Sustainable Future through Natural Disaster Prediction

Damini Prashant Vichare, Manisha Lagisetty, Sweekruthi Balivada, Yuting Sha

***Abstract*—Amidst the escalating environmental difficulties, the convergence of sustainability and disaster management has assumed paramount importance. This ML project, entitled "Sustainable Futures through Natural Disaster Prediction," embarks on a transformative quest towards promoting heightened resilience and sustainable practices globally. The primary goal of this project is to promote the preservation of sustainability. By employing predictive techniques, we strive to reduce the detrimental impact that disasters have on ecosystems, natural resources, and the environment. The objective aligns with worldwide endeavors in addressing climate change. Through our machine learning models such as K-Nearest Neighbors, Random Forest, Support Vector Machine (SVM), and Naive Bayes are used as early warning indicators, thus facilitating proactive approaches for disaster preparedness and response strategies. By thoroughly analyzing historical data on diverse types of calamities such as earthquakes, hurricanes, floods, and wildfires and by employing feature engineering, and machine learning algorithms our aim is to develop precise prediction models that offer practical insights for effective action. We will meticulously assess the performance of our prediction models in order to ensure their reliability and accuracy.**

***Index Terms*—Machine Learning, Natural Disaster, SVM, KNN, Random Forests, Naïve Bayes**

I. Introduction

NATURAL calamities inflict immeasurable anguish upon human beings, resulting in fatalities, injuries, and the displacement of affected individuals. Natural disasters can have dire consequences on our environment also, like wildfires contribute to deforestation while industrial accidents during such crises lead to pollution. It is incumbent upon us to anticipate and counteract these events as part of our duty towards environmental preservation and ensuring a sustainable planet for generations yet unborn. Utilizing early warning systems and promoting preparedness information can greatly strengthen community resilience. By equipping communities with necessary knowledge and tools, they are able to safeguard their residences, economic activities, and cultural significance, thereby fostering sustainable development in the long run. By accurately forecasting these occurrences, we are afforded the chance to not only safeguard lives but also alleviate suffering and extend timely aid to those who require it.

The primary goal of this project is to develop a reliable and accurate machine learning-based system for predicting natural disasters, utilizing insights gained from a historical data set spanning the years 1900 to 2021. Using Machine learning algorithms on historical data spanning more than a century we make an effort to improve our understanding of the intricate patterns and factors leading up to different types of natural disasters.

The proposed system integrates data from Gov website historical disaster dataset. Feature engineering is employed to extract relevant information, while advanced ML algorithms are utilized for accurate pattern recognition and predictive modeling. The project considers both spatial and temporal dimensions, which will in turn enhance the ability to capture patterns and interactions leading to natural disasters.

The outcome of our project will contribute to predicting natural disasters and have significant implications for disaster management and the resilience of society. Enabling early warning systems, for emergency response optimizing resource allocation, and building strategies for mitigation by developing accurate predictive models is what we thrive for. Our project can be utilized by policymakers to craft effective policies, encompassing zoning regulations and investments in early warning technologies. The project's insights can empower communities and policymakers with a good amount of knowledge to build resilience and enhance public safety through education and awareness initiatives.

Objective

The ultimate objective of our project is to predict the upcoming natural disasters there by providing early warnings and insights that can empower communities, emergency responders, and policymakers to take proactive measures in preparation for and response to potential disasters and to maintain sustainable futures. By combining Data Analytics, and Machine Learning techniques, the project aspires to improve our understanding, prediction and lessen the severity of the effects of natural disasters.

In order to achieve our objective, we are implementing four different models on our dataset by exploring different features, that can possibly affect the occurrences of different disasters. These models include Random Forest, Support Vector Machines, K-Nearest Neighbor and Naïve Bayes.

II. Theoretical Bases and Literature Survey

A. Problem Definition

Our project focusses on predicting Natural disasters using innovative Machine Learning models including Ensemble models, precisely. To improve the effectiveness of our system, we are incorporating various features and leveraging advanced methodologies that estimates natural disaster types.

B. Problem Theoretical Background

The primary purpose of this project is natural disaster prediction using machine learning technologies. We aim to develop a robust model using large historical disaster records from 1900 to 2021 as a training dataset, capturing complex relationships and patterns. The main task is to predict natural disasters by using 'Disaster Type' as the target variable and features like 'Year', 'Continent', 'Country', etc. Various supervised classification learning methods like Support Vector Machine, Random Forest, K-Nearest neighbor as well and Naïve Bayes are used in this research. We'll pick the best model by comparing different evaluation metrics and save the model for potential real-time application. In a nut, this research emphasizes disaster prediction and early warning systems enabling local governments and communities to take proactive actions, saving lives and for a sustainable future.

C. Literature Survey

When exploring the technical feasibility of our target problem, we come across a comprehensive review by authors of [1] on how to address disaster management using various machine learning algorithms. The paper presents abundant existing work on natural disaster management in the different phases, including disaster prediction and early warning, immediate responses and crowed evacuation, post-disaster recovery, and future risk mitigation. It shows that supervised learning methods such as Random Forest, SVM, and Naïve Bayes are popular in various types of disaster prediction.

The author of [2] proposes an innovative application that uses environmental conditions to predict different types of natural disasters. Instead of relying solely on location and time data, this application leverages natural signals and features provided by nature. The study investigates various machine learning algorithms to determine the most effective model for this predictive application. Results indicate that SVM achieved the highest accuracy at 92.1%, surpassing other commonly used algorithms such as K-Nearest Neighbors, regression, and ensemble methods. Furthermore, the article discusses encountered challenges in the predictive process and suggests future directions for feature research.

The study in [3] explores the method of predicting whether there is likely to be a sandstorm up to 24 hours ahead of the real-time. They collected past-ten-year weather data for this prediction, a methodology similar to ours, where we use the historical catastrophe dataset. The oversampling skill SMOTE was used in this paper to address the data imbalance. Two approaches to data splitting are employed in the modeling stage. One is splitting the dataset at a ratio of 6:4, which is abandoned due to poor generalization, the other is the 10-fold cross-validation method, which turns out to have a better performance. It concludes that random forest outperforms Naïve Bayes and logistic regression, achieving an accuracy of 96.51% with no false alarm.

In article [4], the author proposed an approach to achieve sustainable development goals through predicting, controlling, and monitoring floods. Machine Learning algorithms like Logistic Regression, Support vector machine are used for this research. Strategies are applied according to three different scenarios. Namely, Pre-Flood Activities (Pre-FA), During Flood Activities (DFA), and Post-Flood. However, for this research, only one country has been considered. In our project, we will overcome these problems by considering all-natural disaster data for all regions. We will use Random Forest, KNN in addition to SVM.

Authors of [5] applied both traditional approaches and state-of-the-art algorithms to develop predictive models of fire outbreaks in the wild. Various methods are taken to reduce the effects of fires in different stages, including fire outbreak prediction, fire detection, and fire spread and burn severity modeling. The study employs diverse performance metrics such as accuracy, Kappa statistic, precision, recall, and AUC. It emphasizes the importance of evaluating with different metrics instead of using accuracy as the only benchmark. The findings presented that bagging decision trees consistently yields the best predictive accuracy of 86%. It achieves higher precision and recall, while Random Forests demonstrate heightened sensitivity.

In the research outlined in [6], the authors addressed the problem of earthquake forecasting using machine learning. Like [5], this research also mentioned that the data collected for earthquake prediction are imbalanced, as well as most of real-world datasets. This imbalance can lead to a bias in our classification toward the majority class, potentially overlooking the importance of the minority class. After the application of SMOTE, the Decision Tree and SVM are used for the oversampled data. Decision Tree achieved a slightly higher success rate, outperforming SVM by 2%, with a hit rate of 0.86 using the ROC metric. Furthermore, metric MMC is also included in addition to the common evaluation metrics, giving good results even in the initial scenario where negative and positive classes are of very different sizes.

Through a comprehensive review of literature surveys, we have gained a clear understanding of the potential impact of machine learning on natural disaster management. Additionally, the existing works have revealed popular classifiers that can be applied to address our target problem and highlighted potential challenges that we may encounter. This process equipped us with valuable insights toward the achievement of our goals successfully and efficiently.

D. Problem Solution

In our project, we have developed robust models that can be valuable in predicting the Natural disaster occurrences in the future. Our feature set stands unique, which is based on careful correlation analysis, different visualization analysis along with mutual information and domain knowledge. Models are inputted with the intricately crafted feature set. We also used advanced machine learning techniques such as Random oversampling and Cross Validation in an attempt to improve the performance of our models. Additionally, we have formulated ensemble models.

E. Why our Solution is Better

Various methods were employed to address missing values in our original dataset. We filled numerical and categorical

null values with the mean and mode, respectively. To tackle the imbalanced dataset, techniques like Random Oversampling were utilized. Four distinct models were constructed, and their performance was assessed using various metrics. The model's generalizability was tested through cross-validation to ensure its robustness, and ensemble models were also developed.

III. METHODOLOGY

A. Data Collection Process

The dataset is collected from Kaggle. It has historical data records which date from 1900 to 2021. It consists of more than 16000 records and 22 columns. The "ALL NATURAL DISASTERS 1900-2021 / EOSDIS" dataset supports classification. Some of the columns in our dataset are Disaster Subgroup, Disaster Type, Disaster Subtype, Country, Longitude, Latitude, Total Death, No Injured, No Affected, Total Damages, etc.

B. Exploratory Data Analysis

A thorough Analysis of dataset is always required for better understanding of the project scope. We have performed exploratory data analysis on our dataset for this purpose. Firstly, we have plotted a stacked bar plot to visualize the frequency of different disaster types across the continents. The stacked bar plot Fig.1, helped us to compare the frequency of disaster occurrences in various continents. The legend of the bar chart clearly shows the colors assigned to each disaster type so us to help us interpret the stacked bars. For the second visualization we have chosen a Horizontal bar chart, Fig.2, where we analyzed different disaster type frequencies in overall. This visualization helped us to conclude the top 5 disasters types.

Fig. 1

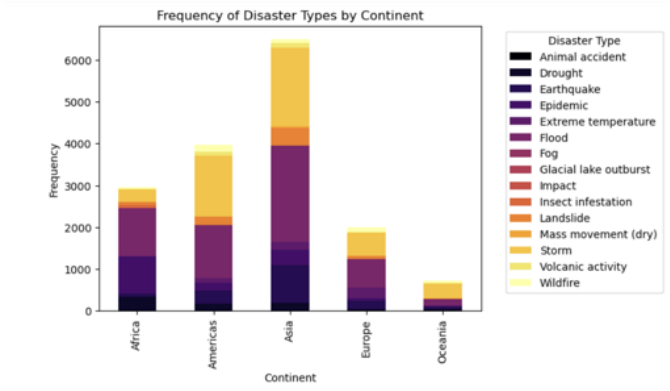

Fig. 1. Frequency of Disaster types

After some basic analysis of the data, we proceeded to the advanced techniques such as correlation analysis. We have plotted a heat map Fig.3, after data preprocessing in where, we have encoded all the categorical features of our dataset. In this heat map we understood the correlation between the target variable and the feature variables. Through this analysis we have gained clarity on variables that are strongly correlated with our target variable such as Disaster Subtype and Disaster Subgroup.

Fig. 2

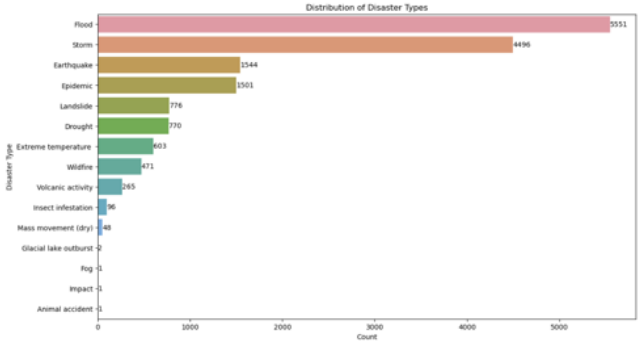

Fig. 2. Distribution of Disaster types

Fig. 3

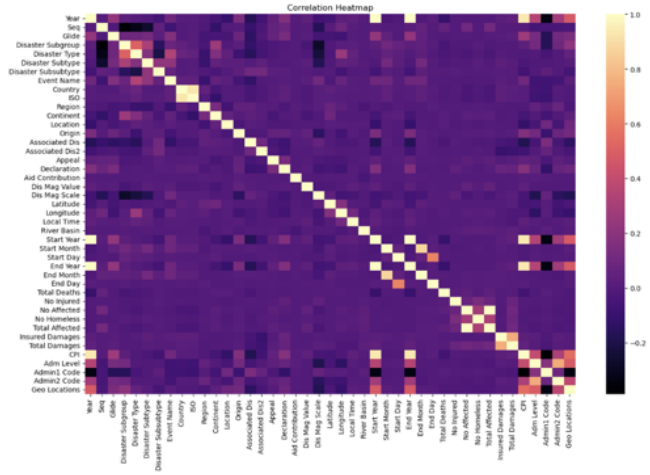

Fig. 3. Correlation Heatmap

Furthermore, we have also analyzed the top 5 disaster types over the years, Fig.4, by plotting a time series line graph. This graph provided us with the insights into how the occurrences of different disasters evolver over the years.

Fig. 4

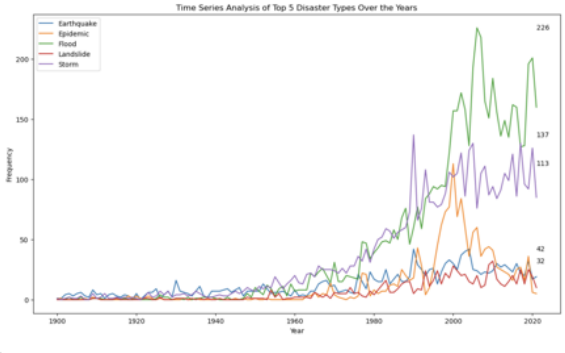

Fig. 4. Top 5 Disaster Types

C. Data Cleaning

Handling Missing values: After the Exploratory data analysis, our primary concern was the missing values present in the data. The count of missing values in some columns was alarming. Deleting the records where the missing values are present leads to data loss. Hence, we imputed the missing values. For the numeric columns containing the missing values, we have imputed using mean of the respective columns and for the categorical columns we have used mode for imputation. This method prevented the project from substantial data loss. The Missing Values are shown in the Figure 5

In Figure 6, it shows the dataset is clean.

D. Data Pre-Processing

Data Encoding: Our target variable is categorical and our dataset contains many categorical columns such as 'Country,' 'Region,' 'Disaster Subtype'. Some Machine learning models could not handle categorical variables. As most of the categorical columns in our dataset are ordinal variables. we have used label encoding to encode our categorical variables. Label encoding converts the categorical variables into numerical format. Most of the categorical columns in our dataset are ordinal variables.

E. Feature Engineering

After the necessary analysis, data cleaning and preprocessing has done on the data we proceeded for feature selection. 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'. Now, we have a preprocessed dataset that is used for the model building.

IV. PROBLEM SOLUTION

In our modeling process, we have explored various classification models that have demonstrated high accuracy in previous studies on disaster prediction. The target variable for our prediction is the 'Disaster Type,' and the selection of features is based on both feature importance and our domain knowledge. Column 'Disaster Type' is the target variable in our prediction, and the selection of features is based on the feature's importance and our domain knowledge. To address data imbalance and enhance model performance, we have employed oversampling techniques. Ensemble methods are applied for the performance comparison. Various evaluation metrics are utilized to conduct a comprehensive examination of these classification algorithms, aiming to capture the complex patterns in the data, improving the robustness of predictive modeling for identifying and classifying different types of disasters.

A. Modeling

Random Forest: Random Forest algorithm is known for its robustness and the ability to handle complex data. Hence, we have chosen Random forest to predict the type of disaster

```
Null values:
Year                      0
Seq                       0
Glide                 14545
Disaster Group            0
Disaster Subgroup         0
Disaster Type             0
Disaster Subtype       3110
Disaster Subsubtype   15049
Event Name            12265
Country                   0
ISO                       0
Region                    0
Continent                 0
Location               1792
Origin                12332
Associated Dis        12778
Associated Dis2       15419
OFDA Response         14432
Appeal                13557
Declaration           12870
Aid Contribution      15449
Dis Mag Value         11180
Dis Mag Scale          1190
Latitude              13397
Longitude             13394
Local Time            15023
River Basin           14839
Start Year                0
Start Month             387
Start Day              3628
End Year                  0
End Month               708
End Day                3556
Total Deaths           4713
No Injured            12231
No Affected            6906
No Homeless           13696
Total Affected         4509
Insured Damages       15030
Total Damages         10881
CPI                     315
Adm Level              8267
Admin1 Code           11545
Admin2 Code           12157
Geo Locations          8267
dtype: int64
```

Fig. 5. Sum of Missing Value

.png

Fig. 6. Clean Dataset

based on the feature set provided. We have chosen parameters such as n_estimators, max_depth, min samples_leaf, random_state to ensure the performance and reliability of the model.

Support Vector Machines: Support vector machines is powerful supervised Machine Learning algorithm used for both regression and classification. In our project we are using SVMs for classification. SVMs used for classification finds the optimal hyperplane that separates different classes in feature space. SVM is efficient on high dimensional data. We have configured the SVM model with linear kenel.

K- Nearest Neighbor: K-NN is a non-deterministic algorithm. It uses multi-layer perception to learn complex functions. It is a simple algorithm which is easy to understand. K-NN follows instance based learning. In our project we have chose some parameters such as number of neighbor, weights, algorithms for more efficient model.

Naïve Bayes: Naive Bayes is a classification algorithm that relies on Bayes' theorem . It has proven to be quite effective for different classification tasks, particularly in situations with less data and high dimensionality. The "naive" assumption of Naive Bayes stipulates that features are conditionally independent given the class.

B. Evaluation of the Models

To assess our model comprehensively, we have chosen the below performance metrics.

Accuracy: One of the most common evaluation metrics. It gives us a picture of the model's correctness. In our class, it refers to correctly predicted disaster types out of the total predictions. While it's widely used, it may not be reliable when the data is imbalanced. The formula for Accuracy is:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

Precision: Also known as Positive Predictive Rate, is a valuable metric in multi-class classification, providing insights into the model's accuracy for each class. In the context of disaster prediction, precision is a crucial metric that capable of identifying actual disaster occurrences accurately, minimizing false alarms. The formula for precision is:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives + False Positives}}$$

Recall: Sometimes called sensitivity, is a metric that gauges the model's ability to capture the positive class. There is a trade-off between precision and recall. In other words, if one metric improves, the other may decline. Recall is generally less sensitive to data distribution compared to precision. Both metrics play a crucial role in evaluating model performance. The formula for recall is:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives + False Negatives}}$$

F1-score: F measure is a metric that considers both false positives and false negatives, striking a balance between precision and recall. It is particularly useful in cases of imbalanced data. The formula for F1-score is:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision + Recall}}$$

The performance of each model is shown in Figure 7.

Performance	Accuracy	Precision	Recall	F1-Score
Random Forest	**0.83**	**0.81**	**0.83**	**0.82**
SVM	0.37	0.16	0.37	0.22
K-NN	0.69	0.67	0.69	0.67
Naïve Bayes	0.14	0.37	0.14	0.16

Fig. 7. Model Performance Without Data Balancing

C. Tuning

After the evaluation of the model we found that our models are not performing well. Random Forest produced F1 score of 82% and accuracy of 83%. SVMs had F2 score of 22% and accuracy of 37%, K-NN's F1 score was 66% and accuracy was 68% and the Navie Bayes has performed the worst by producing an F1 score of 15% and accuracy of 14%. Hence, we have used some techniques to improve the performance of our models.

Balancing the data: First, we checked whether our dataset is imbalanced or not. After checking the count of each 'Disaster type' in the dataset, we concluded that the data is imbalanced. Imbalance of data can be caused when one class is underrepresented, compared to the other classes and this can affect the performance of the models. So, in order to balance the data, we have used Random Oversampler method. This method aims to balance the data by replicating the minority class instances randomly until a balanced distribution is achieved.

In Figure 8, it shows the dataset is balanced.

We have also experimented with the undersampler method as well, but our models performance did not increase instead got decreased. Thus we have come to a conclusion that the undersampling techniques is not ideal for our models. After oversampling is done, we gave the balanced dataset to the models and our model performances increased. Random Forest produced an F1 score and accuracy of 95%. SVMs produced F1 score of 63% and accuracy of 65%. K-NN produced F1 score of 92% and accuracy of 93%. Navie Bayes also performed better with F1 score of 62% and an accuracy of 65%.

```
1     5551
2     5551
13    5551
11    5551
12    5551
5     5551
3     5551
10    5551
14    5551
4     5551
6     5551
9     5551
8     5551
0     5551
7     5551
Name: Disaster Type, dtype: int64
```

Fig. 8. Model Performance Without Data Balancing

D. Ensemble Models

After achieving an encouraging model accuracy, we wanted to apply the Ensemble method, which combines the predictions of multiple models into one model, to enhance the model's robustness. We implemented 2 approaches in ensemble models – hard voting and soft voting. In hard voting, each model gets to vote on the predicted class, and the class with the highest votes gets decided as the predicted class. Hard voting can contribute to a more robust prediction when the single classifier in the ensemble model has a comparable accuracy. In soft voting, instead of making the prediction based on the majority of votes, each model provides a predicted probability for each class, and we take the average probability across all the models as the final prediction. In Hard-Voting we combined 4 individual models (Random forest, SVM, Naïve Bayes, and k-NN) that we had built. For the Soft-Voting approach, we took the Support Vector Machine model as the priority. Both soft Voting and Hard Voting consistently achieve accuracy rates above 92% across all four metrics (F1-score, accuracy, precision, recall). We can see that there is a slight increase in accuracy and F1 score in Ensemble Soft Voting.

E. Model Comparison

In this phase, we have used a grouped bar chart visualization to compare the different models that we have developed, based on their performance metrics. This comparison is done so as to choose the best model so far. By referring to the above Figure 9 we can observer that, Navie Bayes' precision is more than that of SVMs but its overall performance is not good. So we did not consider Navie Bayes'. SVM on the other hand also did not perform well as it was effected by the multicollinearity. Ensemble models and KNN have almost same performance, but the Random Forest out performed all the other models with highest F1 score and accuracy when compared to other models. Thus we have chose Random forest as the best model. The performance of each model is shown in Figure 10:

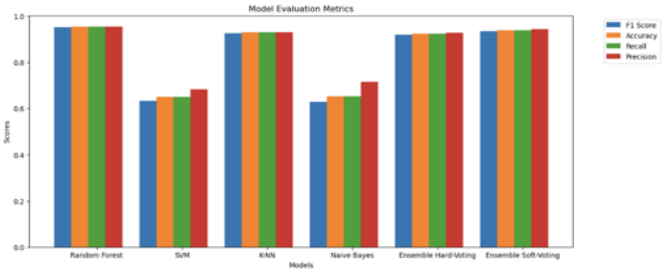

Fig. 9. Evaluation Metrics Comparison

Performance	Accuracy	Precision	Recall	F1-Score
Random Forest	**0.95**	**0.96**	**0.95**	**0.95**
SVM	0.65	0.68	0.65	0.63
K-NN	0.93	0.93	0.93	0.93
Naïve Bayes	0.65	0.72	0.65	0.63
Ensemble Model (Hard Voting)	0.92	0.93	0.92	0.92
Ensemble Model (Soft Voting)	0.94	0.94	0.94	0.94

Fig. 10. Model Performance After Data Balancing

F. Hyperparameter Tuning and Cross Validation

Once we have decided our best performing model, we wanted to validate the generalizability and performance of the model further. To achieve this we have considered two techniques, Hyper parameter tuning and cross validation. In the Hyperparameter tuning we have checked what are the best parameters that can be included in out best model and we got max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 100 We have also performed cross validation on our best model. These two techniques did not increase the performance metric values of our model. So we concluded that the Random forest model is performing well and its best F1 score and accuracy is 95%.

G. Languages and Tools used

We used Python for all aspects of the project including EDA, data cleaning and model development. We utilized Jupyter Notebooks for our project. For implementing different models, we used the scikit-learn python package.

Additionally, we used tools like Trello for Project distribution as well as project status tracking, GitHub for version control, Grammarly for better fluency, GitHub copilot for implementing Pair programming. We used Google Meet for hosting regular project meetings as well as for working together. MS word was used for documentation. LaTeX was used to generate the project report and the presentation was created using Perzi.

V. CHALLENGES AND FURTURE DIRECTIONS

One of our major challenges is data availability and quality. We face difficulties in collecting natural disaster datasets and their corresponding environmental data. Some columns that we

intended to use as features end up with lots of missing values. Hyperparameter tuning is another hurdle in the optimization of our models. Fine-tuning is crucial to enhance the model's effectiveness and robustness. Also, data imbalance is one thing that can not be neglected, we handle our imbalanced data using random oversampling, which might introduce some noises even it hits a favorable accuracy. Achieving an optimal balance between accuracy and mitigating noise proves to be a challenging task when dealing with imbalanced datasets.

For this project, our model has not undergone testing on unseen real-world data. Continuous monitoring and periodic retraining are required in the future. Furthermore, this paper focuses primarily on the initial stage of disasters—prediction and early warning. Future improvements could extend to addressing during-disaster responses, such as crowd evacuation, and post-disaster effects like reconstruction costs, damage recovery, and predictions of total deaths.

VI. CONCLUSION

The outcome of this study can be thought of as a thorough data driven approach to predict natural disasters effectively. Using the results of this study we can say that historical data can be effectively used to predict natural disasters. These models can be used in conjunction with real time weather data to accurately predict natural disasters which will enable a step towards a sustainable future for planet Earth. Also, government officials and agencies can get advance warnings of impending disasters which will help save lives and property.

ACKNOWLEDGMENTS

We would like to thank Professor Dr. Vishnu Pendayala for his constant guidance throughout the course. His dedication of teaching coursework in detail has had a significant impact on our understanding and will undoubtedly shape our careers. We would also like to thank Lohitha Vanteru and Supreetha Naik for their consistent and constructive feedback on our coursework submissions. Their insights played an important role in enhancing our academic growth.

APPENDIX A
CRITERIA MET IN RUBRICS

1. Code Walkthrough: The Jupyter Notebook presents the code in a comprehensive manner, along with clear explanations added as comments. This enhances the readability of the code for others and helps them understand the operations being performed.

2. Presentation Skills: We have created slides that are precise and easy to understand by incorporating appropriate visuals to explain the model's performance. Employed a well-organized narrative supported by evidence from the data and included reflection for continuous improvement. Practiced delivery skills to enhance the timing and overall effectiveness of the presentation.

3. Discussion / Q & A: During the presentation, an open discussion will be strongly encouraged. Participants are welcome to ask any questions they may have throughout the demo. Furthermore, at the conclusion of the demo, a dedicated period will be set aside for a comprehensive Q & A session.

4. Demo: A well-organized demo structure was created, emphasizing the functionality of the working model.

5. Visualization / EDA: In the exploratory data analysis step, by using heat maps, bar charts, and other visualization techniques, we have gained a deeper understanding of our data and identified the features that are most important for modeling.

6. Report: The project report adheres to the IEEE format and is composed in clear, self-written language. It encompasses all necessary information, providing a comprehensive understanding of the problem requirements and the approach taken to address them.

7. Version Control: We have stored all our data, complete code, and a readme file with instructions in publicly accessible GitHub repository. The link to the GitHub repository for our project is https://github.com/daminivichare66/Sustainable-Future-through-Natural-Disaster-Prediction. Additionally, we utilized Trello as a means to monitor and track individual authors' story on their respective stories.

8. Relates to sustainability: Utilizing machine learning technology to forecast natural calamities and partnering with governments, NGOs, and local communities for the formulation of disaster readiness strategies. This initiative contributes towards achieving targeted Sustainable Development Goals such as Climate Action, Sustainable Cities and Communities, Life on Land, Partnerships for the Goals, and Industry, Innovation, and Infrastructure. The project aims to minimize human hardships and financial losses while promoting sustainable progress in an innovative manner that can be expanded easily to wider audiences.

9. Lessons learned: We have learned how to effectively analyze data and gain insights through visualization techniques in order to successfully deploy intricate machine learning models. we developed proficiency in composing reports following the guidelines set forth by IEEE formatting standards. Implementing agile/scrum methodologies in one-week sprints facilitated better project management. We also gained insights into the practical application of innovative tools like GitHub Copilot.

10. Prospects of winning competition / publication: Our project holds strong prospects for winning competitions and securing publications due to its timely and relevant focus on addressing the global challenge of natural disasters. With its emphasis on data analysis, impressive model precision, and clear focus on important sustainability issues, it distinguishes itself as a strong contender within the highly competitive field.

11. Innovation: While most of the existing work focuses on predicting specific disaster types, such as floods, storms, earthquakes, and landslides, our project adopts a broader approach, covering all the natural disaster predictions. We used oversampling skills to tackle the data imbalance to improve the model accuracy. Additionally, we applied ensemble methods that combine four popular and highly accurate classifiers.

12. Evaluation of performance: Model accuracy was evaluated using metrics like precision, recall, and F1 score. Through robust validation techniques such as cross-validation,

we ensured the model's generalizability, and comparisons with baselines highlighted significant performance improvement.

13. Teamwork: Our team actively participated and made valuable contributions throughout all stages of the project. We conducted weekly meetings to assess the progress of our work.

14. Technical difficulty: A significant challenge for us was the availability and quality of data. Many columns which we are thinking of taking as features are filled with missing values. We faced some issues due to the dataset not being balanced. The dataset had a significant amount of data for disaster types like Flood, Storm, and Earthquakes but there was not enough data for other disaster types like Wildfire, Volcano and Drought. This presented a significant challenge which we overcame by using data oversampling methods. Additionally, we faced issues with feature selection and feature engineering which we overcame by using the SelectKBest feature selection technique from the scikit-learn library.

15. Practiced pair programming: Our team has employed GitHub Copilot and engaged in pair programming, resulting in enhanced code quality and improved collaboration skills. We effectively divided tasks using a Trello board, which facilitated our team's success by enabling regular updates.

16. Practiced agile / scrum (1-week sprints): We do our project time management using Trello within the Agile Scrum framework. Five 1-week sprints are completed to address the target problem. Each team member actively contributes to each sprint, and we hold meetings twice a week to stay on track for the final shipping deliverable. Here is the link to access our Trello board: https://trello.com/b/hLDzop7b/data245-machine-learning-project

17. Used Grammarly / other tools for language: Grammarly was used to verify that the project materials adhered to language and grammatical rules. Screenshot is submitted

18. Slides: A detailed presentation was prepared, covering the key aspects of the project.

19. Saving model for a quick demo: We have utilized the Joblib approach to save our trained model. This saved model can be employed for testing on new and unseen data.

20. Used LaTeX: The official IEEE Trans template facilitated adherence to formatting guidelines for font size, margins, and other requirements. Using the LaTeX editor "Overleaf" streamlined collaboration and enhanced the overall writing and editing process.

21. Used creative presentation techniques: Using Prezi tool we have created our presentation and created engaging animations.

22. Literature Survey: A thorough literature survey was conducted, referencing relevant papers and research on natural disaster prediction. This enhanced the overall strength and reliability of our project.

APPENDIX B
AUTHOR CONTRIBUTIONS

All: Collaborated to define the primary objectives and specific goals of the project.

Manisha Lagisetty and Sweekruthi Balivada: Collected relevant datasets for predicting natural disasters, preprocessed the data to address issues like missing values, duplicates and outliers, performed exploratory data analysis to gain valuable insights into dataset characteristics. Identified input features that could enhance model performance and extracted meaningful information from the data to optimize predictive capabilities. Responsible for conducting model development and training by implementing various machine learning techniques such as K-Nearest Neighbor, Random Forest, SVM, and Naive Bayes while ensuring alignment with overall project goals. Additionally, model training was carried out using appropriate algorithms and methodologies.

Yuting Sha and Damini Prashant Vichare: Conducted an analysis of performance metrics, such as accuracy, precision, recall, and F1 score. Employed cross-validation methodologies to validate the model's reliability by addressing overfitting and evaluating its ability to generalize through cross-validation techniques. Responsible for deploying the model and saved it using Joblib approach for performing the testing on new unseen data to ensure functionality and generalization. Created comprehensive documentation covering methodologies, results, and key findings, collaborating with all team members to ensure alignment with the entire project scope.

REFERENCES

[1] Chamola, V., Hassija, V., Gupta, S., Goyal, A., Guizani, M., & Sikdar, B.,*Disaster and Pandemic Management Using Machine Learning: A Survey*. IEEE Internet of Things Journal, vol. 8, pp. 16047 - 16071, 2020, [Online]. Available: https://ieeexplore.ieee.org/document/9295332

[2] Jishnu Saurav Mittapallia, Jainav Amit Muthab, and Maheswari Rc, *NatDisP – An Intelligent Natural Disaster Predictor*. Feb, 2021, [Online]. Available: https://www.researchsquare.com/article/rs-204305/v1

[3] H. A. Shaiba, N. S. Alaashoub, and A. A. Alzahrani, *Applying machine learning methods for predicting sand storms*, in 2018 1st International Conference on Computer Applications Information Security (ICCAIS), 2018, pp. 1–5.

[4] O. Zabihi, M. Siamaki, M. Gheibi, M. Akrami, and M. Hajiaghaei-Keshteli, *A smart sustainable system for flood damage management with the application of artificial intelligence and multi-criteria decision-making computations*, International Journal of Disaster Risk Reduction, vol. 84, p. 103470, Jan. 2023.

[5] D. Stojanova et al., *Estimating the risk of fire outbreaks in the natural environment*, Data Min Knowl Disc, vol. 24, pp. 411-442, 2012.

[6] T. Chelidze, T. Kiria, G. Melikadze, T. Jimsheladze, G. Kobzev,*Earthquake Forecast as a Machine Learning Problem for Imbalanced Datasets: Example of Georgia, Caucasus*,Front. Earth Sci., vol. 10, pp. 847808, Mar. 2022. [Online]. Available: https://doi.org/10.3389/feart.2022.847808

[7] E. Weber et al.,*Detecting natural disasters, damage, and incidents in the wild*,In book: Computer Vision – ECCV, pp. 331-350, 2020

[8] Giovanna Menardi, Nicola Torelli,*Training and assessing classification rules with imbalanced data*,Data Mining and Knowledge Discovery, pp. 92-122, Oct.2012

[9] N. Ahmad et al., *Flood Prediction and Disaster Risk Analysis using GIS based Wireless Sensor Networks, A Review*, J. Basic. Appl. Sci. Res., vol. 3, no. 8, pp. 632-643, 2013. ISSN 2090-4304.