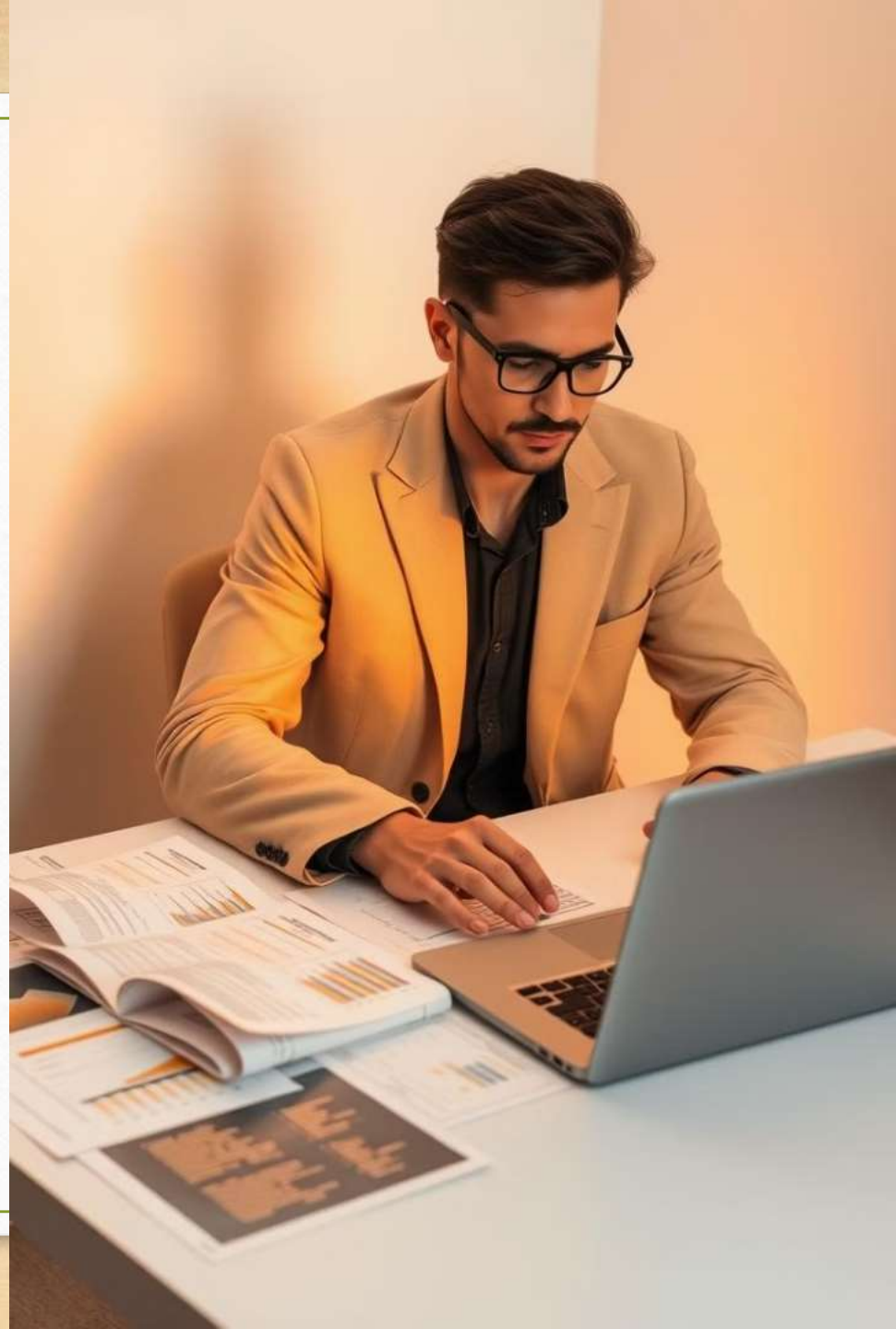
NLP Project: Disaster Tweet Classification

- **NAME: MINAL DEVIKAR**
- **INSTITUTION: DIGICHROME ACADEMY**

PROJECT OBJECTIVE

- **Classify Tweets**: Build a classifier to categorize tweets into Disaster or Not Disaster.
- **Preprocess Data**: Clean noisy text data and extract meaningful features.
- **Deploy Application**: Create a user-friendly Streamlit app for real-time tweet classification.
- **Model Development**: Use a Logistic Regression model for accurate predictions.



Data Cleaning

1. Purpose of Data Cleaning

•To preprocess raw tweet data and remove irrelevant noise for effective feature extraction and modeling.

2. Cleaning Steps

•**Removing Noise**:

- Special characters, emojis, and numbers were removed.
- URLs and hashtags were stripped.

•**Lowercasing**:

- Converted all text to lowercase for uniformity.

•**Stopword Removal**:

- Removed commonly used words like "the," "and," "is," etc., which do not contribute to classification.

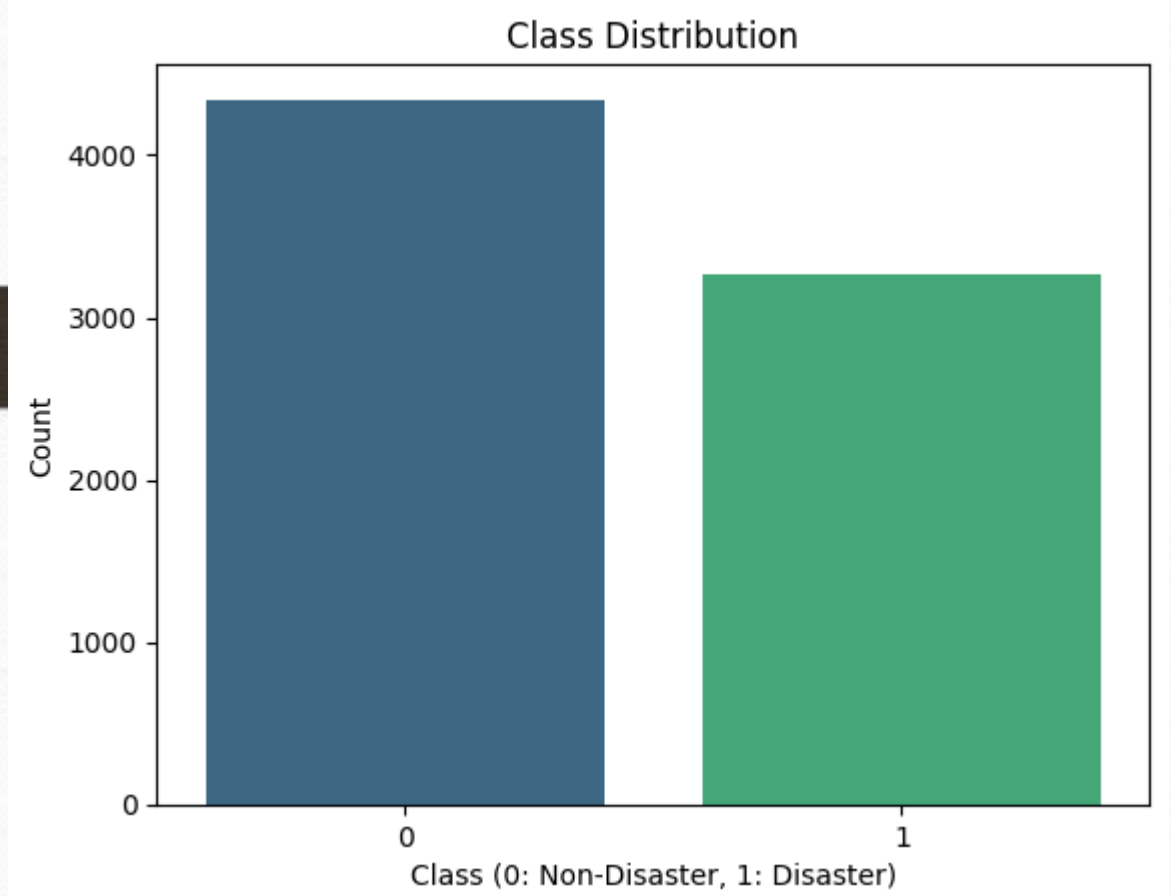
•**Tokenization**:

- Split text into meaningful words or tokens.

Cleaned Data

	text	target	words	numberOfWords	tokenized_words	lemmatized_words	clean_text
3764	there fire catalina look kinda cool pictur doe...	1	[there's, a, fire, in, the, catalinas., looks,...	16	[fire, catalina, look, kinda, cool, pictur, do...	[fire, catalina, look, kinda, cool, pictur, do...	fire catalina look kinda cool pictur doesnt ju...
3368	veldfest announc refund day two extrem weathe...	1	[., @veldfest, announces, refunds, after, day,...	12	[veldfest, announc, refund, day, two, extrem, ...	[veldfest, announc, refund, day, two, extrem, ...	veldfest announc refund day two extrem weather...
994	bomairing elutranscend straight bodi bag	0	[@bomairinge, @elutranscendent, straight, body...	5	[bomairing, elutranscend, straight, bodi, bag]	[bomairing, elutranscend, straight, bodi, bag]	bomairing elutranscend straight bodi bag
1988	complain phoenix mode fire emblem turn ray gig...	0	[and, here, i, was, complaining, about, phoeni...	25	[complain, phoenix, mode, fire, emblem, turn, ...	[complain, phoenix, mode, fire, emblem, turn, ...	complain phoenix mode fire emblem turn ray gig...
765	iphooey time iron michel bachmann brought wro...	0	[@iphooey, @time, ironically, michele, bachman...	24	[iphooey, time, iron, michel, bachmann, brough...	[iphooey, time, iron, michel, bachmann, brough...	iphooey time iron michel bachmann brought wron...

Data Distribution



- **Non-Disaster Tweets (Class 0)**:
- Larger bar, indicating a higher number of non-disaster tweets.
- Represents the majority class in the dataset.
- **Disaster Tweets (Class 1)**:
- Smaller bar, showing fewer tweets related to disasters.
- Represents the minority class in the dataset.

Feature Extraction

1

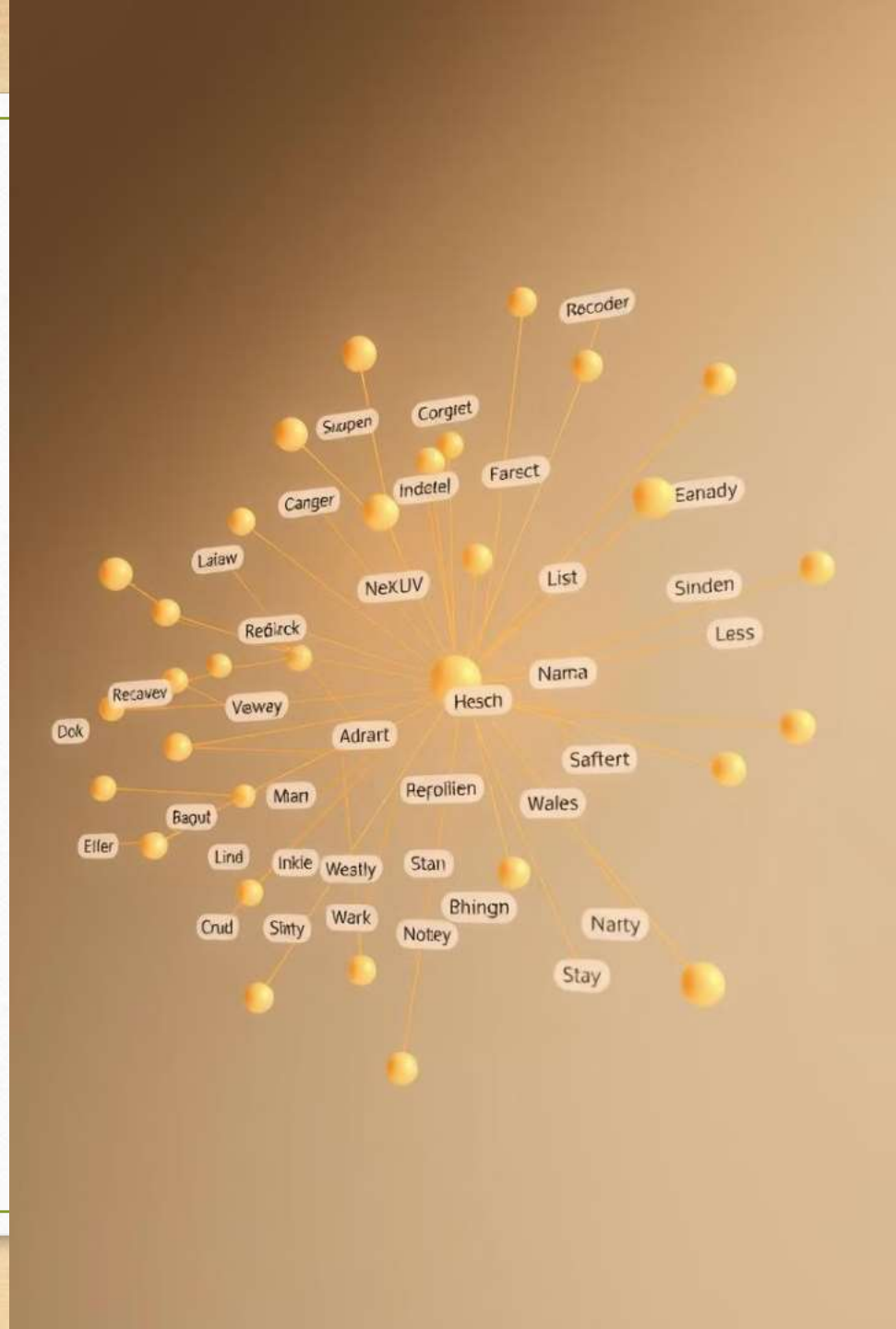
TF-IDF Vectorization

Converts text into numerical features, capturing term importance.

2

Feature Set

6400 features extracted from the dataset.



Word Cloud Interpretation

- **Larger Words**:
- Indicate higher frequency in the dataset.
- Common disaster-related terms: **"fire"**, **"flood"**, **"death"**, **"Hiroshima"**.
- **Smaller Words**:
- Appear less frequently in the dataset.
- Associated with non-disaster contexts.

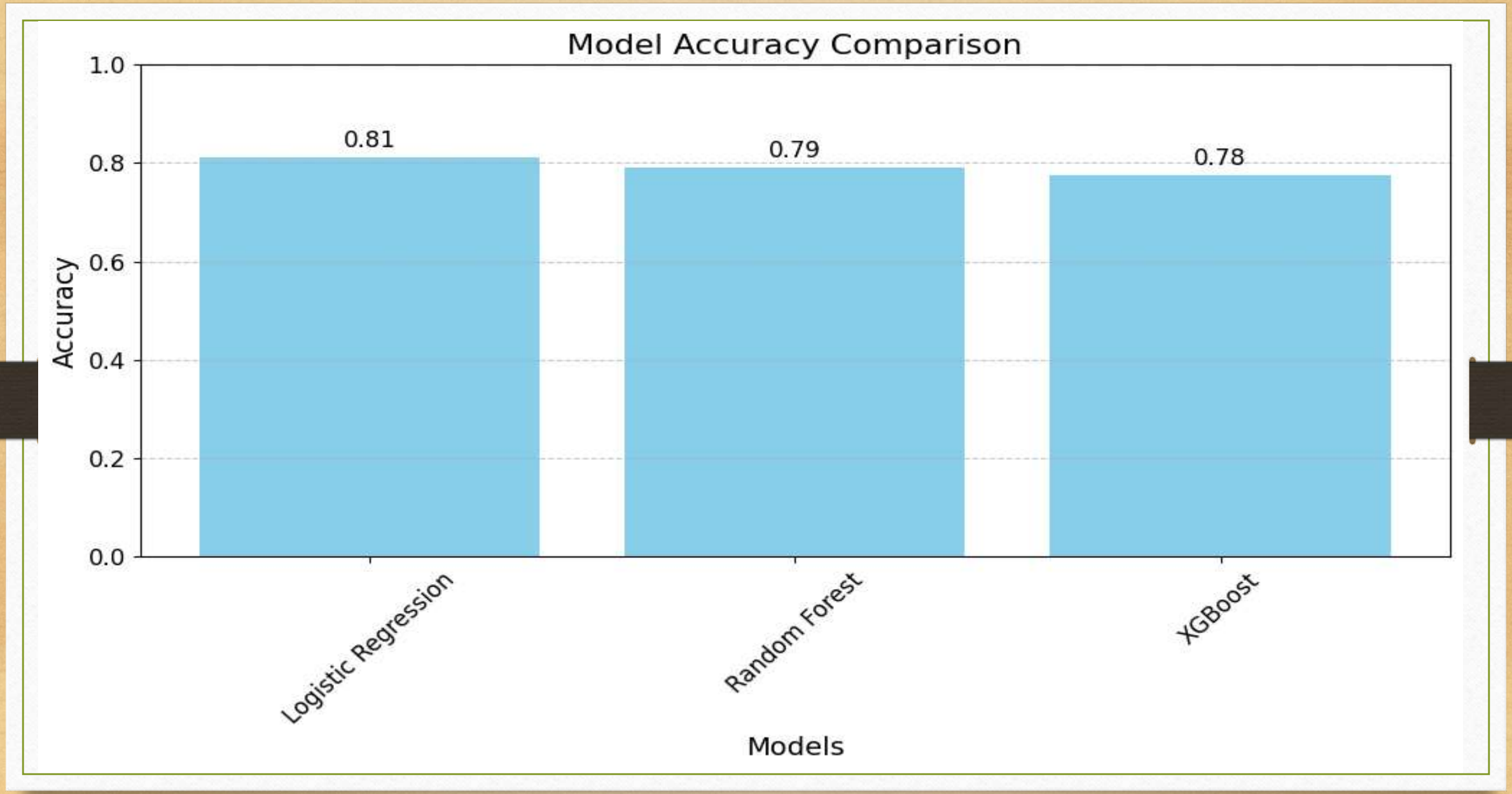
Model Training and Evaluation

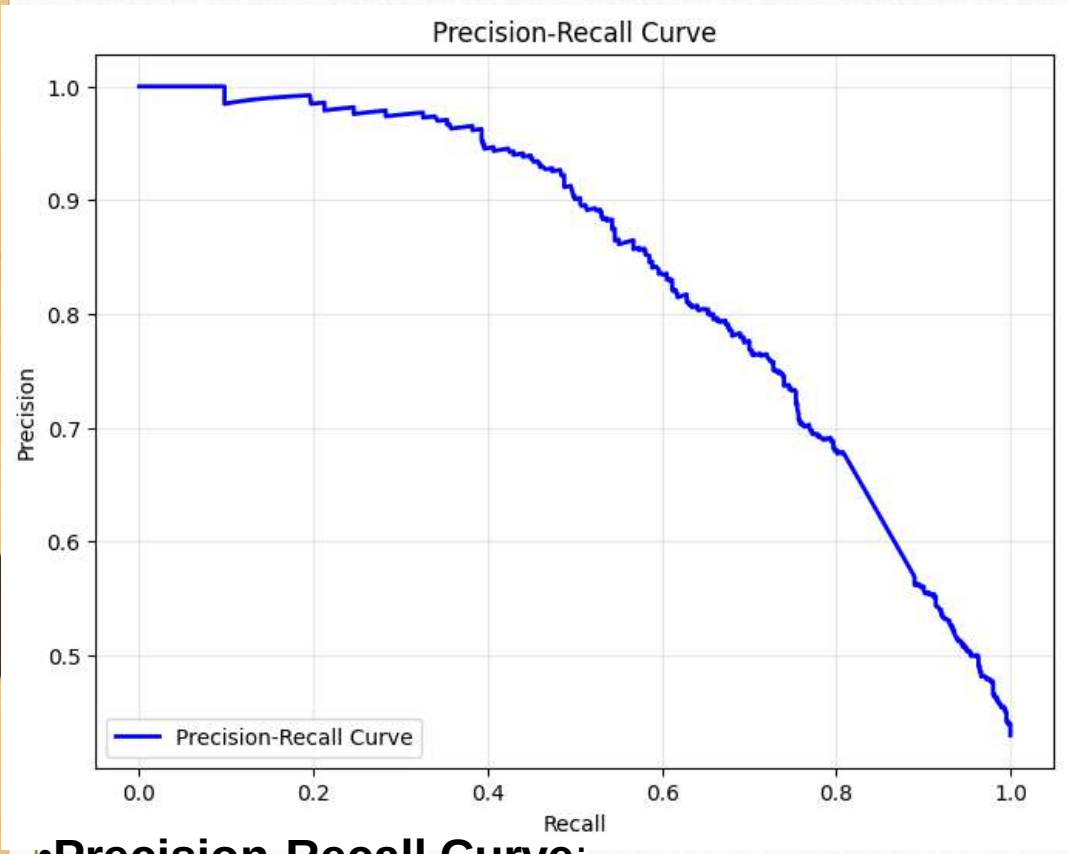
Model Used

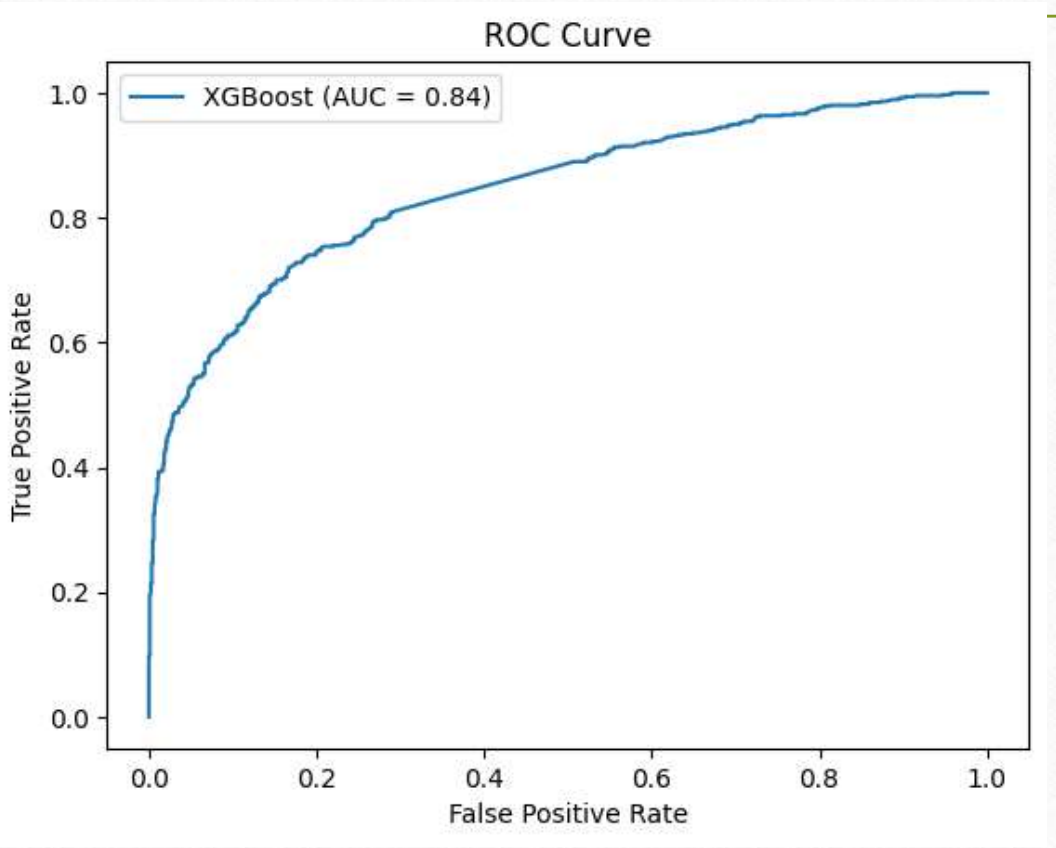
Logistic Regression

Evaluation Metrics

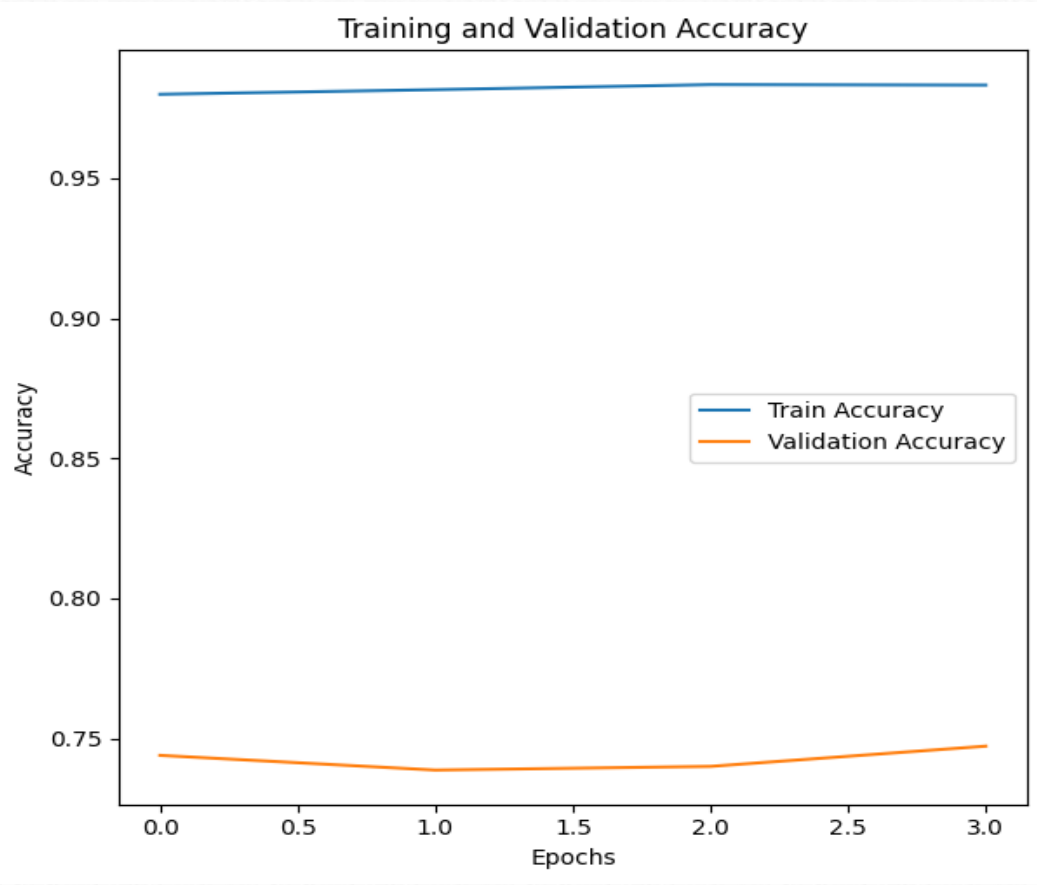
Accuracy: 81.3%, Precision: 81.7%, Recall: 77.0%, F1-Score: 79.2%


Model Accuracy Comparison
1.0
0.8
0.6
0.4
0.2
0.0
Accuracy
0.81
0.79
0.78
Logistic Regression
Random Forest
XGBoost
Models





- **Precision-Recall Curve**:
- Shows the trade-off between precision and recall, with a typical decrease in precision as recall increases in imbalanced datasets.
- **ROC Curve**:
- Plots True Positive Rate (TPR) vs. False Positive Rate (FPR), with an AUC of 0.84 indicating good classification performance.



- **Training Accuracy and Loss**: The training accuracy is consistently high (around 95%), and the training loss remains low, indicating the model fits the training data well.
- **Validation Metrics**: The validation accuracy is significantly lower (around 75%) and shows minimal improvement, while validation loss increases consistently, suggesting overfitting and poor generalization.

App Deployment

App Features

- Accepts tweet input
- Displays predictions

Disaster Tweet

This tweet is classified as **Disaster**.

Project by Minal Devikar

This project aims to classify tweets based on their content into **Disaster** or **Not Disaster** using **Natural Language Processing (NLP)** techniques.

Enter the tweet here:

FloodAlerts tweets flood alerts and warnings for England, Wales and Scotland with local accounts for each County

Predict

Prediction: Disaster

Disaster Tweet

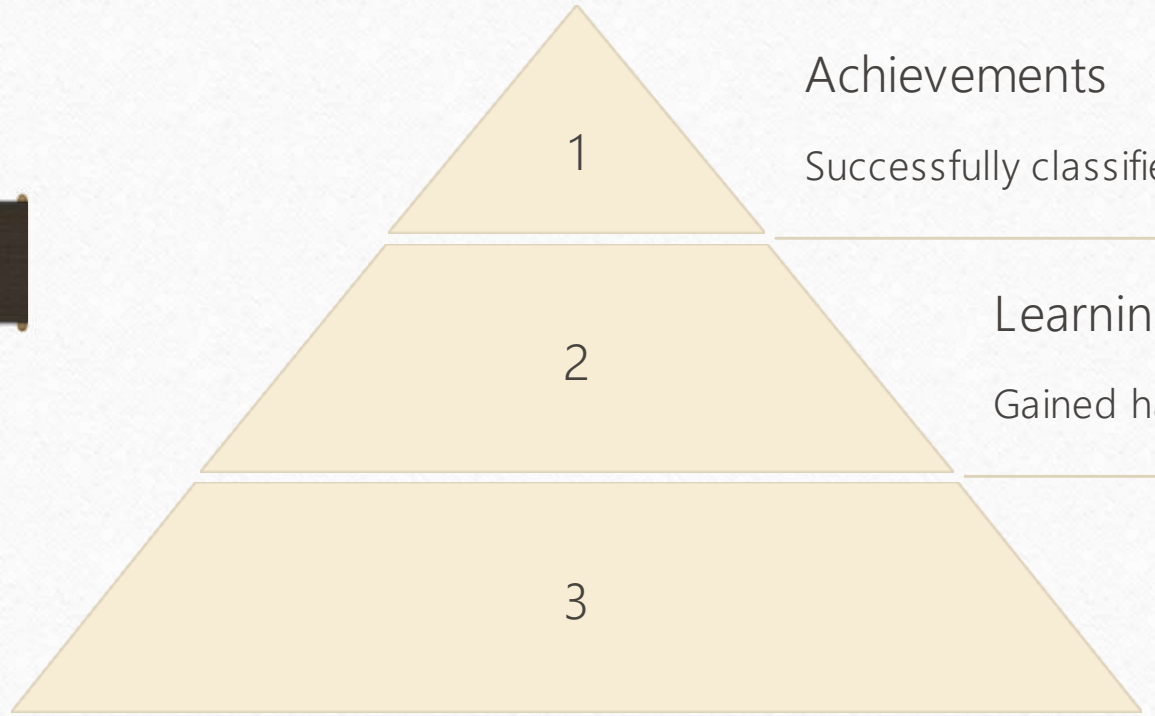
This tweet is classified as **Disaster**.

Project by Minal Devikar

This project aims to classify tweets based on their content into **Disaster** or **Not Disaster** using **Natural Language Processing (NLP)** techniques.

CONCLUSION

- **Achievements**:
  - Successfully classified tweets into Disaster and Non-Disaster categories.
  - Deployed a functional and user-friendly app.
- **Learnings**:
  - Gained hands-on experience in text preprocessing, feature extraction, and model training.
  - Deployed a web-based NLP solution.
- **Future Work**:
  - Extend classification to multi-class disasters (e.g., floods, earthquakes, fires).
  - Experiment with advanced models like **BERT** for better accuracy.
  - Integrate real-time tweet scraping for live predictions.



Achievements

Successfully classified tweets.

Learnings

Gained hands-on experience.

Future Work

Extend to multi-class disasters.