**Sustainable Future through Natural Disaster Prediction**

**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-means clustering, 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.

1. **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 dataset spanning the years 1901 to 2001. 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 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.

* 1. **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.

1. **Theoretical Bases and Literature survey**
   1. **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.

* 1. **Theoretical background of the problem**

------🡪 Have to write

* 1. **Literature Review**

**-------🡪 Have to write**

* 1. **Solution for our problem**

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.

* 1. **Why is 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.

1. **Methodology**
   1. **Data Collection Process**

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. 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.

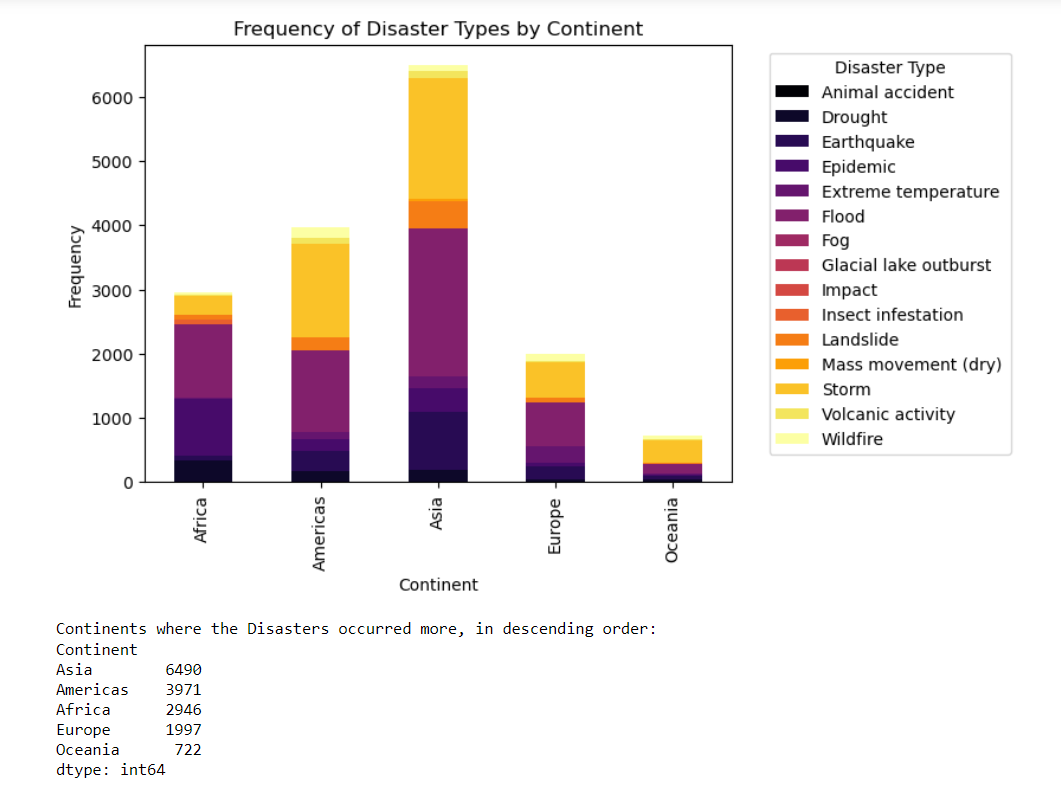
* 1. **Exploratory Data Analysis Process**

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.

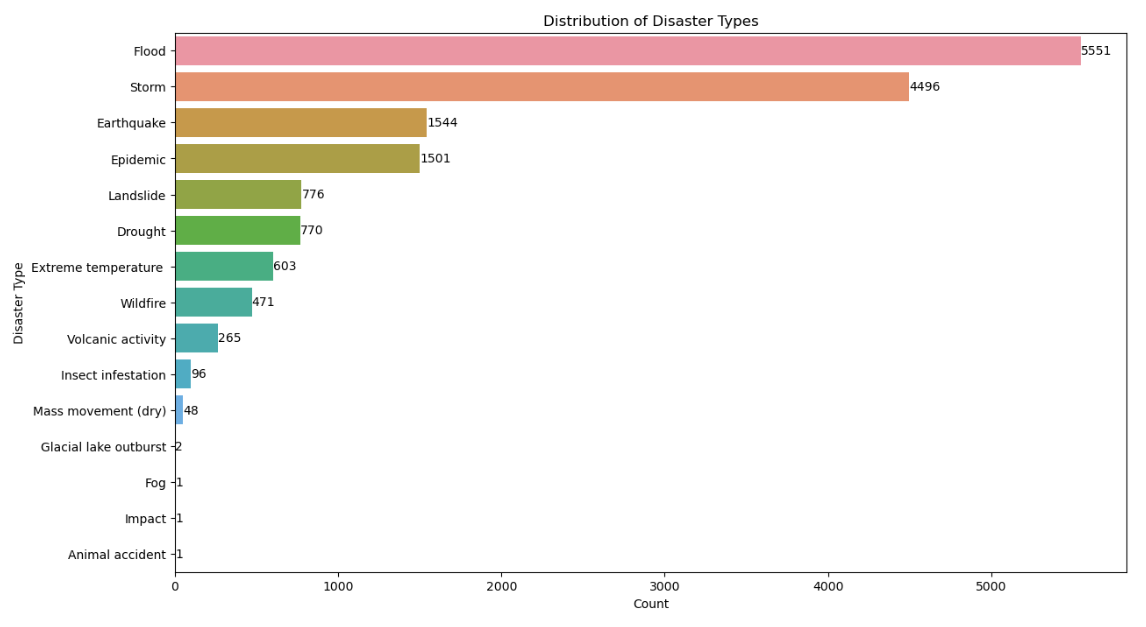
After some basic analysis of the data, we proceeded to the advanced techniques such as correlation analysis. We have plotted a heat map Fig.3, after data preprocessing in where, we have encoded all the categorical features of our dataset. In this heat map we understood the correlation between the target 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.

Furthermore, we have also analyzed the top 5 disaster types over the years, Fig.3, 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.

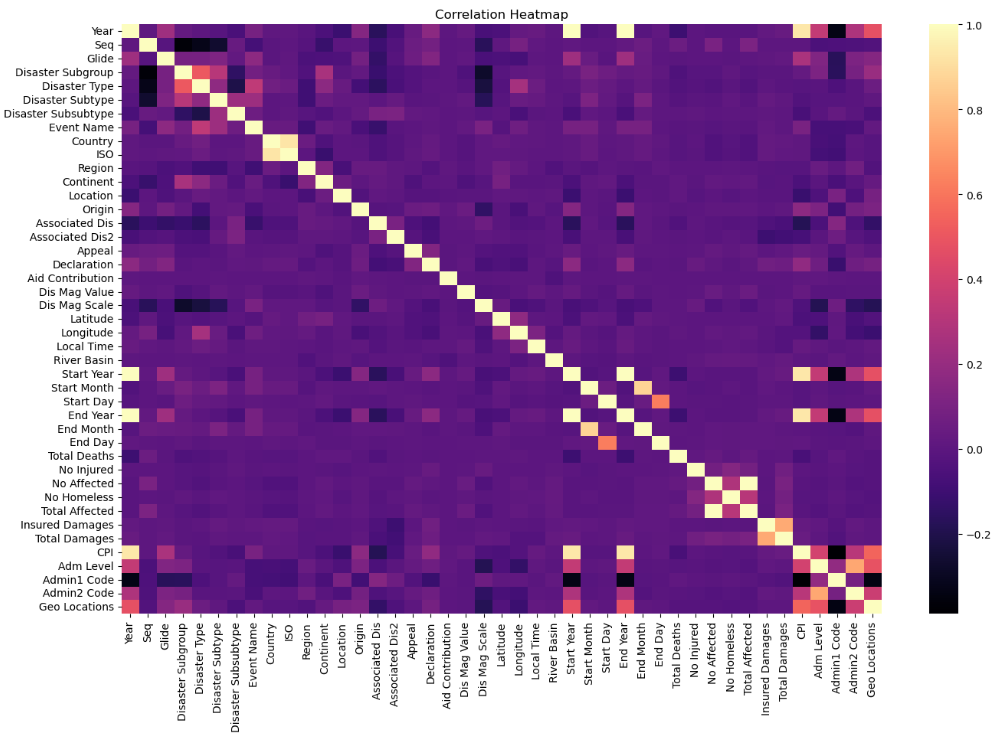
**Fig. 1**



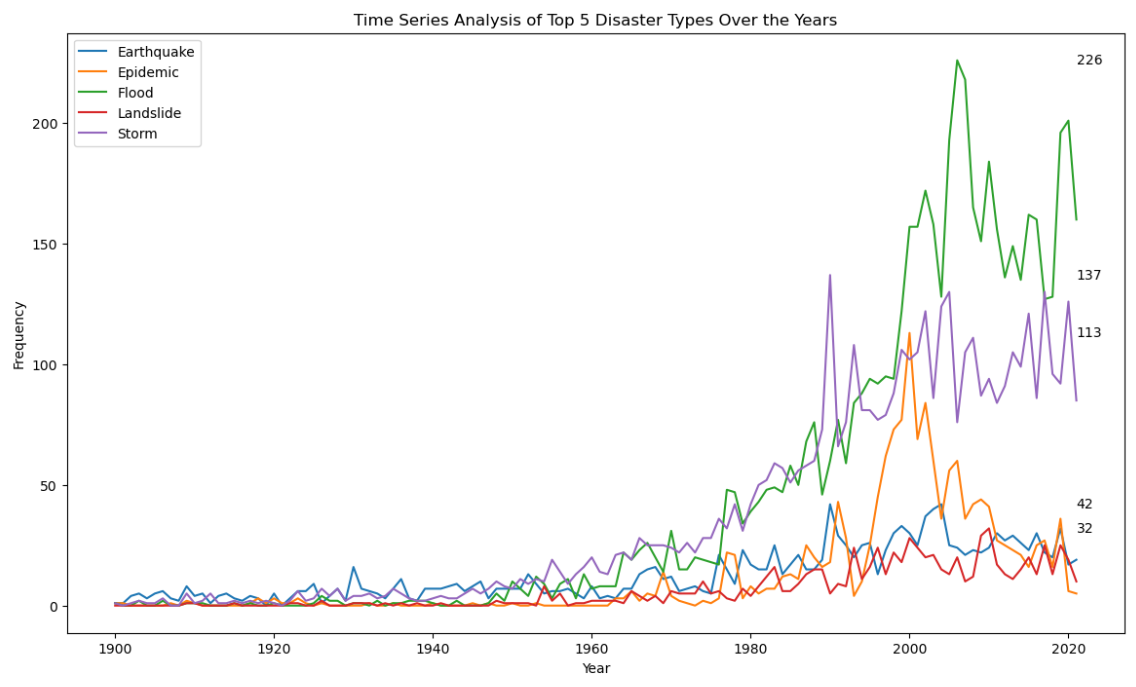
**Fig. 2**



**Fig. 3**



**Fig. 4**

****

* 1. **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.

* 1. **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.

* 1. **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.

1. **Problem Solution** 
   1. **Modeling --------🡪 *Write more if possible, add some more lines but don’t be careful with the redundancy and ai generated text***

For Modelling we have considered different classification models which are built for the prediction of the target variable, ‘Disaster Type’. The comprehensive examination of these classification techniques seeks to capture the complex patterns in the data, improving the robustness of predictive modeling for identifying and classifying different types of disasters.

* + 1. **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 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.

* + 1. **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.

* + 1. **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 n\_neighbor, weights, algorithms for more efficient model.

* + 1. **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.

* 1. **Evaluation of the Models**

For the Evaluation of our models we have chosen the below performance metrics.

* + 1. **F1 score**

**-----------🡪** *write proper (not ai generated) definition and formula for all the metrics*

* + 1. **Accuracy**
    2. **Recall**
    3. **Precision**
  1. **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.

* + 1. **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.

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. **Ensemble Models**

After getting the satisfactory performance metrics, we wanted to build Ensemble models as well. Ensemble models --------------🡪 *See Pendayala’s video and write some text*

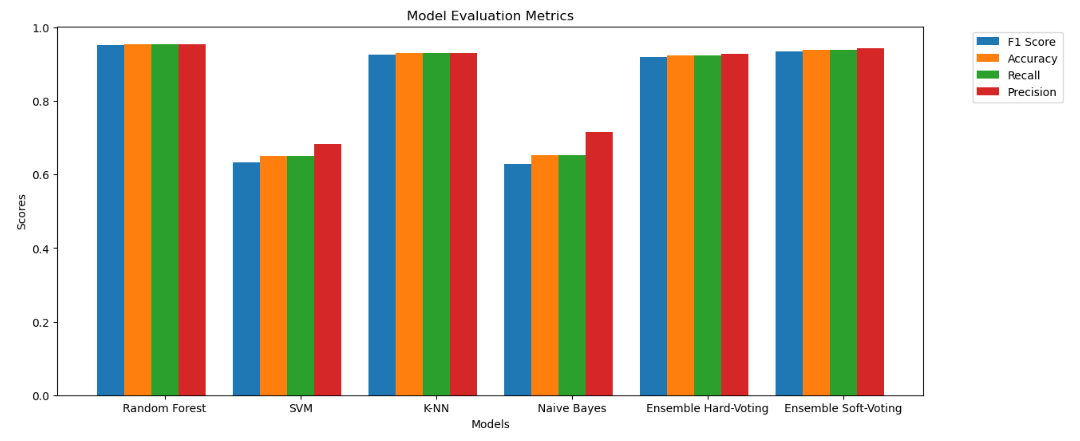
For Hard Voting we combined 4 models

For soft voting we took SVM as the priority

*--------------🡪 Add text and clear explanation here*

* 1. **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. -------🡪 Give Figure number accordingly



By referring to the above figure 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.

* 1. **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%.

* 1. **Languages and Tools used**

Jupyter Notebook was utilized for our project. For implementing different models, sklearn is used. As for the language the code is entirely written in Python.

Used Trello for Project distribution, Github for version control, Grammarly for better fluency, Github copilot for Pair programming. Gmeet for hosting project meetings, working together. MS word for documentation. Perzi for the Presentation.

-----------🡪 Add some more points-----🡪 don’t add too many keep it minimal

**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. Visualizations: 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 -IEEE format: 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. Additionally, we utilized Trello as a means to monitor and track individual authors' story on their respective stories. The project created is named as - <https://trello.com/b/hLDzop7b/data245-machine-learning-project>.

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: ——> WE HAVE TO WRITE THESE BASED ON OUR RESEARCH ON LITERATURE SURVEY\*\*\*\*

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: —> WE HAVE TO WRITE\*\*\*\*

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): —> YUTING’S INPUT\*\*\*\*

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

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

19. Saving the model for 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 LaTeX 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**

Manisha Lagisetty and Sweekruthi Balivada: Collaborated for defining the primary objectives and specific goals of the project. 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.