Predicting Flood Impacts: Analyzing Flood Dataset using Machine Learning Algorithms

Naga Ravindra Babu M[1], B Naresh[2], A Satya Vamsi Kumar[3], G Ganga Bhavani[4], A Sai Ram[5], G Chakradhara Rao[6]

[1]Assoc. Prof., Dept. of Computer Science, B V Raju College, Bhimavaram, AP, India
[2]Assoc. Prof., Dept. of Computer Science, B V Raju College, Bhimavaram, AP, India
[3]Asst. Prof., Dept. of Computer Science, B V Raju College, Bhimavaram, AP, India
[4]Asst. Prof., Dept. of Computer Science, B V Raju College, Bhimavaram, AP, India
[5]Asst. Prof., Dept. of Computer Science, B V Raju College, Bhimavaram, AP, India
[6]Asst. Prof., Dept. of Computer Science, B V Raju College, Bhimavaram, AP, India

---***---

Abstract - *Floods are one of the most destructive and challenging to anticipate natural catastrophes. The frequency or severity of floods has grown in recent years due to climate change and urbanization, as have the associated fatalities and financial losses. Machine learning-based flood forecasting models have slowly started to appear as a result of the fast expansion of computing power. Rich information is included in these models since they were trained on historical data, which is advantageous for data analysis and use. Machine-learning-based models are more effective than the conventional physical flood forecasting model in achieving satisfying results. This study provides a summary of contemporary machine learning-based flood prediction techniques to illustrate recent developments in flood forecasting.*

We list a range of current works in flood prediction and construct the model based on several methodologies.

For flood warnings, flood reduction, or flood prevention, machine learning (ML) models are useful. Machine-learning (ML) addresses have become more well-liked in this regard because to their minimal computing demands and predominance of observational data.

Key Words: Flood, human injure, urbanization, Random Forest, Support Vector, Neural Network etc.

1.INTRODUCTION

The number of natural and man-made disasters has grown globally in recent years. The most common natural catastrophe is a flood, which happens when an excess of water submerges normally dry ground. Floods are typically caused by protracted periods of intense rain, rapid snowmelt, storm surges from tropical cyclones, or tsunamis in coastal areas.

Floods may wreak havoc across a large area, causing fatalities as well as damage to private property and vital public health facilities. Worldwide, more than 2 billion people were impacted by floods between 1998 and 2017. Most at risk from floods are those who reside in floodplains, in non-flood proof structures, lack access to warning systems, or are unaware of the risk of flooding [1].

75% of those who pass away in flood catastrophes drown. Disasters caused by flooding are occurring increasingly often, and this tendency is predicted to continue. Flooding increases the danger of drowning, especially in low- and middle-income nations where residents live in flood-prone locations and flood warning, evacuation, and protection systems are still underdeveloped or insufficient [2].

2. LITERATURE SURVEY

The author of this research analyzed several machine learning-based flood forecasting systems, including linear regression, decision trees, and SVM-based approaches, as well as deep learning-based algorithms like BP and LSTM models. The study shows that the applicability of various approaches varies. Furthermore, since the most recent algorithms are largely influenced by the sophisticated models in deep learning, this paper's conclusion from the current research is that the advancement of deep learning technology has a significant impact on further improving the accuracy of flood prediction performance [3].

Based on historical rainfall datasets spanning 33 years, the goal of this project was to develop a machine learning model that can forecast floods in Kebbi state so that it may be applied to other Nigerian states with high flood risk. In this study, three machine learning algorithms—Decision Tree, Logistic Regression, and Support Vector Classification (SVR)—were assessed and their Accuracy, Recall, and Receiver Operating Characteristics (ROC) scores were compared. When compared to the other two methods, logistic regression yields more accurate findings and offers excellent performance accuracy and recall. The Decision Tree fared better than the Support Vector Classifier as well. Due to Decision Tree's above-average accuracy and below-average recall ratings, it did pretty well [4].

This study introduces an innovative method for using the ensemble model to estimate water level in relation to flood severity. Our method makes advantage of the most recent advancements in machine learning and the Internet of Things (IoT) to automate the analysis of flood data that might be valuable in the mitigation of natural catastrophes. According to research findings, ensemble learning is a more accurate method of predicting the severity of flooding. With sensitivity, specificity, and accuracy of 71.4%, 85.9%, and 81.13%, respectively, the experimental findings show that ensemble learning utilising the Long-Short Term Memory model and random forest outperformed individual models [5].

The various model architectures and their performances are shown in this research. Based on IoT data and weather forecasts, machine learning models using deep learning neural networks have been constructed to identify potential risks [6].

3. PROPOSED METHODOLOGY

Based on your dataset and the goal of flood prediction, here is a proposed methodology for analyzing and predicting flood-related outcomes using machine learning:

Training Phase: The system is trained using the data from the data set and the correct method of model fitting.

Testing Phase: The system is given inputs before its functionality is evaluated. It is put to the test.

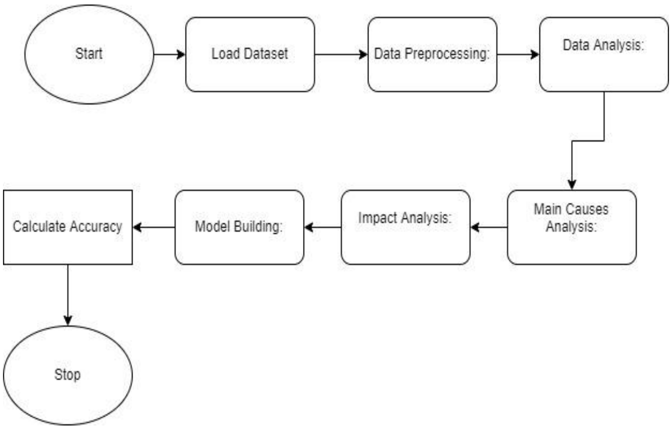

Fig: 1 System Architecture of flood

3.1. TECHNOLOGIES USED:

3.1.1 Python:

For building websites and apps as well as doing data analysis, Python is a widely-liked and approachable programming language. We can readily reuse code thanks to the distinctive features of this high level programming language. Python is not limited to any one area of study and may be used to develop a wide range of original projects and programmes that are dynamic and interesting. Both the dynamic typing and garbage collection are done [7]. Some of its characteristics include simple sharing and collaboration, free access to computer resources.

3.1.2 Google Colab

Python projects and programmes may be written and run using Google Colab, also referred to as Collaborator. We can perform high-level programming, data analysis, and machine learning algorithms thanks to it [8].

3.2 DATASET COLLECTION, PRE-PROCESSING, AND ANALYSIS

3.2.1 Data Collection

In this paper the data is collected from Kaggle. The data set cointains a lot of information like starting date and end date of floods and heavy rainfall occurred. And also the data set cointains the information about number of days rain fall, number humans injured, number of death occurred.

3.2.2 Pre-processing of Data

Pre-processing refers to changing or eliminating unclean and raw data by detecting missing or unnecessary elements of the data and subsequently. Encoding Categorical Variables like "maincause," are encode them into numerical values for the machine learning algorithms to understand. One common approach is one-hot encoding, where each category is represented by a binary value (0 or 1) in separate columns.

3.2.3 Data Analysis

In this dataset, many columns data from across India has collected. By using these columns data it is used to predict the human injured, human deaths, animal deaths can be predicted. By using this data set it is analysied how many human injured, human deaths, animal deaths can be forecasted.

4. IMPLEMENTATION

The model is tested in this stage using the predefined machine learning classifiers. There have been a few models created, and their correctness has been confirmed. For this project, we're using the four classifiers indicated below.

- **Decision Tree:** Calculate the accuracy of the Decision Tree model by comparing the predicted labels to the true labels of the test data. Additionally, you can compute metrics such as precision, recall, and F1-score for a more comprehensive evaluation.

- **Random Forest:** Similar to the Decision Tree, compute the accuracy, precision, recall, and F1-score of the Random Forest model based on the predictions and true labels of the test data.

- **SVM:** For SVM, use the accuracy metric to measure how well the model performs in classifying the test data.

- **Gradient Boosting:** Calculate the accuracy, precision, recall, and F1-score for the Gradient Boosting model using the predicted labels and true labels of the test data.

- **Neural Network:** Evaluate the Neural Network model's accuracy by comparing its predicted labels to the true labels of the test data. Additionally, you can compute other metrics such as precision, recall, and F1-score.

- **Isolation Forest:** For Isolation Forest, the typical evaluation metric used is the Mean Squared Error (MSE). This metric measures the average squared difference between the actual values and the predicted anomaly scores generated by the Isolation Forest algorithm.

5. RESULTS

The data from various cities in India, such as Arunachal Pradesh,Assam,Bihar,Chattisgarh,Himachal Pradesh,Jammu & Kashmir, Kerala, Madhya Pradesh, Maharashtra, Meghalaya, Odisha, Punjab, Rajasthan, Sikkim, Tamilnadu, Telangana, Uttar Pradesh, Uttarakhand, West Bengal, Puducherry, Andhra Pradesh, Chhattisgarh, Gujarat, Jharkhand, Mizoram, Tripura, Nagaland, Andaman & Nicobar Islands, New Delhi, were collected. The data includes different variables, such as start date ,end date, Duration, Main_Cause, Districts, State, Human_fatality, Human_injured, Animal_Fatality, Description_of_Casualties/injured, Extent_of_damage.

The correlation coefficients between two sets of variables are shown graphically in a correlation map.

Some Insights:

- According to a heatmap, Human_injured and Animal_Fatality have a significant positive linear link with a correlation coefficient of 0.83 between them.

- A heatmap's Human_Injured and Main_Cause correlation coefficient of 0.072 indicates that there is a marginally positive linear association between these two variables.

- According to a heatmap, there is a somewhat positive linear link between Animal_Fatality and Human_injured , with a correlation value of 0.83.

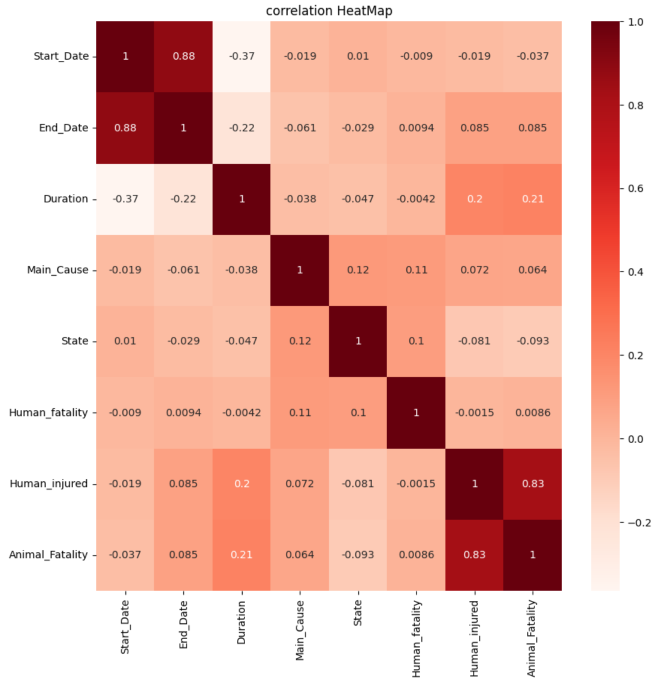

Fig:2 Correlation Heat Map for Flood Dataset

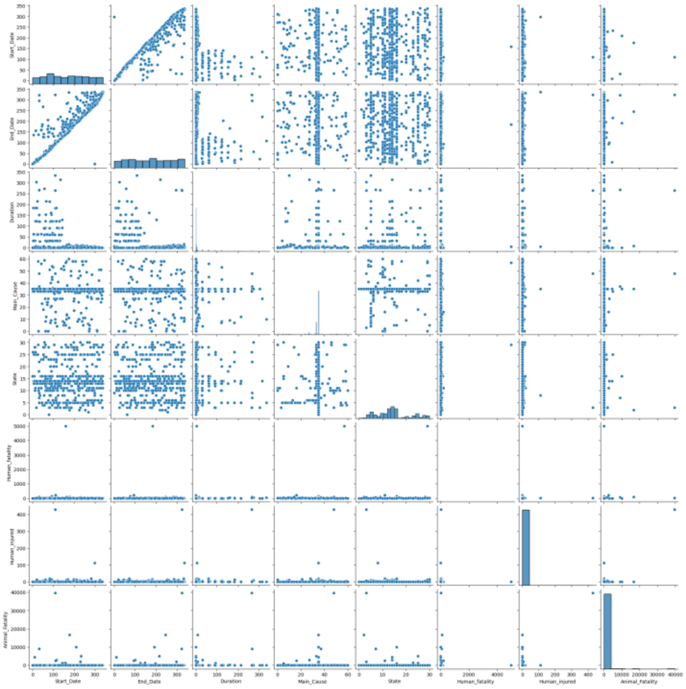

Fig:3 Pair Plot for flood dataset

From the above graph it shown the different columns data in the graph

A pair plot is a graphical representation that allows you to visualize the relationships and patterns between pairs of variables in a dataset. It provides a matrix of scatter plots, where each scatter plot represents the relationship between two variables. Here's a description of the pair plot for the flood dataset columns you provided:

1. 'Start_Date' vs. 'End_Date': This scatter plot will show the relationship between the start and end dates of the flood events. It can help identify any patterns or trends in the duration of the floods.

2. 'Start_Date' vs. 'Duration': This scatter plot will display the relationship between the start date of the floods and their duration. It can reveal whether there is a correlation between the start date and the length of the flood events.

3. 'Start_Date' vs. 'Human_fatality': This scatter plot will show how the number of human fatalities varies with the start date of the floods. It can help identify if there are any time-specific factors that contribute to higher or lower human fatalities.

4. 'Duration' vs. 'Human_fatality': This scatter plot will illustrate the relationship between the duration of the floods and the number of human fatalities. It can indicate whether longer-lasting floods are associated with a higher likelihood of human casualties.

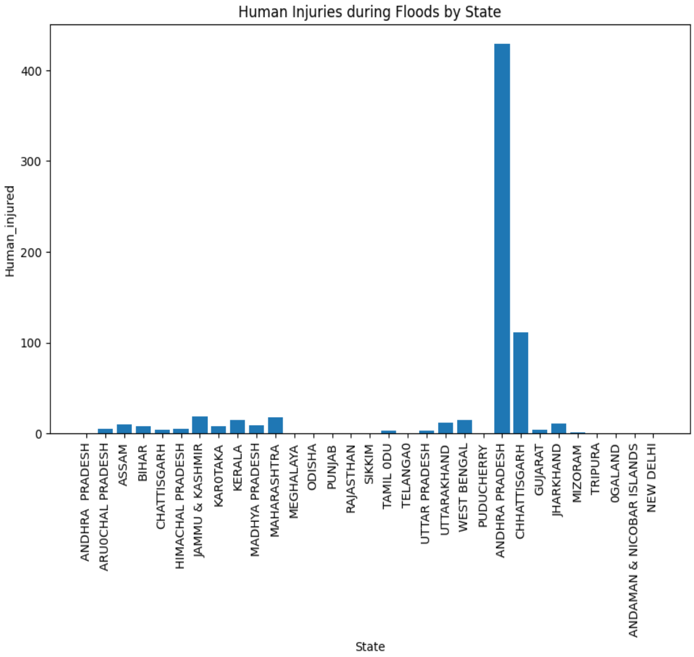

Fig:4 Human Injuries during flood by state

The bar chart displays the number of human injuries during floods, categorized by state. The x-axis represents the different states, while the y-axis represents the count of human injuries. Each bar represents a state and its corresponding value indicates the number of human injuries reported in that state during flood events. The height of each bar represents the magnitude of human injuries, allowing for a visual comparison between states. The bar chart provides insights into the states that experienced a higher number of human injuries during floods, helping identify regions that were more severely affected.

By examining the chart, you can identify states with the highest and lowest counts of human injuries, facilitating a better understanding of the impact of floods on human populations across different regions.

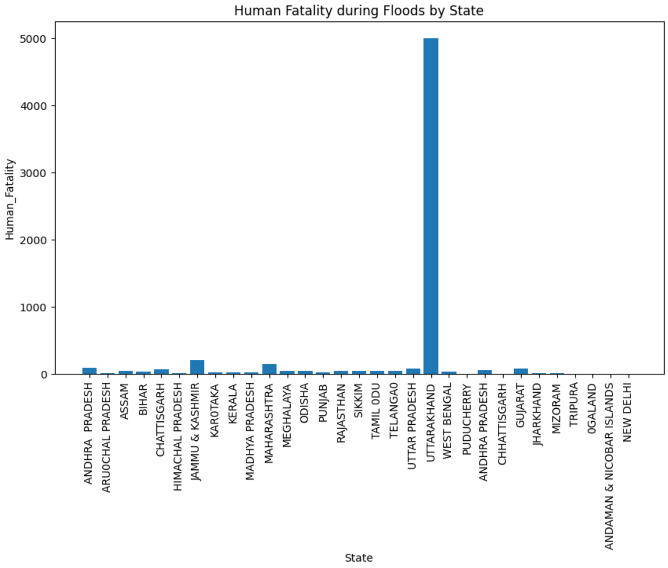

Fig:5 Human Fatality during Floods by State

The bar chart illustrates the number of human fatalities during floods, categorized by state. The x-axis represents the different states, while the y-axis represents the count of human fatalities. Each bar corresponds to a state, and its height represents the number of human fatalities reported in that state during flood events.

The bar chart allows for a visual comparison of the magnitude of human fatalities across different states. It provides insights into the states that experienced a higher number of human fatalities during floods, enabling the identification of regions that were more severely impacted.

By examining the chart, you can identify states with the highest and lowest counts of human fatalities, gaining a better understanding of the impact of floods on human life in different regions.

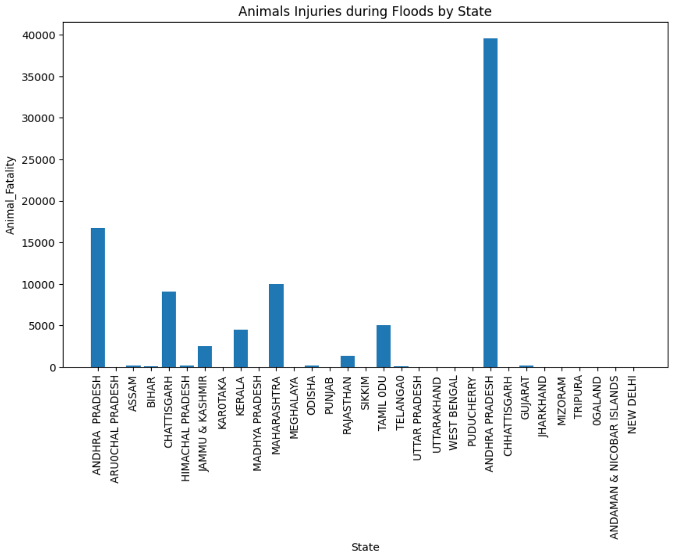

Fig:6 Animals injuries during flood by state

The bar chart depicts the number of animal fatalities during floods, categorized by state. The x-axis represents the different states, while the y-axis represents the count of animal fatalities. Each bar corresponds to a state, and its height indicates the number of animal fatalities reported in that state during flood events.

The bar chart facilitates a visual comparison of the impact of floods on animal life across different states. It provides insights into the states that experienced a higher number of animal fatalities during floods, enabling the identification of regions where animals were more severely affected.

By examining the chart, you can identify states with the highest and lowest counts of animal fatalities, gaining a better understanding of the impact of floods on wildlife in different regions.

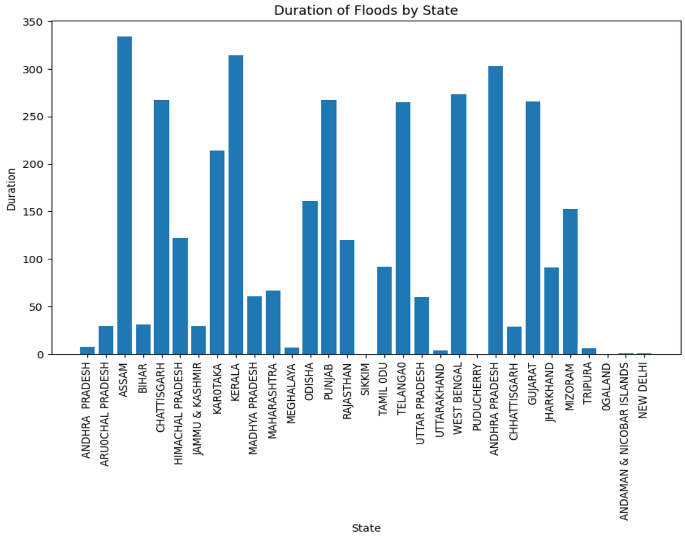

Fig 7: Duration of flood by state

The x-axis displays the different states, while the y-axis represents the duration of the floods in terms of days. Each bar corresponds to a state, and its height indicates the average or total duration of floods reported in that state.

The bar chart allows for a visual comparison of the duration of floods across different states, providing insights into regions that experienced longer or shorter flood events. It helps identify states with the highest and lowest average or total duration of floods, highlighting areas that were more or less affected by prolonged flooding.

By examining the chart, you can identify states with the most extended or shortest durations of floods, facilitating a better understanding of the temporal impact of floods in different regions.

6. COMPARISON

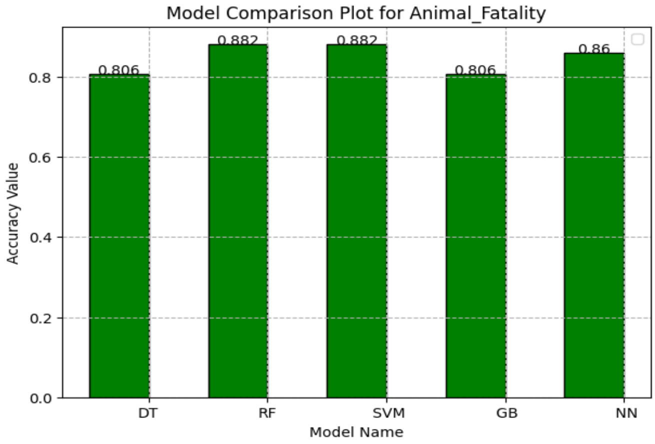

Fig:8 Model Comparison for animal fatality

By observing the below bar chat, the Random Forest Classifier and SVM gives the best accuracy 88.2% among all the remaining algorithms.

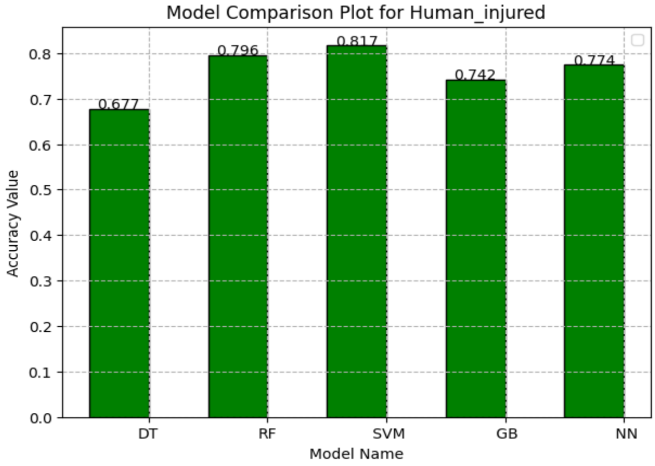

Fig:8 Model Comparison for human injured

By observing the below bar chat, the SVM gives the best accuracy 81.7% among all the remaining algorithms.

7. CONCLUSION

In conclusion, the project demonstrates the effectiveness of machine learning algorithms in analyzing flood datasets and predicting flood impacts. The findings emphasize the importance of using advanced algorithms to inform decision-making processes in flood management and response, ultimately aiding in reducing the adverse effects of floods on human life, animal welfare, and infrastructure.

REFERENCES

[1] World Health Organization. (n.d.). Floods. Retrieved from https://www.who.int/health-topics/floods#tab=tab_1

[2] World Health Organization. (n.d.). Floods. In Health topics. Retrieved Month Day, Year, from https://www.who.int/health-topics/floods#tab=tab_2

[3] Di, Q., Jinbo, Q., & Mingti, C. (2022, October). Application of Machine Learning in Flood Forecast: A Survey. In *2022 International Conference on Virtual Reality, Human-Computer Interaction and Artificial Intelligence (VRHCIAI)* (pp. 177-181). IEEE.

[4] Lawal, Z. K., Yassin, H., & Zakari, R. Y. (2021, December). Flood prediction using machine learning models: a case study of Kebbi state Nigeria. In *2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)* (pp. 1-6). IEEE.

Z. K. Lawal, H. Yassin and R. Y. Zakari, "Flood Prediction Using Machine Learning Models: A Case Study of Kebbi State Nigeria," 2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), Brisbane, Australia, 2021, pp. 1-6, doi: 10.1109/CSDE53843.2021.9718497.

[5] Khalaf, M., Alaskar, H., Hussain, A. J., Baker, T., Maamar, Z., Buyya, R., ... & Al-Jumeily, D. (2020). IoT-enabled flood severity prediction via ensemble machine learning models. *IEEE Access*, *8*, 70375-70386.

[6] Wang, Q. (2022, October). Machine learning model design for IoT-based flooding forecast. In *2022 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)* (pp. 97-103). IEEE.

[7] Python (programming language). (n.d.). In Wikipedia. Retrieved May 8, 2023, from https://en.wikipedia.org/wiki/Python_(programming_language)

[8] Google. (n.d.). Frequently asked questions (FAQ).Google Research. https://research.google.com/colaboratory/faq.html

BIOGRAPHIES

Naga Ravindra Babu M working as an Associate Professor in B V Raju College, Bhimavaram, AP India. He have hands on experience on doing web based projects using PHP and MySQL and ML Projects.

B Naresh working as an Associate Professor in B V Raju College, Bhimavaram, AP India. He have hands on experience on doing web based projects using PHP and MySQL.

A Satya Vamsi Kumar working as an Asst.,Professor in B V Raju College, Bhimavaram, AP India.

G Ganga Bhavani working as an Asst.,Professor in B V Raju College, Bhimavaram, AP India.

A Sai Ram working as an Asst.,Professor in B V Raju College, Bhimavaram, AP India.

G Chakradara Rao working as an Asst.,Professor in B V Raju College, Bhimavaram, AP India.