOTE or adjusting class weights could be used. Lastly, incorporating external dataset sources such as weather or geological data could provide more context and improve mode's robustness.

Discussion

Model Performance

The classification models were evaluated based on accuracy, precision, recall and f1 score. Based on those metrics, Random Forest Classifier performed better than Decision Tree Classifier on all metrics. The reason may be Random Forest Classifier ability to manage more complex patterns and reduce overfitting with the help ensemble of multiple decision trees. While the Decision Tree is faster and simpler than Random Forest, its single tree nature made it prone to overfitting, which reduced its performance on test set.

Impact of hyperparameter tuning and feature selection

The hyperparameters and feature selection play a crucial role in improving the model's performance, however in this case, after the implementation of hyperparameters, the metrics of model evaluation have decreased stating that model's performance has worsened. It is likely that reducing the features has caused the model unable to capture the patterns as well as it did before.

Result Interpretation

The Random Forest Classification accuracy has dropped by 0.033 and f1 score by 0.04, suggesting the feature selection does not improve the model performance especially when the dataset has less features.

Limitations and suggestions for future research

Despite successfully building the model, some limitations remain, these include less features, which might not have provided enough variance to allow models to learn patterns. For further research, balancing the class labels and experimenting with different classification algorithms such as Gradient Boosting or Neural Network, may find the complex relationship between features and target, leading to better performance.