Enhancing Sustainability: Leveraging Machine Learning for Natural Disaster Prediction

Introduction

The dissertation titledembarks on a transformative journey aimed at enhancing global resilience and fostering sustainable practices. The primary objective of this research endeavor is to advocate for the preservation of sustainability by leveraging predictive techniques to mitigate the adverse impacts of disasters on ecosystems, natural resources, and the environment. This objective is in alignment with global initiatives to address the challenges posed by climate change. Through the utilization of machine learning models including K-nearest neighbors, Random Forest, Support Vector Machine (SVM), and Naive Bayes, this study endeavors to serve as an early warning system, enabling proactive disaster preparedness and response strategies. By conducting comprehensive analyses of historical data pertaining to various types of calamities such as earthquakes, hurricanes, floods, and wildfires, and employing advanced techniques such as feature engineering and machine learning algorithms, the aim is to develop accurate prediction models that provide actionable insights for effective decision-making and intervention.

- Our project seeks to improve global resilience by using predictive methods to reduce the negative effects of disasters on ecosystems and natural resources.

- The project employs machine learning models such as K-means clustering, Random Forest, SVM, and Naive Bayes for predicting disaster types by thorough analysis of historical data

- This project aims to develop practical insights and effective actions for different disasters

- To ensure the reliability and accuracy of the developed prediction models, a thorough performance assessment has been carried out.

- The project's transformative mission emphasizes its dedication to promoting increased resilience, sustainable methods, and a proactive approach in addressing growing environmental challenges.

Literature Review

- The authors [1] provide a thorough analysis of machine learning's application in disaster management, covering every stage from recovery to prediction. They highlight well-known supervised techniques like Naïve Bayes, SVM, and Random Forest.

Link: https://ieeexplore.ieee.org/document/9295332

- The authors [2] present a unique method that uses machine learning to forecast various natural disasters based on environmental signals. With 92.1% accuracy, SVM performs noticeably better than other methods. Problems are emphasized to direct future feature-focused research.

LINK:https://www.researchsquare.com/article/rs-204305/v1

- Using weather data from the previous ten years, [3] examines 24-hour sandstorm prediction, using SMOTE to address data imbalance. By using 10-fold cross-validation, Random Forest achieves 96.51% accuracy with zero false alarms, outperforming both Naïve Bayes and logistic regression.

LINK:https://ieeexplore.ieee.org/document/8441998

- The authors of [5] examine several approaches for predicting fire outbreaks. They emphasize a variety of criteria in addition to accuracy, recommending 86% accuracy, greater precision, and recall for bagging decision trees; Random Forests show stronger sensitivity.

LINK:https://link.springer.com/article/10.1007/s10618-011-0213-2

- While noting problems with data imbalance, [6] explores the use of machine learning for earthquake prediction. The application of SMOTE improves the performance of SVM and Decision Trees. With a 0.86 ROC hit rate, Decision Tree outperforms SVM by 2%; the MMC metric exhibits potential despite the differences in class sizes.

LINK:https://www.frontiersin.org/articles/10.3389/feart.2022.847808/full

Data

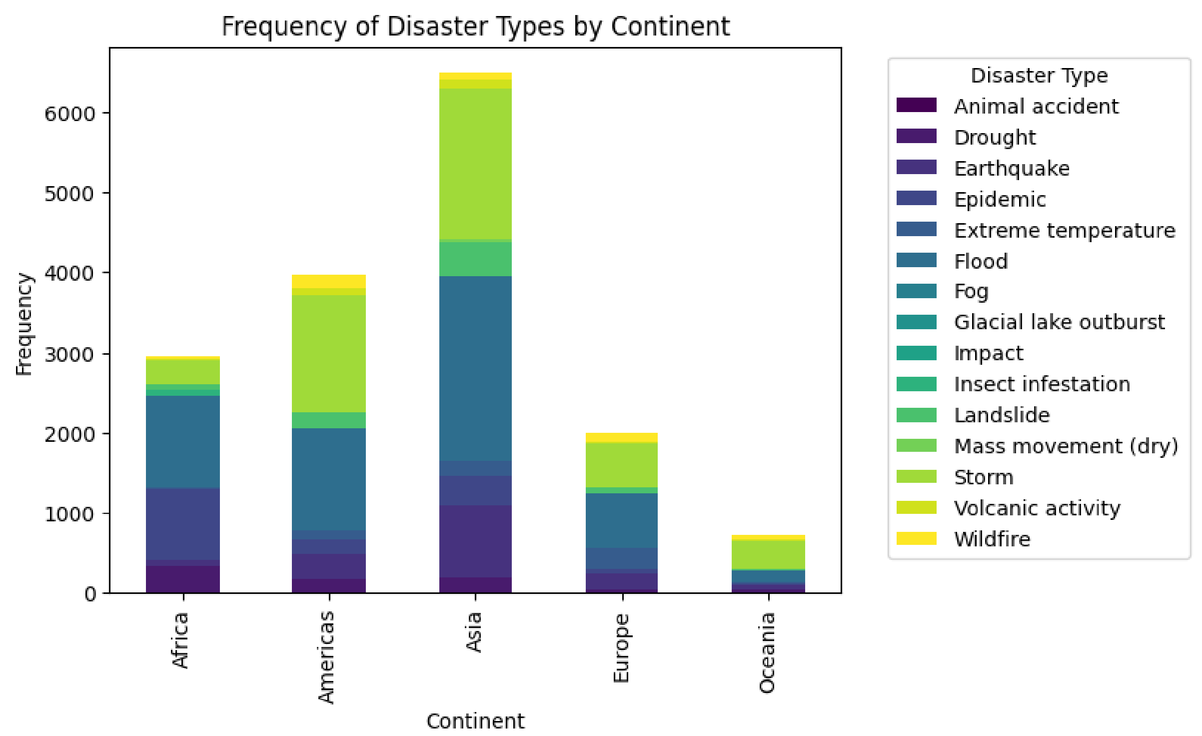
1: Data Collection

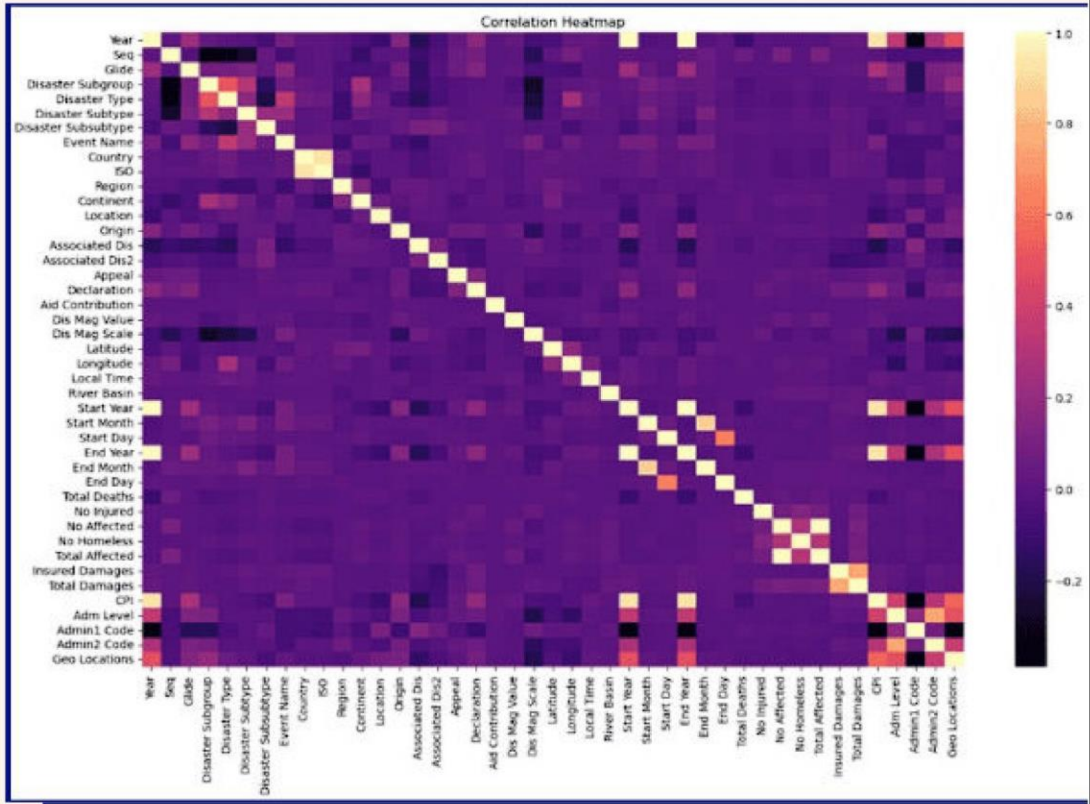
- The dataset is collected form Kaggle.
- It has historical data records which date form 1900 to 2021. It consists of more than 16000 records and 22 columns.
- Link: https://www.kaggle.com/datasets/brsdincer/all-natural-disasters-19002021-eosdis
- After collecting the data, exploratory data analysis was performed on it for better understanding.

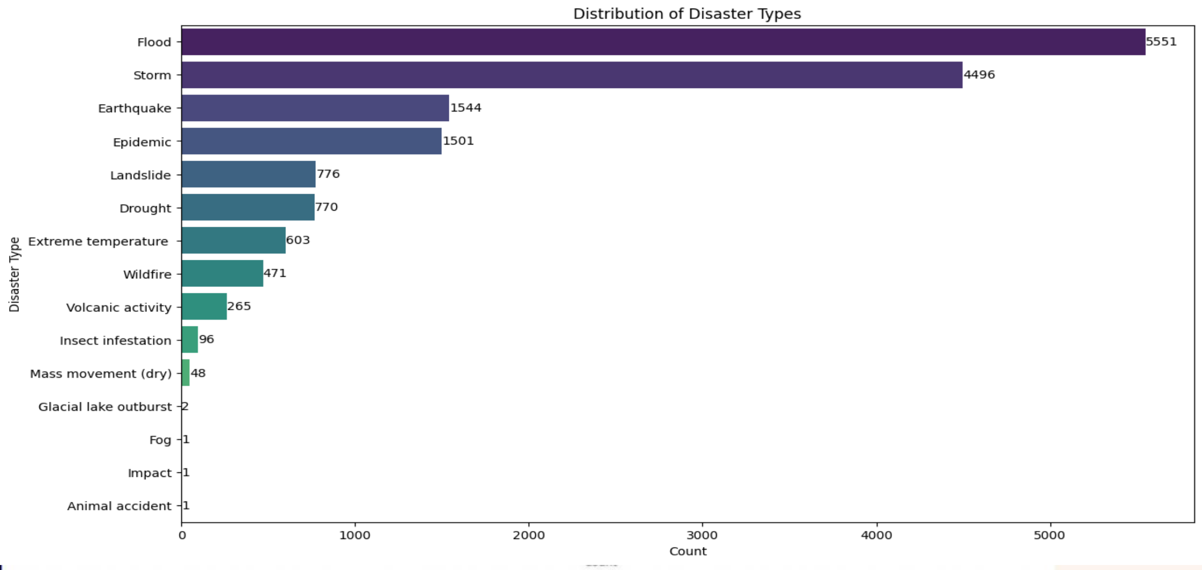
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 16126 entries, 0 to 16125
Data columns (total 45 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   Year                          16126 non-null  int64
 1   Seq                           16126 non-null  int64
 2   Glide                         1581 non-null   object
 3   Disaster Group                16126 non-null  object
 4   Disaster Subgroup             16126 non-null  object
 5   Disaster Type                 16126 non-null  object
 6   Disaster Subtype              13016 non-null  object
 7   Disaster Subsubtype           1077 non-null   object
 8   Event Name                    3861 non-null   object
 9   Country                       16126 non-null  object
 10  ISO                           16126 non-null  object
 11  Region                        16126 non-null  object
 12  Continent                     16126 non-null  object
 13  Location                      14334 non-null  object
 14  Origin                        3794 non-null   object
 15  Associated Dis                3348 non-null   object
 16  Associated Dis2               707 non-null    object
 17  OFDA Response                 1694 non-null   object
 18  Appeal                        2569 non-null   object
 19  Declaration                   3256 non-null   object
 20  Aid Contribution              677 non-null    float64
 21  Dis Mag Value                 4946 non-null   float64
 22  Dis Mag Scale                 14936 non-null  object
 23  Latitude                      2729 non-null   object
 24  Longitude                     2732 non-null   object
 25  Local Time                    1103 non-null   object
 26  River Basin                   1287 non-null   object
 27  Start Year                    16126 non-null  int64
 28  Start Month                   15739 non-null  float64
 29  Start Day                     12498 non-null  float64
 30  End Year                      16126 non-null  int64
 31  End Month                     15418 non-null  float64
 32  End Day                       12570 non-null  float64
 33  Total Deaths                  11413 non-null  float64
 34  No Injured                    3895 non-null   float64
 35  No Affected                   9220 non-null   float64
 36  No Homeless                   2430 non-null   float64
 37  Total Affected                11617 non-null  float64
 38  Insured Damages ('000 US$)    1096 non-null   float64
 39  Total Damages ('000 US$)      5245 non-null   float64
 40  CPI                           15811 non-null  float64
 41  Adm Level                     7859 non-null   object
 42  Admin1 Code                   4581 non-null   object
 43  Admin2 Code                   3969 non-null   object
 44  Geo Locations                 7859 non-null   object
dtypes: float64(14), int64(4), object(27)
memory usage: 5.5+ MB
None
```

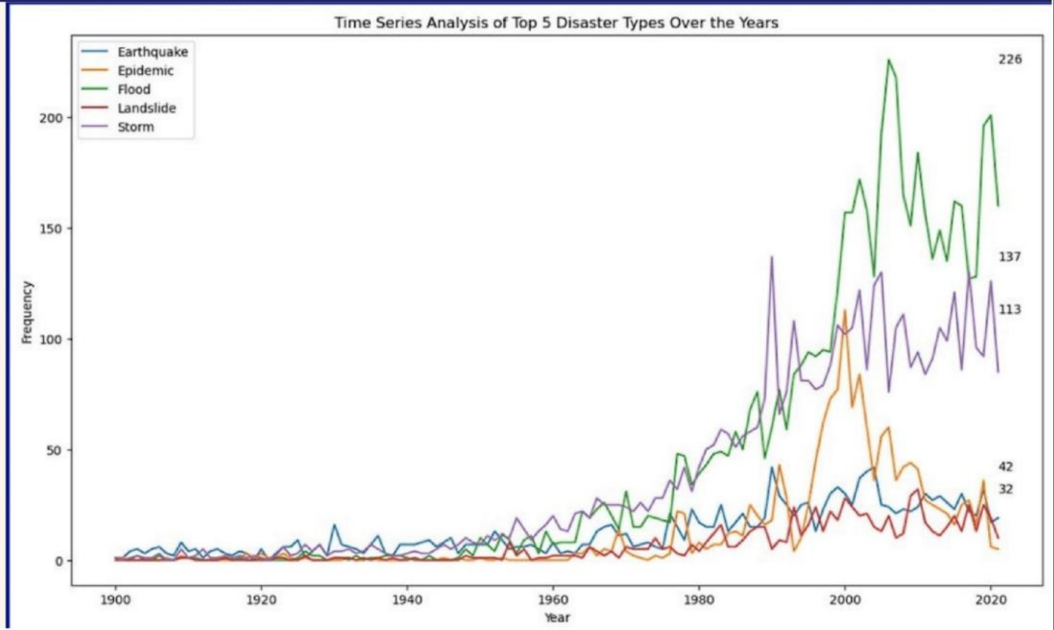
Methodology

EDA


Frequency of Disaster Types by Continent
Frequency
6000
5000
4000
3000
2000
1000
0
Africa
Americas
Asia
Europe
Oceania
Continent
Disaster Type
Animal accident
Drought
Earthquake
Epidemic
Extreme temperature
Flood
Fog
Glacial lake outburst
Impact
Insect infestation
Landslide
Mass movement (dry)
Storm
Volcanic activity
Wildfire


Correlation Heatmap
Year
Seq
Glide
Disaster Subgroup
Disaster Type
Disaster Subtype
Disaster Subsubtype
Event Name
Country
ISO
Region
Continent
Location
Origin
Associated Dis
Associated Dis2
Appeal
Declaration
Aid Contribution
Dis Mag Value
Dis Mag Scale
Latitude
Longitude
Local Time
River Basin
Start Year
Start Month
Start Day
End Year
End Month
End Day
Total Deaths
No Injured
No Affected
No Homeless
Total Affected
Insured Damages
Total Damages
CPI
Adm Level
Admin1 Code
Admin2 Code
Geo Locations
1.0
0.8
0.6
0.4
0.2
0.0
-0.2


Distribution of Disaster Types
Disaster Type
Flood 5551
Storm 4496
Earthquake 1544
Epidemic 1501
Landslide 776
Drought 770
Extreme temperature 603
Wildfire 471
Volcanic activity 265
Insect infestation 96
Mass movement (dry) 48
Glacial lake outburst 2
Fog 1
Impact 1
Animal accident 1
0
1000
2000
3000
4000
5000
Count


Time Series Analysis of Top 5 Disaster Types Over the Years
Earthquake
Epidemic
Flood
Landslide
Storm
Frequency
200
150
100
50
0
226
137
113
42
32
1900
1920
1940
1960
1980
2000
2020
Year

Data Preprocessing

- For the data preprocessing we have encoded all the categrorical columns in our dataset using Label encoding method.

Data Cleaning

```
Year                              0
Seq                               0
Glide                             0
Disaster Group                    0
Disaster Subgroup                 0
Disaster Type                     0
Disaster Subtype                  0
Disaster Subsubtype               0
Event Name                        0
Country                           0
ISO                               0
Region                            0
Continent                         0
Location                          0
Origin                            0
Associated Dis                    0
Associated Dis2                   0
OFDA Response                     0
Appeal                            0
Declaration                       0
Aid Contribution                  0
Dis Mag Value                     0
Dis Mag Scale                     0
Latitude                          0
Longitude                         0
Local Time                        0
River Basin                       0
Start Year                        0
Start Month                       0
Start Day                         0
End Year                          0
End Month                         0
End Day                           0
Total Deaths                      0
No Injured                        0
No Affected                       0
No Homeless                       0
Total Affected                    0
Insured Damages ('000 US$)        0
Total Damages ('000 US$)          0
CPI                               0
Adm Level                         0
Admin1 Code                       0
Admin2 Code                       0
Geo Locations                     0
dtype: int64
```

- As our dataset was having missing values in many columns, we imputed them.
- For numerical values we have used "Mean" of for imputation.
- For categorical values we have used "Mode" for imputation.

Feature Engineering

```
In [20]: data_selected.head(10)
```

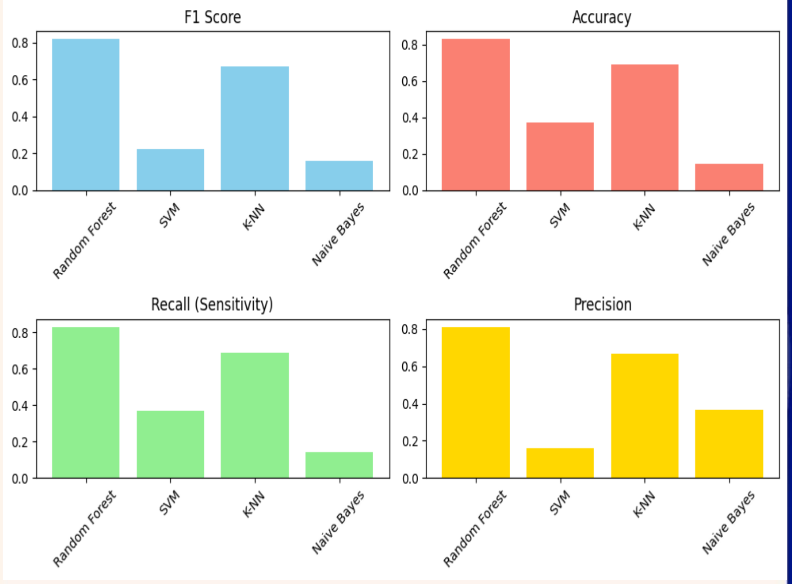
Out[20]:

	Year	Dis Mag Scale	Dis Mag Value	Country	Longitude	Latitude	Disaster Type
0	1900.0	0	47350.380307	31	2376	1242	1
1	1900.0	0	47350.380307	89	2376	1242	1
2	1902.0	2	8.000000	80	482	662	2
3	1902.0	0	47350.380307	80	2376	1242	13
4	1902.0	0	47350.380307	80	2376	1242	13
5	1903.0	0	47350.380307	34	2376	1242	11
6	1903.0	0	47350.380307	42	2376	1242	13
7	1904.0	1	47350.380307	15	2376	1242	12
8	1905.0	0	47350.380307	34	2376	1242	11
9	1905.0	2	8.000000	89	2151	1329	2

- For the feature selection we have used mutual and domain knowledge and we have made a feature set.
- Our feature set consists of the following features: 'Year', 'Dis Mag Scale', 'Dis Mag Value', 'Country', 'Longitude', 'Latitude', 'Disaster Type'.

Navie Bayes

- We chose Naive Bayes, which is a probabilistic model based on Bayes' theorem and assumes that features are independent.
- Specifically, the Gaussian Naive Bayes variant was well-suited for this task because it works effectively with continuous numerical features commonly found in natural disasters.
- Despite its simplicity, our model showed strong performance metrics and accurately classified different types of natural disasters.
- The probabilistic approach and simplicity of the Gaussian Naive Bayes model make it a practical choice for predicting various scenarios related to natural disasters.

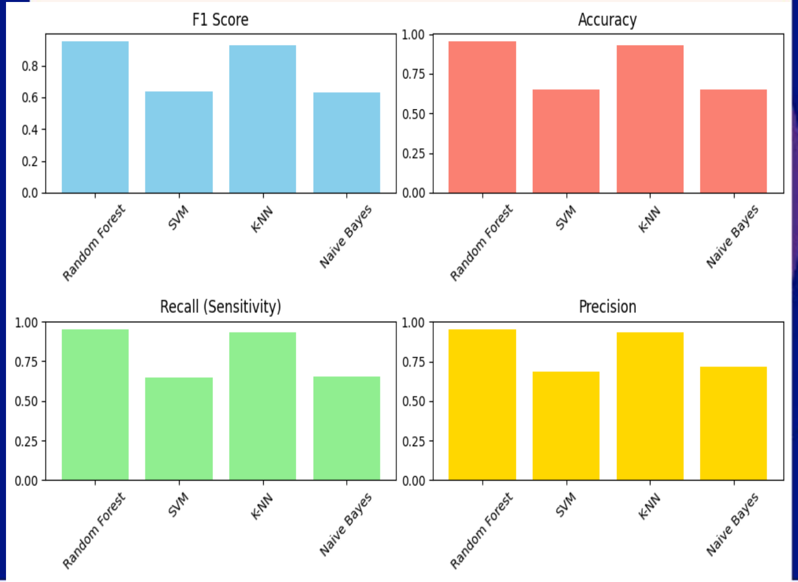


K-Nearest Neighbor

- The K-Nearest Neighbors algorithm was selected due to its simplicity and ability to capture local patterns. It classifies instances by considering the majority class among their k-nearest neighbors.
- Before training, features were scaled using StandardScaler in order to prevent dominance by larger-scaled features in the distance metric.
- Evaluation metrics showed that the KNN model had high accuracy and other favorable indicators, further supporting its appropriateness for predicting various types of natural disasters.

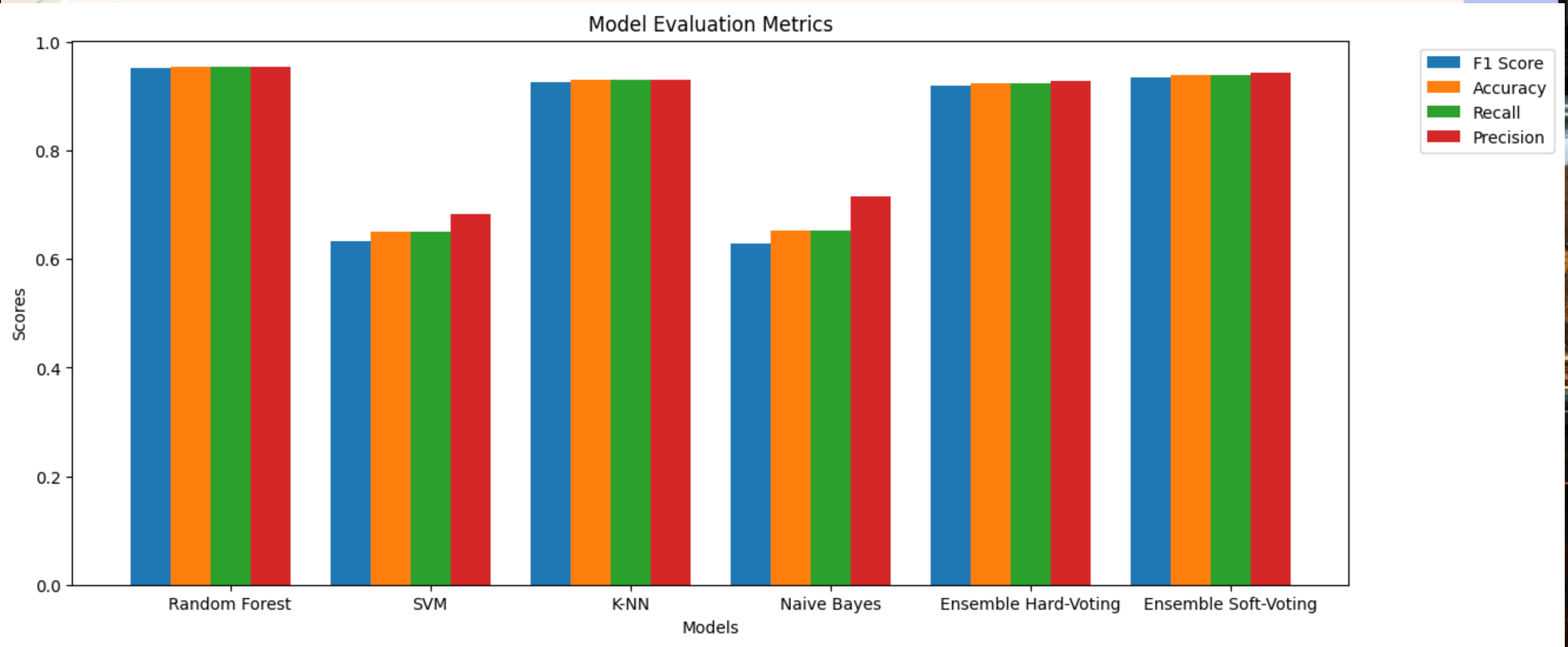
Support Vector Machines

- The Support Vector Machine with a linear kernel was chosen for its ability to handle high-dimensional data and separate classes effectively.
- One important preprocessing step involved scaling the features using StandardScaler to ensure equal contribution from each feature in the SVM model.
- Additionally, when deployed, the SVM model showed effectiveness as part of a soft voting ensemble.

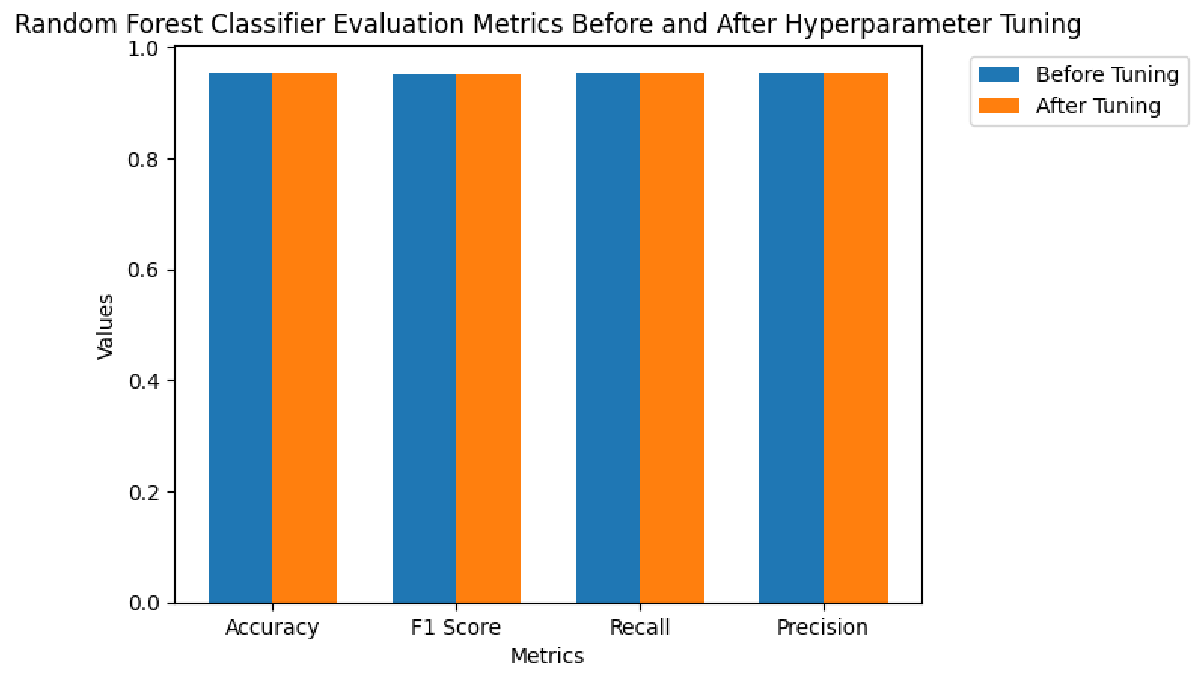


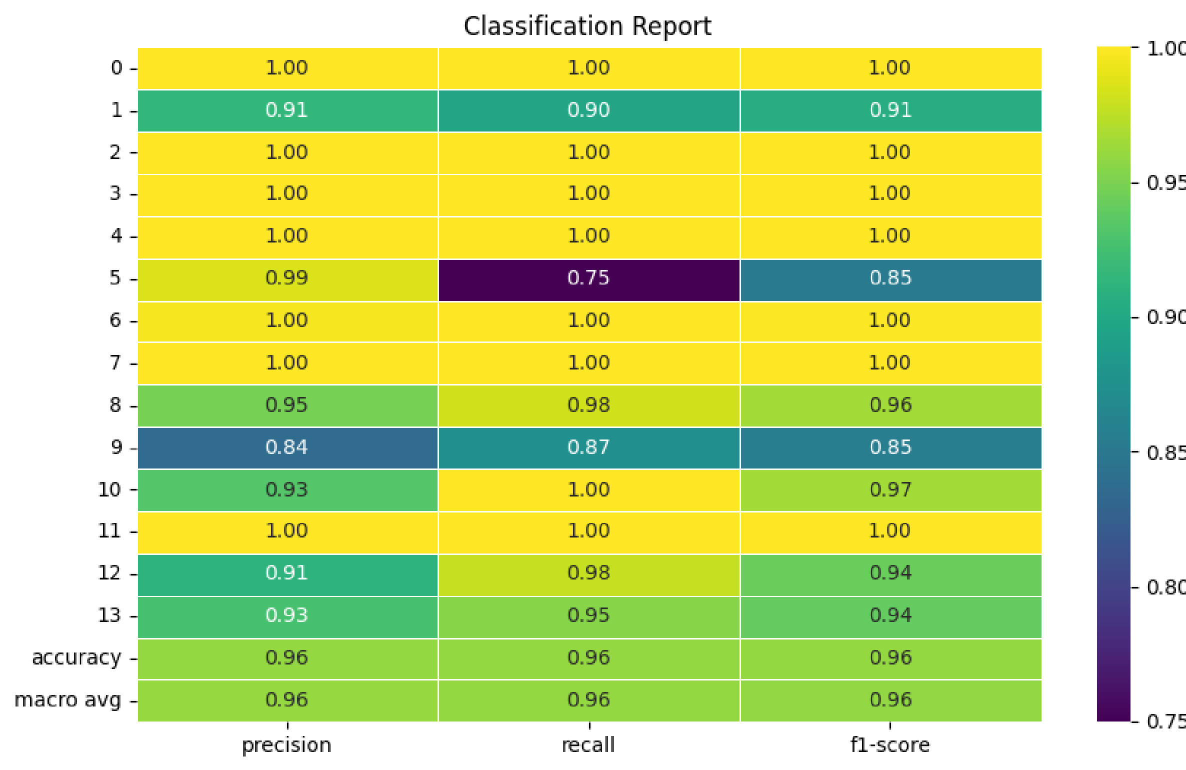
Random Forest

- The Random Forest model was selected due to its ability to effectively handle intricate data relationships and prevent overfitting.
- To optimize the model's accuracy and generalization, we have selected parameters like the number of trees (n_estimators) and maximum tree depth (max_depth) were fine-tuned using GridSearchCV.
- The implementation of the Random Forest model resulted in impressive performance measures such as accuracy, F1 score, recall, and precision.
- These metrics collectively highlight the effectiveness of the model in accurately classifying various forms of natural disasters.



Comparison after


Random Forest Classifier Evaluation Metrics Before and After Hyperparameter Tuning
Before Tuning
After Tuning
Values
1.0
0.8
0.6
0.4
0.2
0.0
Accuracy
F1 Score
Recall
Precision
Metrics


Classification Report
precision
recall
f1-score
0 1.00 1.00 1.00
1 0.91 0.90 0.91
2 1.00 1.00 1.00
3 1.00 1.00 1.00
4 1.00 1.00 1.00
5 0.99 0.75 0.85
6 1.00 1.00 1.00
7 1.00 1.00 1.00
8 0.95 0.98 0.96
9 0.84 0.87 0.85
10 0.93 1.00 0.97
11 1.00 1.00 1.00
12 0.91 0.98 0.94
13 0.93 0.95 0.94
accuracy 0.96 0.96 0.96
macro avg 0.96 0.96 0.96
1.00
0.95
0.90
0.85
0.80
0.75

Conclusion

In conclusion, the integration of **machine learning** into natural disaster prediction represents a significant step forward in enhancing **sustainability**. By harnessing the power of technology, we can better prepare for and respond to natural disasters, ultimately contributing to a safer and more resilient future.