Data Collection and Preprocessing Phase

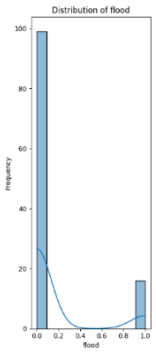
Date	29 September 2024
Team ID	LTVIP2024TMID24876
Project Title	Rising Waters: A Machine Learning Approach to Flood Prediction
Maximum Marks	6 Marks

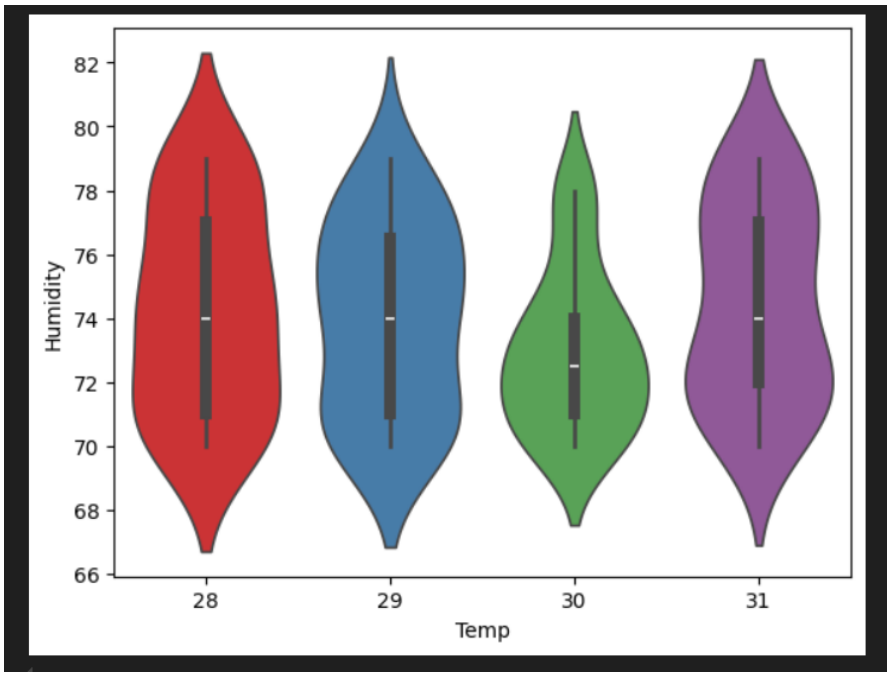
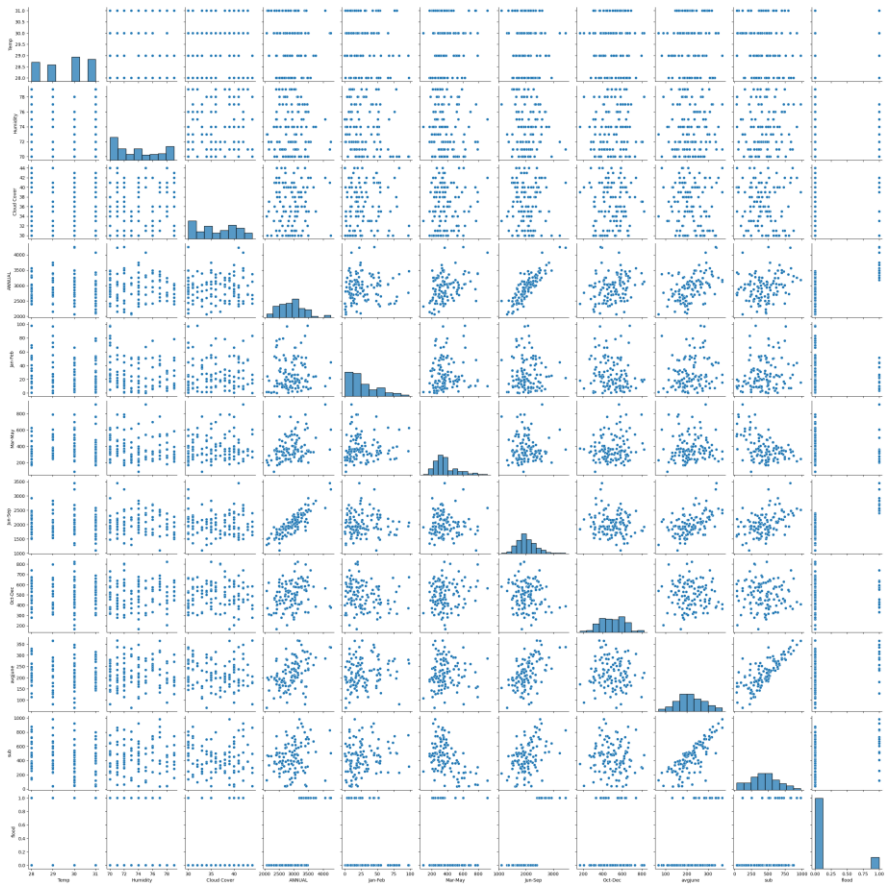
Data Exploration and Preprocessing :

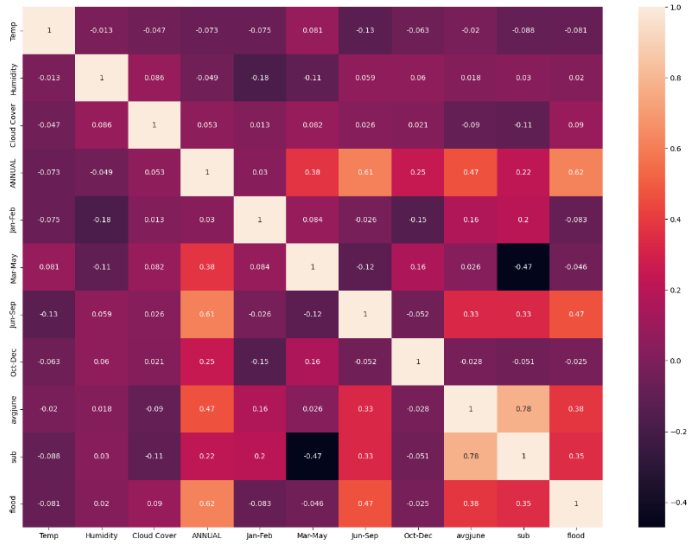
Dataset variables will be statistically analyzed to identify patterns and outliers, with Python employed for preprocessing tasks like normalization and feature engineering. Data cleaning will address missing values and outliers, ensuring quality for subsequent analysis and modeling, and forming a strong foundation for insights and predictions.

Section	Description
Data Overview	Dimension: 115rows X 11columns Descriptive statistics:

	Temp	Humidity	Cloud Cover	ANNUAL	Jan-Feb	Mar-May	Jun-Sep	Oct-Dec	avgjune	sub	flood
count	115.000000	115.000000	115.000000	115.000000	115.000000	115.000000	115.000000	115.000000	115.000000	115.000000	115.000000
mean	29.600000	73.852174	36.286957	2925.487826	27.739130	377.253913	2022.840870	497.636522	218.100870	439.801739	0.139130
std	1.122341	2.947623	4.330158	422.112193	22.361032	151.091850	386.254397	129.860643	62.547597	210.438813	0.347597
min	28.000000	70.000000	30.000000	2068.800000	0.300000	89.900000	1104.300000	166.600000	65.600000	34.200000	0.000000
25%	29.000000	71.000000	32.500000	2627.900000	10.250000	276.750000	1768.850000	407.450000	179.666667	295.000000	0.000000
50%	30.000000	74.000000	36.000000	2937.500000	20.500000	342.000000	1948.700000	501.500000	211.033333	430.600000	0.000000
75%	31.000000	76.000000	40.000000	3164.100000	41.600000	442.300000	2242.900000	584.550000	263.833333	577.650000	0.000000
max	31.000000	79.000000	44.000000	4257.800000	98.100000	915.200000	3451.300000	823.300000	366.066667	982.700000	1.000000

Univariate Analysis	

Bivariate Analysis	
Multivariate Analysis	



Outliers and Anomalies

```
Q3 = df.ANNUAL.quantile(0.75)
Q1 = df.ANNUAL.quantile(0.25)
IQR = Q3-Q1
Upper_limit = Q3+1.5*IQR
Upper_limit
df['ANNUAL']=np.where(df['ANNUAL']>Upper_limit,3968.4000000000005,df['ANNUAL'])
```

Data Preprocessing Code Screenshots

Loading Data

```
df=pd.read_excel('/content/flood dataset.xlsx')
```

```
df.head()
```

	Temp	Humidity	Cloud Cover	ANNUAL	Jan-Feb	Mar-May	Jun-Sep	Oct-Dec	avgjune	sub	flood
0	29	70	30	3248.6	73.4	386.2	2122.8	666.1	274.866667	649.9	0
1	28	75	40	3326.6	9.3	275.7	2403.4	638.2	130.300000	256.4	1
2	28	75	42	3271.2	21.7	336.3	2343.0	570.1	186.200000	308.9	0
3	29	71	44	3129.7	26.7	339.4	2398.2	365.3	366.066667	862.5	0
4	31	74	40	2741.6	23.4	378.5	1881.5	458.1	283.400000	586.9	0

Handling Missing Data

```
df.isnull().sum()
```

	0
Temp	0
Humidity	0
Cloud Cover	0
ANNUAL	0
Jan-Feb	0
Mar-May	0
Jun-Sep	0
Oct-Dec	0
avgjune	0
sub	0
flood	0

dtype: int64

Data Transformation	There is no need of Standardization and Normalization of our dataset, as we using Ensemble Technique.
Feature Engineering