Flood Detection and Prediction Using Satellite Imagery and Machine Learning

DATA 270 – Project Presentation

Professor: Dr. Eduardo Chan
San Jose State University

Saniya Lande, Shilpa Shivarudraiah, Iqra Bismi, Divya Nalam

(Project Group- 4)

Project Motivation & Needs

- Floods are the most frequently occurring natural disasters (46%) resulting in human suffering, human loss, property damage.
- People have suffered due to the lack of a proper flood management system which indicates the need for flood forecasting.
- To reduce vulnerabilities and loss of lives, flood forecasting is crucial, for which we are implementing machine learning algorithms to predict floods accurately.

Targeted Problem

- The flood prediction model is complex due to the consideration of various factors like geographical location, hydrology and human activities.
- Previous research paper review: Supervised Machine learning algorithms can face challenges in predictions due to complexity of the data, the amount of time required for computation.
- Goal - To implement supervised as well as unsupervised machine learning models on satellite imagery, tidal, rainfall data to understand which method is more robust and gives better flood prediction accuracy.

Deliverables

- A machine learning model prototype being developed using SVM, Random Forest, Ensemble and K-means.
- Comparison of models for their accuracy to decide on which model performs better in predicting floods accurately in timely manner.
- Detailed report containing the project background, scope, literature survey, EDA visualizations, charts , figures describing the model and the technical aspects of the project will be delivered.

Project Requirements

Functional Requirements

- ❑ The system should be able to predict a flood based on the available data immediately.
- ❑ It should generate alerts and forward them to the respective stakeholders.
- ❑ Regarding data, no legal or regulatory requirements needed as data is extracted from open sources.

AI Requirements

❑ *Support Vector Machine Classifier*

No assumptions about data, handles outliers.

❑ *Random Forest Classifier*

Eliminates overfitting.

❑ *XGBoost Model*

Works well on large datasets, eliminates overfitting.

❑ *K - Means Clustering*

Works with unlabelled data.

Data Requirements

- ❑ INSAT Images extracted with Python Script from IMD website. https://mausam.imd.gov.in Contains data such as Visual, IR, time and water-vapor.
- ❑ Weather Data from OpenWeather API. https://openweathermap.org
- ❑ Tidal Data such as height and state collected from MareaAPI https://api.marea.ooo/

Technology Survey - Comparison among different models

Models	Type of Data	Data pre-processing technique	Approach	Performance metrics
SVM	Historical Rainfall data (numeric)	SMOTE	Single SVM model	RSME 0.05 R^2 0.91
SVM-FA	Monthly river flow data (numeric)	Not Specified	SVM combined with Firefly algorithm	RSME 0.04 R^2 0.98
WSVM	Historical monthly, daily Rainfall data (numeric)	SMOTE	Hybrid Wavelet input to SVM model.	RSME 0.05 R^2 0.92
XGBoosting	hydrological data (numeric)	Normalization using Min Max Scaler	Single XGboosting model with learning rate 0.01	RMSE: 0.0089 R2: 0.9945
Random Forest	Sentinel-1 SAR data & Sentinel-2 optical data	Sentinel-1 preprocessing: DS1 & DS2 data, TOPSAR-split, S1 slice assembly,Radiometric Calibration, C2 matrix generation, Polarimetric Speckle Filter, Terrain Correction Sentinel 2 preprocessing: DS3 & DS4 data, cloud masking	Random Forest classifier with fused Sentinel -1 & Sentinel -2 data	Classification results obtained through images and further analyzed visually. Fusing datasets yield most accurate results
Random Forest	Sample points collected through site visits, NDWI and sample points generated from Sentinel-2 satellite image	-	Random Forest algorithm implemented in Rstudio	97.57% overall accuracy and 95.14% Kappa coefficient
Incremental k-means clustering	Air Pollution dataset (dynamic in nature) of West Bengal, India for years 2009 and 2010	Removal of NaN values or other invalid data points	Incremental k-means clustering applied to means	Accuracy of approximately 83.3% achieved through the model
K Means Clustering	Factors influencing low elevation like water inflow, etc. are considered and used.	Removal of NaN values or other invalid data points	K means used to categorize drought levels and fitting functions implemented to determine relationships between parameters (i.e. weights) and to find elevation though volume	Assumptions regarding the environment like flat lake surface and others are made initially. The model can be used for general estimations but not for technical or serious purposes

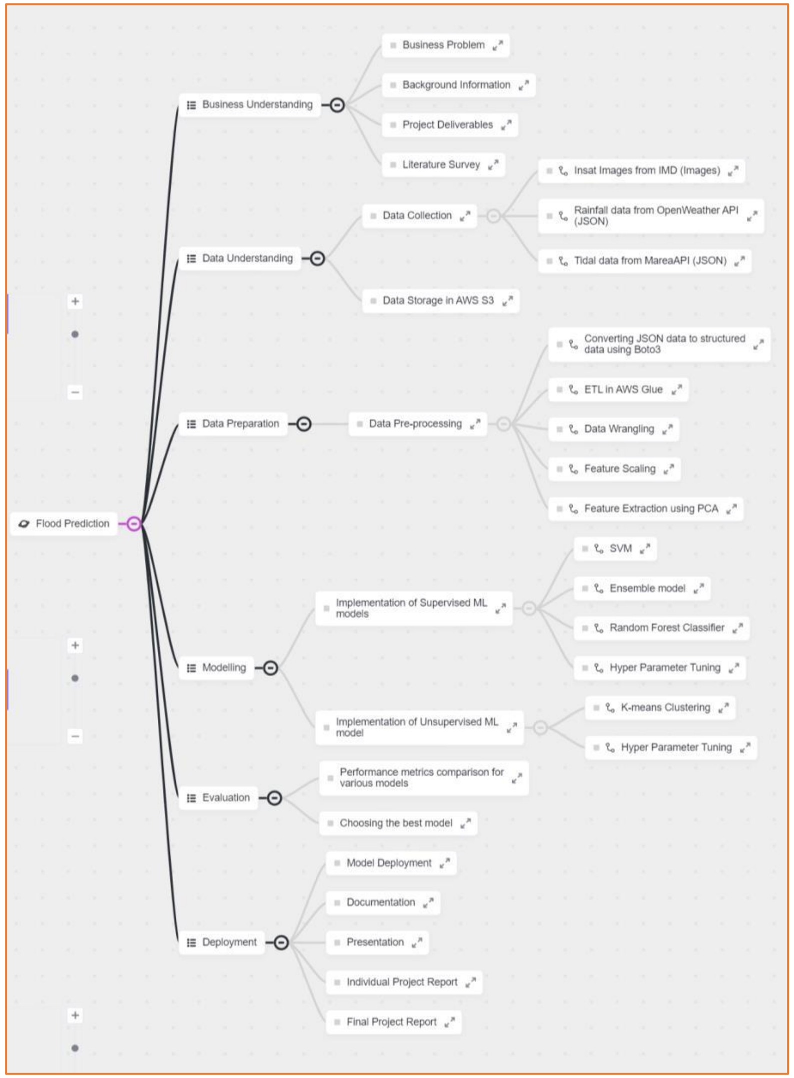
Literature Survey - Models summary

Previous papers Authors	Models	Type of Data	Data pre-processing technique	Performance metrics
Dai et al.	LSVM	Hydrological data captured from IoT devices	Flood intensity classification and K-fold cross validation	RSME = 0.3991 $R^2 = 0.9943$
	QSVM			RSME = 0.3432 $R^2 = 0.9972$
	RF			RSME = 0.1571 $R^2 = 0.9559$
Prasad et al.	Random Forest Regressor	Rainfall level(numeric)	Used statistical methods to clean data	R2 = 1.00
Ghorpade et al.	Decision Tree	hydro data (numeric)	-	RMSE= 0.216; R2= 0.9945
Jain et al.	Deep Convolutional Neural Networks	MediaEval 2019 flood dataset-Sentinel-2 data (image)	-	F1 - 0.96 Kappa - 0.92
Moumtzidou et al.	Deep Neural Networks	Sentinel-2 data (image)	-	F-Score 62%

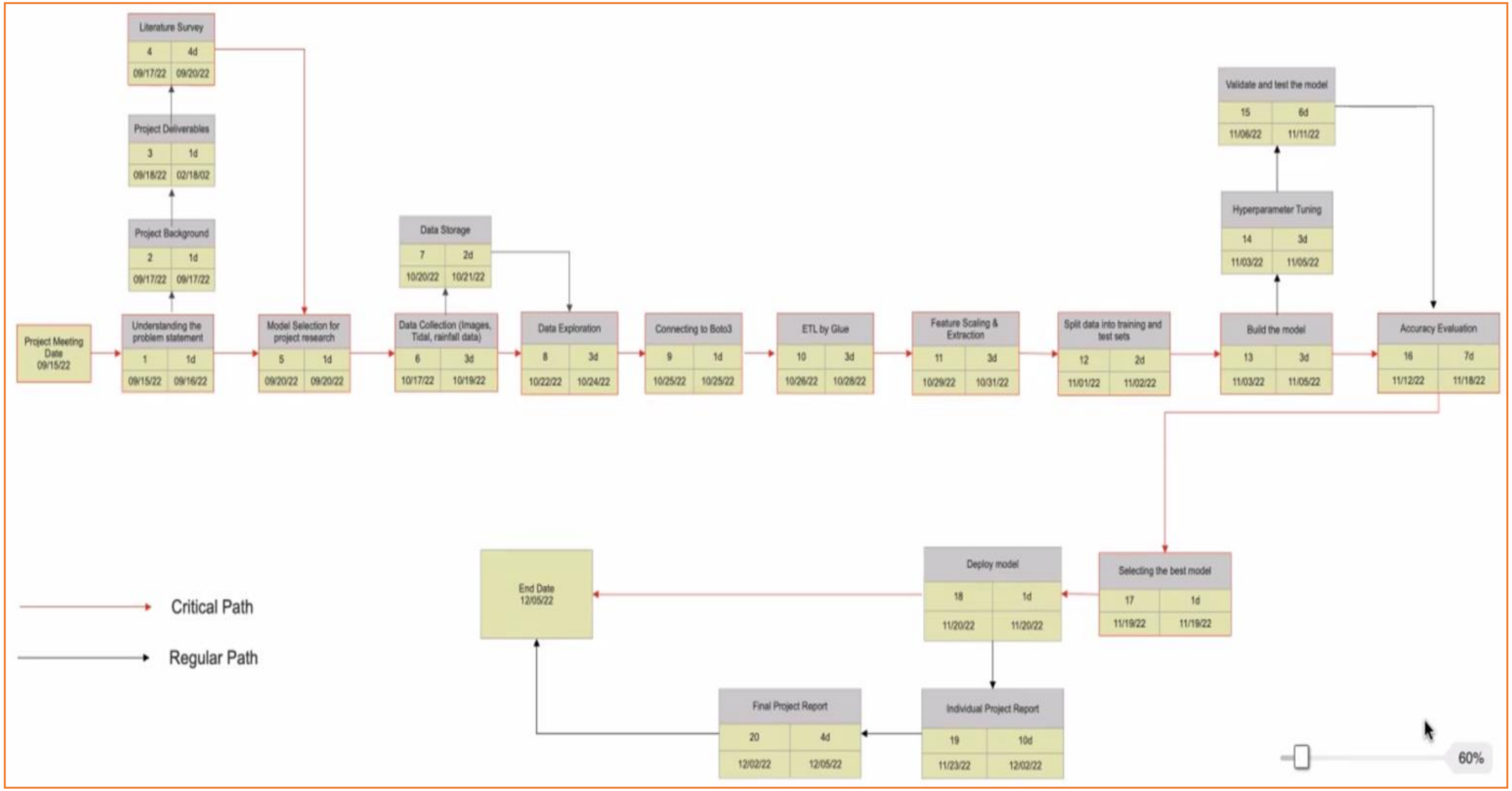
Tanim et al.	RF	Data From road closure reports along with SAR images for flood detection	Image pre-processing done using SNAP tool for noise removal and filtering.	Precision ~ 0.5 Accuracy ~ 0.65
	SVM			Precision ~ 0.85 Accuracy ~0.85
	MLC			Precision ~0.75 Accuracy ~0.82
	Unsupervised (CD, iso-clustering, fuzzy rules)			Precision ~ 0.8 Accuracy ~ 0.84
Razafipahatelo D. et al.	Digital Elevation Model (DEM), kernel k-means, non - linear clustering	SAR Images of flooded regions	-	Manual result = 807.63 ha Log Ratio in fs = 844.74 ha Ratio in fs = 1429.27 ha Ratio = 682 ha
Xu H. et al.	Urban flood inundation model, k-means clustering and improved entropy weight method	Information about river, rains and drainage, previous storm records and physical parameters like slope, distance from river etc,	Removal of invalid data	High risk zones overlap when compared with the references therefore making it a feasible approach (but limited by data provided).

Project Plan – Work Breakdown Structure

- The WBS has been made using Click up- Mind map and adheres to CRISP-DM methodology which involves six phases in successfully completing an ML project.
- Business Understanding: Motivation, Goals, Needs, Targeted Problem
- Data Understanding: Data requirement, Data collection approaches
- Data Preparation: Data cleaning, Data Transformation using Python
- Modeling: Build Machine learning models i.e., SVM, Random Forest, Ensemble, K-Means
- Evaluation: Assess the models using evaluation metrics like accuracy, recall, precision and F1 score
- Deployment: Deploy the machine learning model to end users

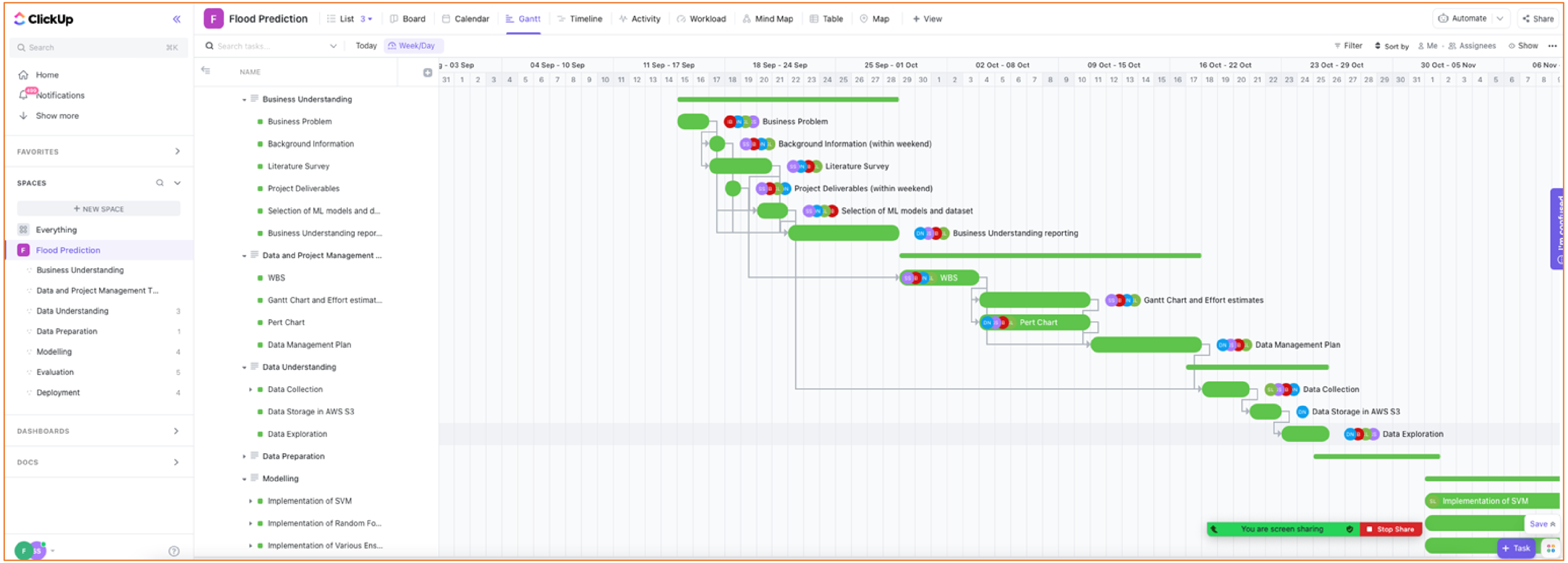


Pert Chart



Literature Survey
4
4d
09/17/22
09/20/22
Project Deliverables
3
1d
09/18/22
02/18/02
Project Background
2
1d
09/17/22
09/17/22
Project Meeting Date 09/15/22
Understanding the problem statement
1
1d
09/15/22
09/16/22
Model Selection for project research
5
1d
09/20/22
09/20/22
Data Storage
7
2d
10/20/22
10/21/22
Data Collection (Images, Tidal, rainfall data)
6
3d
10/17/22
10/19/22
Data Exploration
8
3d
10/22/22
10/24/22
Connecting to Boto3
9
1d
10/25/22
10/25/22
ETL by Glue
10
3d
10/26/22
10/28/22
Feature Scaling & Extraction
11
3d
10/29/22
10/31/22
Split data into training and test sets
12
2d
11/01/22
11/02/22
Build the model
13
3d
11/03/22
11/05/22
Hyperparameter Tuning
14
3d
11/03/22
11/05/22
Validate and test the model
15
6d
11/06/22
11/11/22
Accuracy Evaluation
16
7d
11/12/22
11/18/22
Selecting the best model
17
1d
11/19/22
11/19/22
Deploy model
18
1d
11/20/22
11/20/22
Individual Project Report
19
10d
11/23/22
12/02/22
Final Project Report
20
4d
12/02/22
12/05/22
End Date 12/05/22
Critical Path
Regular Path
60%

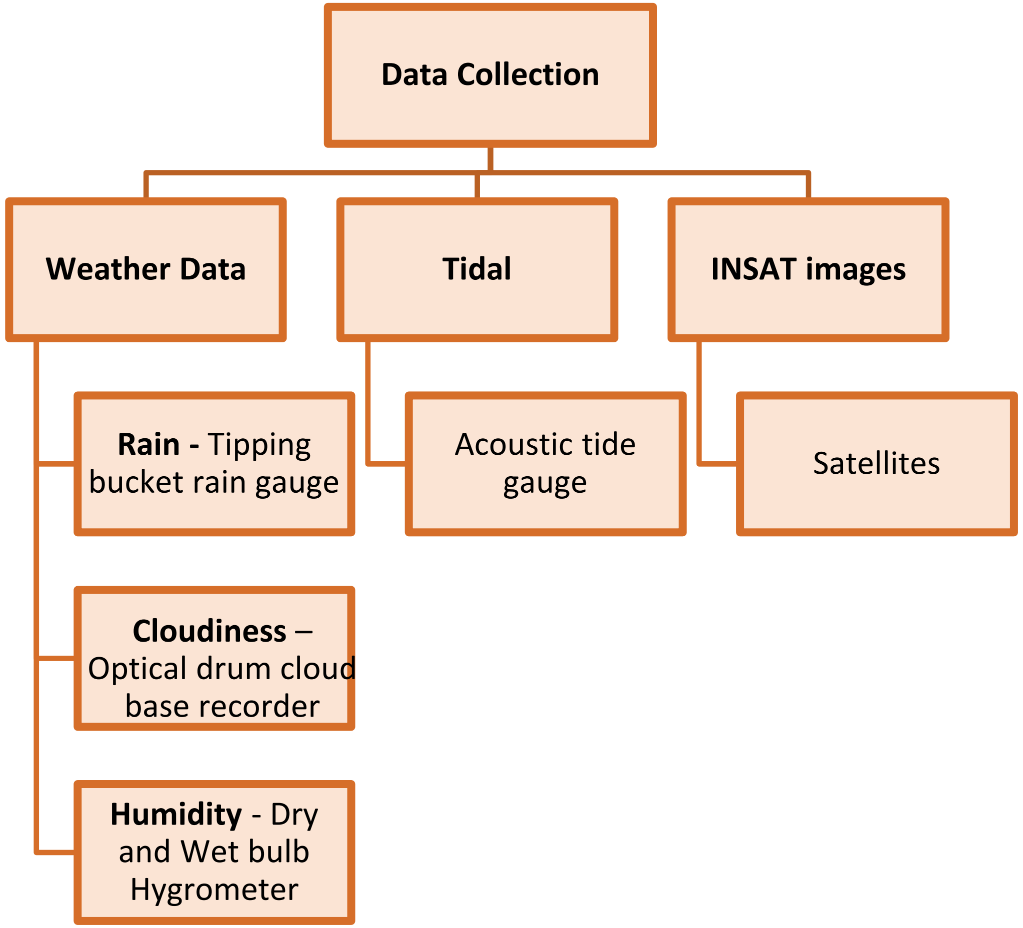
Project Schedule – Task breakdown

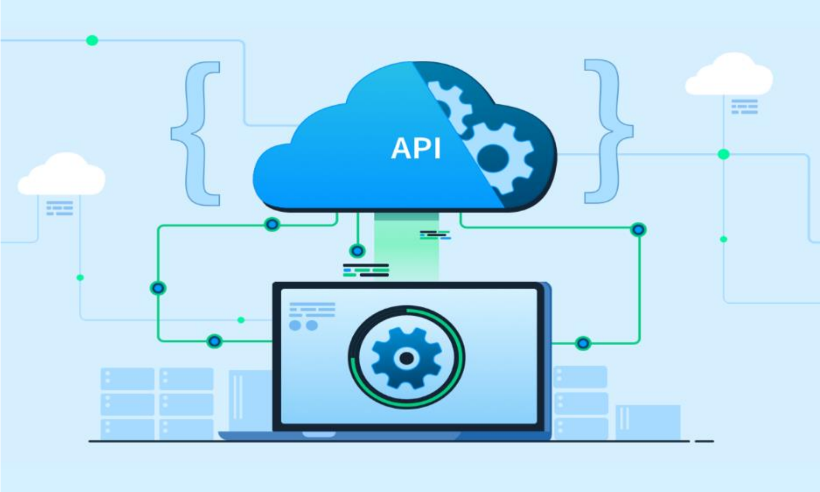


Required Resources, Technology and Platform

Utility	Type of Resources	Tools used	Time Taken	Estimating the Cost
Data Source	DataSet	API (OpenWeather API, MareaAPI) IMD Website	1 month	Free
Local machine	Hardware	64 bit	4 months	500 USD
Data Pre-processing	Software	Pandas, Jupyter Notebook	2 months	Free
Machine Learning Framework	Software	Sklearn	2 months	Free
Visualization tool	SaaS	Tableau	2 months	Student License

Data Collection Process

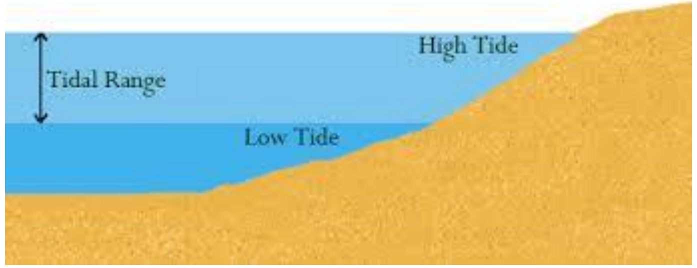




- Data Collected after regular interval of 15 minutes

Data Sources	Data collected	Duration	Data Format	Quantity of data
Open Weather API	Weather data	One month	JSON	1273 X 15
Marea API	Tidal data	One month	JSON	1273 X 15
Satellite images	IMD Website	One month	JPG	1270

Raw Datasets



Tidal data

API Server https://api.marea.ooo/v2
Authentication API Key (x-marea-api-token) in header

Response Status: 200
Took 209 milliseconds

RESPONSE RESPONSE HEADERS CURL


```
      "state": "HIGH TIDE",
      "datetime": "2022-09-16T22:46:52+00:00"
    },
    {
      "timestamp": 1663387356,
      "height": -0.0419750423,
      "state": "LOW TIDE",
      "datetime": "2022-09-17T04:02:36+00:00"
    },
    {
      "timestamp": 1663406455,
      "height": 0.2059674117,
      "state": "HIGH TIDE",
      "datetime": "2022-09-17T09:20:55+00:00"
    },
    {
      "timestamp": 1663431293,
      "height": -0.4890664111,
```


disclaimer	status	latitude	longitude	origin	datums	timestamp	datetime	unit	timezone	datum	extremes	heights	source	copyrigl
NOT SUITABLE FOR NAVIGATIONAL PURPOSES. Marea ...	200	76.26	9.94	{'latitude': 76.25, 'longitude': 9.9375, 'dist...	{'LAT': -0.84, 'HAT': 0.752, 'MLLW': -0.475000...	2022-10-26 08:00:18	2022-10-26 08:00:18+00:00	m	UTC	MSL	[{'timestamp': 1666787226, 'height': 0.5704027...	[{'timestamp': 1666771218, 'height': -0.459390...	FES2014	©202 Marea Generate usir AVISC Products.
NOT SUITABLE FOR NAVIGATIONAL PURPOSES. Marea ...	200	76.26	9.94	{'latitude': 76.25, 'longitude': 9.9375, 'dist...	{'LAT': -0.84, 'HAT': 0.752, 'MLLW': -0.475000...	2022-10-26 08:16:51	2022-10-26 08:16:51+00:00	m	UTC	MSL	[{'timestamp': 1666787226, 'height': 0.5704027...	[{'timestamp': 1666772211, 'height': -0.388316...	FES2014	©202 Marea Generate usir AVISC Products.
NOT SUITABLE FOR NAVIGATIONAL PURPOSES. Marea ...	200	76.26	9.94	{'latitude': 76.25, 'longitude': 9.9375, 'dist...	{'LAT': -0.84, 'HAT': 0.752, 'MLLW': -0.475000...	2022-10-26 08:31:51	2022-10-26 08:31:51+00:00	m	UTC	MSL	[{'timestamp': 1666787226, 'height': 0.5704027...	[{'timestamp': 1666773111, 'height': -0.318326...	FES2014	©202 Marea Generate usir AVISC Products.
NOT SUITABLE FOR NAVIGATIONAL PURPOSES. Marea ...	200	76.26	9.94	{'latitude': 76.25, 'longitude': 9.9375, 'dist...	{'LAT': -0.84, 'HAT': 0.752, 'MLLW': -0.475000...	2022-10-26 08:46:51	2022-10-26 08:46:51+00:00	m	UTC	MSL	[{'timestamp': 1666787226, 'height': 0.5704027...	[{'timestamp': 1666774011, 'height': -0.244201...	FES2014	©202 Marea Generate usir AVISC Products.
NOT SUITABLE FOR NAVIGATIONAL PURPOSES. Marea ...	200	76.26	9.94	{'latitude': 76.25, 'longitude': 9.9375, 'dist...	{'LAT': -0.84, 'HAT': 0.752, 'MLLW': -0.475000...	2022-10-26 09:01:51	2022-10-26 09:01:51+00:00	m	UTC	MSL	[{'timestamp': 1666787226, 'height': 0.5704027...	[{'timestamp': 1666774911, 'height': -0.167119...	FES2014	©202 Marea Generate usir AVISC Products.

1273 rows × 15 columns

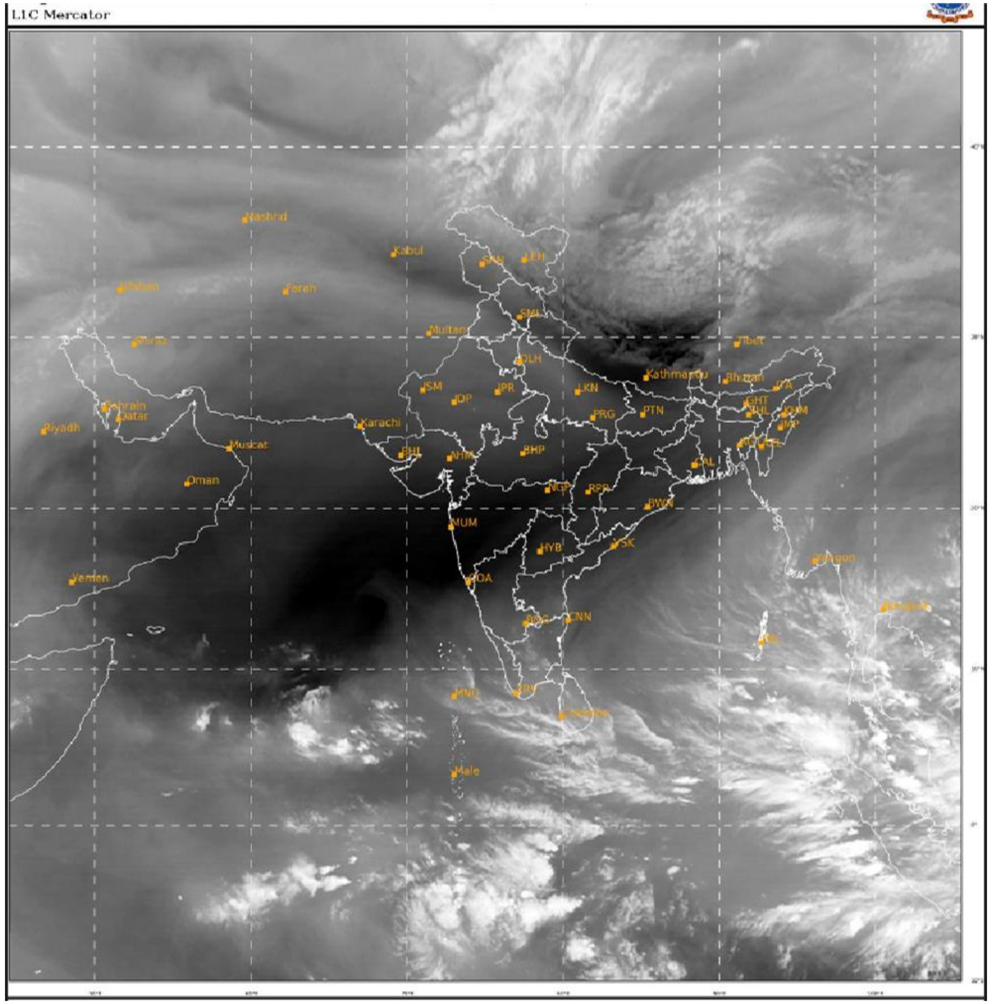
Raw Datasets

Weather Data

	coord	weather	base	main	visibility	wind	clouds	dt	sys	timezone	id	name	cod	snow	rain
0	{'lon': 9.94, 'lat': 76.26}	[{'id': 804, 'main': 'Clouds', 'description': ...	stations	{'temp': 271.89, 'feels_like': 266.75, 'temp_m...	7077	{'speed': 4.85, 'deg': 87, 'gust': 4.71}	{'all': 100}	1666771218	{'sunrise': 1666774297, 'sunset': 1666790195}	3600	0		200	NaN	NaN
1	{'lon': 9.94, 'lat': 76.26}	[{'id': 804, 'main': 'Clouds', 'description': ...	stations	{'temp': 271.89, 'feels_like': 266.75, 'temp_m...	7077	{'speed': 4.85, 'deg': 87, 'gust': 4.71}	{'all': 100}	1666772211	{'sunrise': 1666774297, 'sunset': 1666790195}	3600	0		200	NaN	NaN
2	{'lon': 9.94, 'lat': 76.26}	[{'id': 804, 'main': 'Clouds', 'description': ...	stations	{'temp': 272.08, 'feels_like': 266.87, 'temp_m...	10000	{'speed': 5.04, 'deg': 84, 'gust': 4.55}	{'all': 100}	1666773111	{'sunrise': 1666774297, 'sunset': 1666790195}	3600	0		200	NaN	NaN
3	{'lon': 9.94, 'lat': 76.26}	[{'id': 804, 'main': 'Clouds', 'description': ...	stations	{'temp': 272.08, 'feels_like': 266.87, 'temp_m...	10000	{'speed': 5.04, 'deg': 84, 'gust': 4.55}	{'all': 100}	1666774011	{'sunrise': 1666774297, 'sunset': 1666790195}	3600	0		200	NaN	NaN
4	{'lon': 9.94, 'lat': 76.26}	[{'id': 804, 'main': 'Clouds', 'description': ...	stations	{'temp': 272.08, 'feels_like': 266.87, 'temp_m...	10000	{'speed': 5.04, 'deg': 84, 'gust': 4.55}	{'all': 100}	1666774911	{'sunrise': 1666774297, 'sunset': 1666790195}	3600	0		200	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1268	{'lon': 9.94, 'lat': 76.26}	[{'id': 804, 'main': 'Clouds', 'description': ...	stations	{'temp': 273.05, 'feels_like': 271.53, 'temp_m...	10000	{'speed': 1.35, 'deg': 41, 'gust': 1.05}	{'all': 100}	1667924955	{'sunrise': 0, 'sunset': 0}	3600	0		200	NaN	NaN

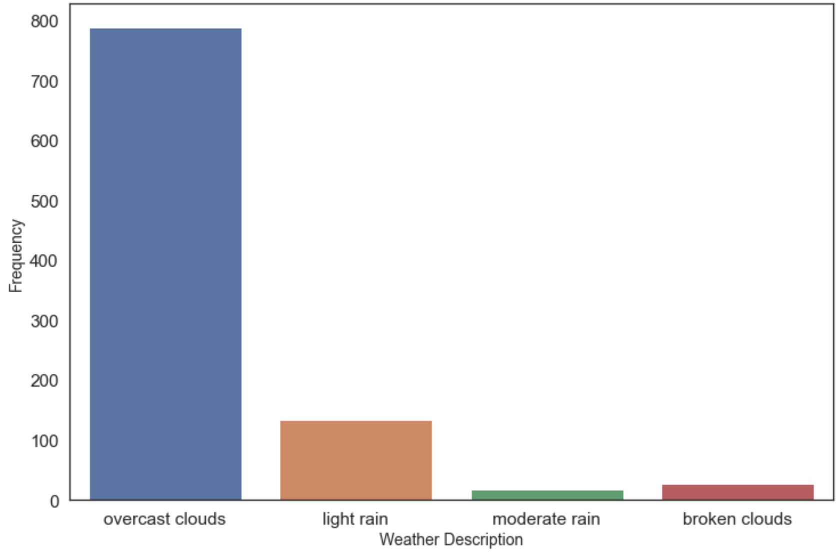
1273 rows × 15 columns

Satellite images

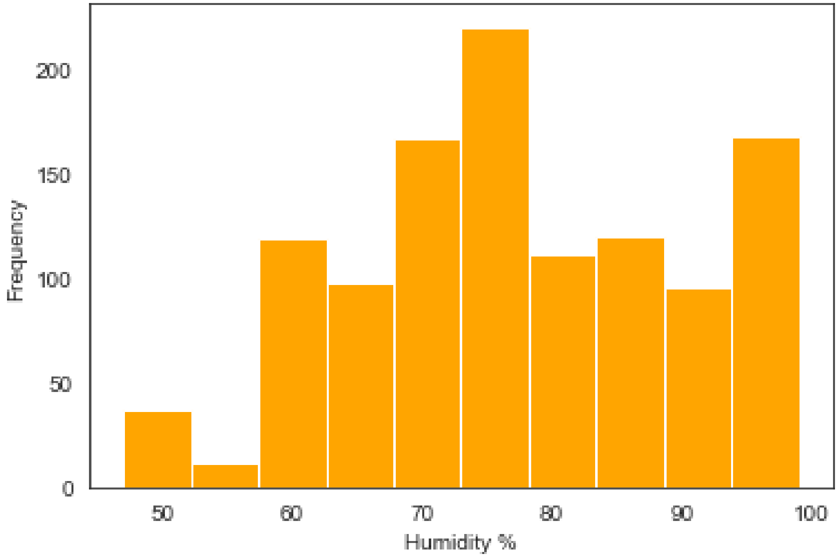


Data Exploration

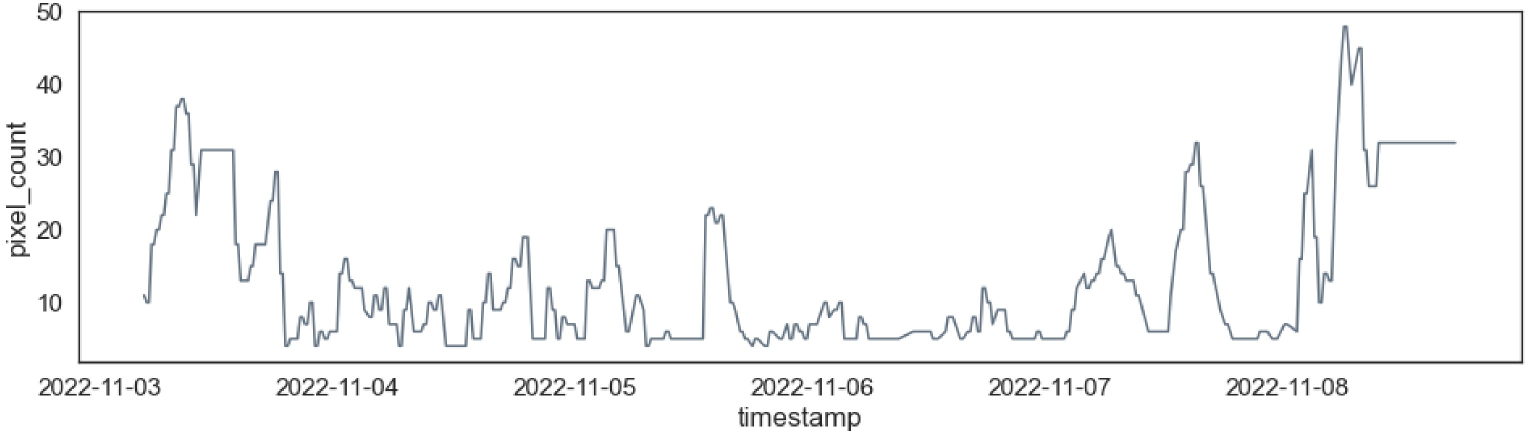
Weather Description



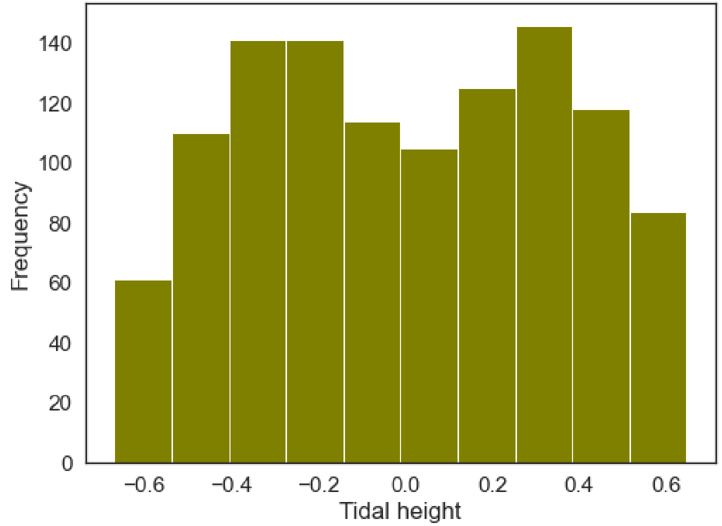
Distribution of humidity %



Trend for pixel count



Distribution of tidal height



Data Pre-Processing

Tidal Data

Dropping noisy data → Extracting relevant information → No missing values

	timestamp	timestamps	tidal_height	tidal_state
0	2022-10-26 08:00:18	1666771218	-0.459391	RISING
1	2022-10-26 08:16:51	1666772211	-0.388316	RISING
2	2022-10-26 08:31:51	1666773111	-0.318326	RISING
3	2022-10-26 08:46:51	1666774011	-0.244202	RISING
4	2022-10-26 09:01:51	1666774911	-0.167120	RISING
...	...	...	...	...
1268	2022-11-08 16:29:16	1667924956	-0.422229	FALLING
1269	2022-11-08 16:44:15	1667925855	-0.464931	FALLING
1270	2022-11-08 16:59:15	1667926755	-0.500146	FALLING
1271	2022-11-08 17:14:25	1667927665	-0.527511	FALLING
1272	2022-11-08 17:29:19	1667928559	-0.545890	FALLING

1273 rows × 4 columns

Weather Data

Dropping noisy data → Extracting relevant information → Handling missing values

	dt	rain	Weather_description	humidity	cloudiness
0	1666771218	0	overcast clouds	56	100
1	1666772211	0	overcast clouds	56	100
2	1666773111	0	overcast clouds	52	100
3	1666774011	0	overcast clouds	52	100
4	1666774911	0	overcast clouds	52	100
...	...	...	...	...	...
1268	1667924955	0	overcast clouds	59	100
1269	1667925856	0	overcast clouds	59	100
1270	1667926755	0	overcast clouds	59	100
1271	1667927670	0	overcast clouds	59	100
1272	1667928559	0	overcast clouds	59	100

1273 rows × 5 columns

Satellite images

Cropping image to eliminate noisy/irrelevant data → Image enhancing → Extracting relevant information

	timestamp	pixel_count
0	2022-10-26 01:00:00	7
1	2022-10-26 01:16:00	5
2	2022-10-26 01:31:00	5
3	2022-10-26 01:46:00	5
4	2022-10-26 02:01:00	5
...	...	...
1265	2022-11-08 08:59:00	26
1266	2022-11-08 09:14:00	32
1267	2022-11-08 09:29:00	32
1268	2022-11-10 10:29:00	24
1269	2022-11-10 10:44:00	24

1270 rows × 2 columns

Data Integration and transformation

Data Integration → Thresholding to create target feature → Hot encoding → Data Normalization

Threshold condition : Probability of floods if Humidity > 90, tidal height > 0.6,
Weather Description = moderate/light rain

	rain	humidity	cloudiness	tidal_height	pixel_count	tidal_state_RISING	broken_clouds	light_rain	moderate_rain	overcast_clouds	target
0	0.0	0.173077	1.0	0.158979	0.071429	1.0	0	0.0	0.0	1.0	0
1	0.0	0.173077	1.0	0.213084	0.035714	1.0	0	0.0	0.0	1.0	0
2	0.0	0.096154	1.0	0.266363	0.035714	1.0	0	0.0	0.0	1.0	0
3	0.0	0.096154	1.0	0.322790	0.035714	1.0	0	0.0	0.0	1.0	0
4	0.0	0.096154	1.0	0.381467	0.017857	1.0	0	0.0	0.0	1.0	0
...	...	...	...	...	...	...	...	...	...	...	...
1140	0.0	0.230769	1.0	0.468670	0.500000	0.0	0	0.0	0.0	1.0	0
1141	0.0	0.230769	1.0	0.313749	0.500000	0.0	0	0.0	0.0	1.0	0
1142	0.0	0.230769	1.0	0.267441	0.500000	0.0	0	0.0	0.0	1.0	0
1143	0.0	0.230769	1.0	0.127954	0.500000	0.0	0	0.0	0.0	1.0	0
1144	0.0	0.230769	1.0	0.093132	0.500000	0.0	0	0.0	0.0	1.0	0

1145 rows × 11 columns

Data Preparation

Training Dataset – 80%

	rain	humidity	cloudiness	tidal_height	pixel_count	tidal_state_RISING	broken_clouds	light_rain	moderate_rain	overcast_clouds
0	0.0	0.173077	1.000	0.158979	0.071429	1.0	0	0.0	0.0	1.0
1	0.0	0.173077	1.000	0.213084	0.035714	1.0	0	0.0	0.0	1.0
2	0.0	0.096154	1.000	0.266363	0.035714	1.0	0	0.0	0.0	1.0
3	0.0	0.096154	1.000	0.322790	0.035714	1.0	0	0.0	0.0	1.0
4	0.0	0.096154	1.000	0.381467	0.017857	1.0	0	0.0	0.0	1.0
...	...	...	...	...	...	...	...	...	...	...
911	0.0	0.538462	0.950	0.427510	0.321429	0.0	0	0.0	0.0	1.0
912	0.0	0.538462	0.950	0.387400	0.321429	0.0	0	0.0	0.0	1.0
913	0.0	0.903846	0.975	0.348877	0.339286	0.0	0	0.0	0.0	0.0
914	0.0	0.903846	0.975	0.312611	0.339286	0.0	0	0.0	0.0	0.0
915	0.0	0.903846	0.975	0.279249	0.303571	0.0	0	0.0	0.0	0.0

916 rows × 10 columns

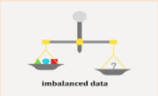
SMOTE

	rain	humidity	cloudiness	tidal_height	pixel_count	tidal_state_RISING	broken_clouds	moderate_rain	light_rain	overcast_clouds
0	0.000000	0.173077	1.0	0.158979	0.071429	1.0	0.0	0.0	0.0	1.0
1	0.000000	0.173077	1.0	0.213084	0.035714	1.0	0.0	0.0	0.0	1.0
2	0.000000	0.096154	1.0	0.266363	0.035714	1.0	0.0	0.0	0.0	1.0
3	0.000000	0.096154	1.0	0.322790	0.035714	1.0	0.0	0.0	0.0	1.0
4	0.000000	0.096154	1.0	0.381467	0.017857	1.0	0.0	0.0	0.0	1.0
...	...	...	...	...	...	...	...	...	...	...
1555	0.151789	0.744737	1.0	0.702416	0.607143	0.0	0.0	1.0	0.0	0.0
1556	0.620000	0.980769	1.0	0.389860	0.079703	0.0	0.0	1.0	0.0	0.0
1557	1.196200	0.927051	1.0	0.747924	0.132381	0.0	0.0	0.0	1.0	0.0
1558	0.110000	0.980769	1.0	0.668999	0.038718	0.0	0.0	1.0	0.0	0.0
1559	0.400000	0.884615	1.0	0.525841	0.125000	1.0	0.0	1.0	0.0	0.0

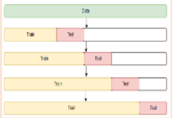
1560 rows × 10 columns

Time Series Splitting

80% - 20% Split



Imbalanced Dataset.
Hence SMOTE performed



3-fold time series Cross Validation

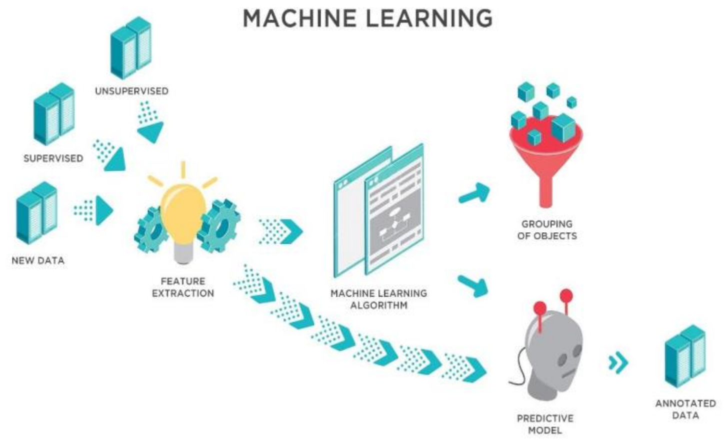
Testing Dataset – 20%

	rain	humidity	cloudiness	tidal_height	pixel_count	tidal_state_RISING	broken_clouds	light_rain	moderate_rain	overcast_clouds
916	0.0	0.903846	0.975	0.249399	0.303571	0.0	0	0.0	0.0	0.0
917	0.0	0.961538	0.975	0.223616	0.321429	0.0	0	0.0	0.0	0.0
918	0.0	0.961538	0.975	0.202388	0.321429	0.0	0	0.0	0.0	0.0
919	0.0	0.961538	0.975	0.175149	0.178571	0.0	0	0.0	0.0	0.0
920	0.0	0.865385	1.000	0.169679	0.107143	0.0	0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...
1140	0.0	0.230769	1.000	0.468670	0.500000	0.0	0	0.0	0.0	1.0
1141	0.0	0.230769	1.000	0.313749	0.500000	0.0	0	0.0	0.0	1.0
1142	0.0	0.230769	1.000	0.267441	0.500000	0.0	0	0.0	0.0	1.0
1143	0.0	0.230769	1.000	0.127954	0.500000	0.0	0	0.0	0.0	1.0
1144	0.0	0.230769	1.000	0.093132	0.500000	0.0	0	0.0	0.0	1.0

229 rows × 10 columns

Machine Learning Models

- Support Vector Classifier
- Random Forest Classifier
- XGBoost Model
- K-Means Clustering



Model Implementation

Split the data into Training and Testing → Hyper-Parameter Tuning of each model using GridSearchCV → Finding the best parameters using time series cross validation → Training the model with best parameters → Performance Evaluation on test data

Support Vector Classifier

Advantages

- SVC works well when the data is linearly separable
- It is effective in high dimensional spaces

Disadvantages

- SVC doesn’t work well when the classes overlap i.e. when there is noise
- It is not efficient for large datasets

Justification

- Output is categorical which consist of two classes
- It is most widely used model in flood prediction as per previous research papers
- It is fast and efficient

Random Forest Classifier

Advantages

- It reduces overfitting in decision trees and helps to improve the accuracy
- It doesn't suffer from curse of dimensionality
- Normalizing of data is not required

Disadvantages

- Its computational power is high
- Due to ensemble of trees, It also suffers from interpretability
- It takes time in training the data

Justification

- As there is noise in the hydrology data, random forest model reduces the model complexity thus prevent overfitting

XGBoost Classifier

Advantages

- It prevents overfitting easily
- XGBoost performs very well on large dataset with not too many features

Disadvantages

- XGBoost does not perform well on unstructured and sparse data
- As gradient Boosting is very sensitive to outliers since each classifier is forced to fix the errors produced by the previous model

Justification

- As target variable in hydrology data is mostly imbalanced, XGBoost model is efficient in such cases

K-Means Clustering

Advantages

- Scales to large data sets
- It is simple to implement and easily adapts to new examples

Disadvantages

- It is sensitive to outliers
- It suffers from curse of dimensionality
- Efficiency of the model depends on initial value of K i.e., number of clusters defined

Justification

As K-means clustering works well with unlabeled data, it can be used in flood prediction when target class is not defined

Model Evaluation and Conclusion

--------Random Forest Classifier-------

	precision	recall	f1-score	support
0	0.91	1.00	0.95	264
1	1.00	0.79	0.88	126
accuracy			0.93	390
macro avg	0.95	0.89	0.92	390
weighted avg	0.94	0.93	0.93	390

---------XGBoost Classifier-------

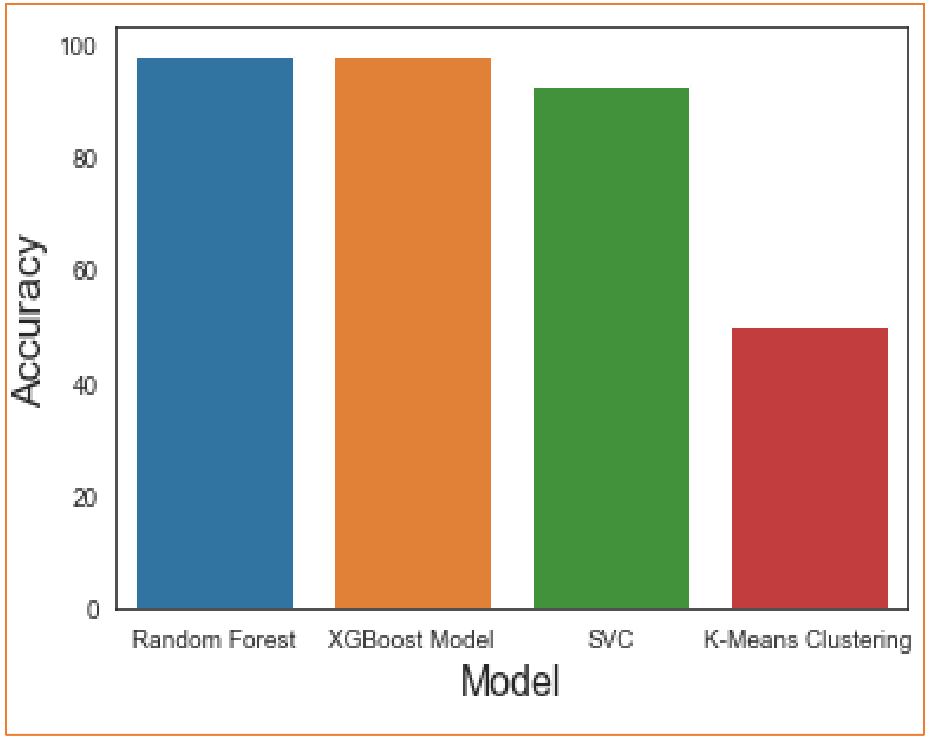
	precision	recall	f1-score
0	0.92	1.00	0.96
1	1.00	0.82	0.90
accuracy			0.94
macro avg	0.96	0.91	0.93
weighted avg	0.95	0.94	0.94

--------SVC-------

	precision	recall	f1-score	support
0	0.76	1.00	0.86	264
1	1.00	0.33	0.50	126
accuracy			0.78	390
macro avg	0.88	0.67	0.68	390
weighted avg	0.84	0.78	0.75	390

-------K-Means Clustering-------

	precision	recall	f1-score	support
0	0.62	0.48	0.54	264
1	0.26	0.39	0.31	126
accuracy			0.45	390
macro avg	0.44	0.43	0.43	390
weighted avg	0.50	0.45	0.47	390



References

- https://arxiv.org/pdf/1406.4756.pdf
- https://doi.org/10.1007/s44196-021-00023-y
- https://doi.org/10.1016/j.ijdrr.2021.102154
- https://doi.org/10.1109/iccike47802.2019.9004419
- https://doi.org/10.1109/icscc51209.2021.9528099
- https://doi.org/10.5194/isprs-annals-v-3-2020-609-2020
- https://doi.org/10.1145/3341105.3374023
- https://doi.org/10.5194/isprs-archives-xliii-b3-2020-641-2020
- https://doi.org/10.1016/j.eswa.2009.09.037
- https://doi.org/10.5281/zenodo.3702303
- https://doi.org/10.1016/j.autcon.2021.103916
- https://doi.org/10.1109/acit53391.2021.9677084
- https://doi.org/10.1007/s12665-011-1504-z
- https://doi.org/10.1109/smartgencon51891.2021.9645804
- https://doi.org/10.11591/ijai.v9.i1.pp73-80
- https://doi.org/10.22115/scce.2022.317761.1383
- https://doi.org/10.1145/3486626.3493432
- https://doi.org/10.1145/3531028.3531037
- https://doi.org/10.1109/CEC.2016.7743944
- https://doi.org/10.3390/w14071140
- https://doi.org/10.1109/JIOT.2020.3044966
- https://doi.org/10.1016/j.jhydrol.2018.06.06

Image Reference

- https://www.pgi.com/blog/2017/08/neuroscience-motivation-increase-motivation/
- https://daubejb.medium.com/are-you-anxious-at-the-start-of-the-project-discover-the-secret-simplicity-of-a-problem-statement-fee9cc268863
- https://www.istockphoto.com/photo/project-deliverables-are-shown-on-the-business-photo-using-the-text-gm1343888857-422433567?phrase=deliverables
- https://thenounproject.com/

Thank you