

**Title:Fire Weather Index Predictor**  
(A Machine Learning model to predict Fire Weather)



**Infosys Springboard Virtual Internship Program**

Submitted By  
**T.Srikavya**

Under the guidance of Mentor **Praveen**

## **Project Statement**

This project focuses on building a Fire Weather Index (FWI) Predictor that estimates wildfire risk using essential environmental and fire-danger features such as temperature, relative humidity, wind speed, rainfall, FFMC, DMC, DC, ISI, and BUI. The dataset is thoroughly cleaned, pre-processed, and analysed to ensure reliable inputs, followed by visual exploration to understand feature distributions and relationships. Machine learning techniques are then applied to learn how weather conditions influence fire danger levels, enabling accurate prediction of the FWI value. The system also incorporates regional information to study area-wise fire behaviour and improve prediction relevance. Overall, the model aims to support early detection, risk assessment, and informed decision-making for effective wildfire management.

## **Expected Outcome**

- A predictive ML model trained using Ridge Regression to forecast FWI.
- A pre-processing pipeline using StandardScaler for normalisation.
- A Flask-based web app where users can input environmental values and get FWI predictions.
- A system that can help forest departments, emergency planners, and climate researchers make data-driven decisions.

## **Modules to be Implemented**

- Data Collection
- Data Exploration (EDA) and Data Preprocessing
- Feature Engineering and Scaling
- Model Training using Ridge Regression
- Evaluation and Optimisation
- Deployment via Flask App
- Presentation and Documentation

## **System Requirements**

### **Software:**

- Python
- Python libraries (pandas, numpy, matplotlib, seaborn etc)
- Flask

# Milestone 1

## 1. Module 1 (Data Collection)

The dataset was collected by exploring multiple online sources and selecting one that contained the essential environmental features required for FWI prediction, including Temperature, Relative Humidity, Wind Speed, Rain, FFMC, DMC, ISI, and Region. After loading the chosen dataset into a Pandas DataFrame, an initial inspection was carried out to understand its structure and quality. This included checking datatypes, identifying null values, reviewing memory usage, and generating statistical summaries to examine the distribution and range of numerical features. The dataset's shape and duplicate entries were also analysed to ensure completeness and reliability, providing a solid foundation for further preprocessing and modelling.

**Conducted initial inspection to understand feature distributions and data quality.**

1. Load the dataset

```
import pandas as pd  
df = pd.read_excel("/kaggle/input/fwidataset/FWI Dataset.xlsx")  
print("Dataset loaded successfully")
```

Dataset loaded successfully

## 2. Verify Datatypes

```
Basic Information About Dataset:  
<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 244 entries, 0 to 243  
Data columns (total 15 columns):  
 #   Column      Non-Null Count  Dtype     
 ---  --          -----          ----  
 0   day         244 non-null    int64    
 1   month        244 non-null    int64    
 2   year         244 non-null    int64    
 3   Temperature  244 non-null    int64    
 4   RH           244 non-null    int64    
 5   Ws           244 non-null    int64    
 6   Rain          244 non-null    float64  
 7   FFMC          244 non-null    float64  
 8   DMC           244 non-null    float64  
 9   DC            244 non-null    object    
 10  ISI           244 non-null    float64  
 11  BUI           244 non-null    float64  
 12  FWI           244 non-null    object    
 13  Classes        243 non-null    object    
 14  Region         244 non-null    object    
 dtypes: float64(5), int64(6), object(4)  
memory usage: 28.7+ KB  
None
```

### 3. Basic Information of the Dataset

```
print("\nFirst 4 Rows of Dataset")
display(df.head())
print("\nLast 4 Rows of Dataset")
display(df.tail())
```

| First 4 Rows of Dataset |     |       |      |             |    |    |      |      |     |      |     |     |     |          |        |
|-------------------------|-----|-------|------|-------------|----|----|------|------|-----|------|-----|-----|-----|----------|--------|
|                         | day | month | year | Temperature | RH | Ws | Rain | FFMC | DMC | DC   | ISI | BUI | FWI | Classes  | Region |
| 0                       | 1   | 6     | 2012 | 29          | 57 | 18 | 0.0  | 65.7 | 3.4 | 7.6  | 1.3 | 3.4 | 0.5 | not fire | Bejaia |
| 1                       | 2   | 6     | 2012 | 29          | 61 | 13 | 1.3  | 64.4 | 4.1 | 7.6  | 1.0 | 3.9 | 0.4 | not fire | Bejaia |
| 2                       | 3   | 6     | 2012 | 26          | 82 | 22 | 13.1 | 47.1 | 2.5 | 7.1  | 0.3 | 2.7 | 0.1 | not fire | Bejaia |
| 3                       | 4   | 6     | 2012 | 25          | 89 | 13 | 2.5  | 28.6 | 1.3 | 6.9  | 0.0 | 1.7 | 0   | not fire | Bejaia |
| 4                       | 5   | 6     | 2012 | 27          | 77 | 16 | 0.0  | 64.8 | 3.0 | 14.2 | 1.2 | 3.9 | 0.5 | not fire | Bejaia |

| Last 4 Rows of Dataset |     |       |      |             |    |    |      |      |      |      |     |      |     |          |                |
|------------------------|-----|-------|------|-------------|----|----|------|------|------|------|-----|------|-----|----------|----------------|
|                        | day | month | year | Temperature | RH | Ws | Rain | FFMC | DMC  | DC   | ISI | BUI  | FWI | Classes  | Region         |
| 239                    | 26  | 9     | 2012 | 30          | 65 | 14 | 0.0  | 85.4 | 16.0 | 44.5 | 4.5 | 16.9 | 6.5 | fire     | Sidi-Bel Abbes |
| 240                    | 27  | 9     | 2012 | 28          | 87 | 15 | 4.4  | 41.1 | 6.5  | 8    | 0.1 | 6.2  | 0   | not fire | Sidi-Bel Abbes |
| 241                    | 28  | 9     | 2012 | 27          | 87 | 29 | 0.5  | 45.9 | 3.5  | 7.9  | 0.4 | 3.4  | 0.2 | not fire | Sidi-Bel Abbes |
| 242                    | 29  | 9     | 2012 | 24          | 54 | 18 | 0.1  | 79.7 | 4.3  | 15.2 | 1.7 | 5.1  | 0.7 | not fire | Sidi-Bel Abbes |
| 243                    | 30  | 9     | 2012 | 24          | 64 | 15 | 0.2  | 67.3 | 3.8  | 16.5 | 1.2 | 4.8  | 0.5 | not fire | Sidi-Bel Abbes |

**FIG 1: Information of Dataset**

The first and last four rows of the dataset were displayed to gain an initial understanding of the data structure, feature values, and overall formatting. This helped verify that all columns loaded correctly, values are aligned, and the dataset is ready for further inspection and preprocessing.

### 4. Statistical Summary

```
display(df.describe().T)
```

|             | count | mean        | std       | min    | 25%      | 50%     | 75%      | max    |
|-------------|-------|-------------|-----------|--------|----------|---------|----------|--------|
| day         | 244.0 | 15.754098   | 8.825059  | 1.0    | 8.000    | 16.00   | 23.000   | 31.0   |
| month       | 244.0 | 7.500000    | 1.112961  | 6.0    | 7.000    | 7.50    | 8.000    | 9.0    |
| year        | 244.0 | 2012.000000 | 0.000000  | 2012.0 | 2012.000 | 2012.00 | 2012.000 | 2012.0 |
| Temperature | 244.0 | 32.172131   | 3.633843  | 22.0   | 30.000   | 32.00   | 35.000   | 42.0   |
| RH          | 244.0 | 61.938525   | 14.884200 | 21.0   | 52.000   | 63.00   | 73.250   | 90.0   |
| Ws          | 244.0 | 15.504098   | 2.810178  | 6.0    | 14.000   | 15.00   | 17.000   | 29.0   |
| Rain        | 244.0 | 0.760656    | 1.999406  | 0.0    | 0.000    | 0.00    | 0.500    | 16.8   |
| FFMC        | 244.0 | 77.887705   | 14.337571 | 28.6   | 72.075   | 83.50   | 88.300   | 96.0   |
| DMC         | 244.0 | 14.673361   | 12.368039 | 0.7    | 5.800    | 11.30   | 20.750   | 65.9   |
| ISI         | 244.0 | 4.774180    | 4.175318  | 0.0    | 1.400    | 3.50    | 7.300    | 19.0   |
| BUI         | 244.0 | 16.664754   | 14.204824 | 1.1    | 6.000    | 12.25   | 22.525   | 68.0   |

**FIG 2: Statistical summary of the Dataset**

A statistical summary was generated to examine key numerical characteristics such as mean, median, minimum, maximum, and standard deviation for each feature. This helped assess data variability, detect potential outliers, and understand the overall distribution of numerical values in the dataset.

## 5. Duplicate Values

```
print("\nDuplicate rows:", df.duplicated().sum())
Duplicate rows: 0
```

## 2. Module 2 (Data Exploration (EDA) and Data Preprocessing)

During the preprocessing stage, the dataset was first examined for missing or null values, and appropriate handling techniques were applied to ensure completeness. Outlier detection was then performed using boxplots and statistical thresholds to identify abnormal values that could affect model performance. To better understand feature behaviour, data distributions were visualised through histograms and density plots, while correlation matrices and scatterplots were used to explore relationships between variables. Categorical features, such as *Region*, were encoded using label encoding to make them suitable for machine learning algorithms. Finally, the cleaned and processed dataset was saved for use in building and evaluating predictive models.

### 1. Handle missing values

```
print("Before:\n", df.isnull().sum())
num_cols = df.select_dtypes(include=['number']).columns
cat_cols = df.select_dtypes(include=['object','category']).columns
for c in num_cols:
    df[c] = df[c].fillna(df[c].median())
for c in cat_cols:
    df[c] = df[c].fillna(df[c].mode()[0])

print("\nAfter:\n", df.isnull().sum())
```

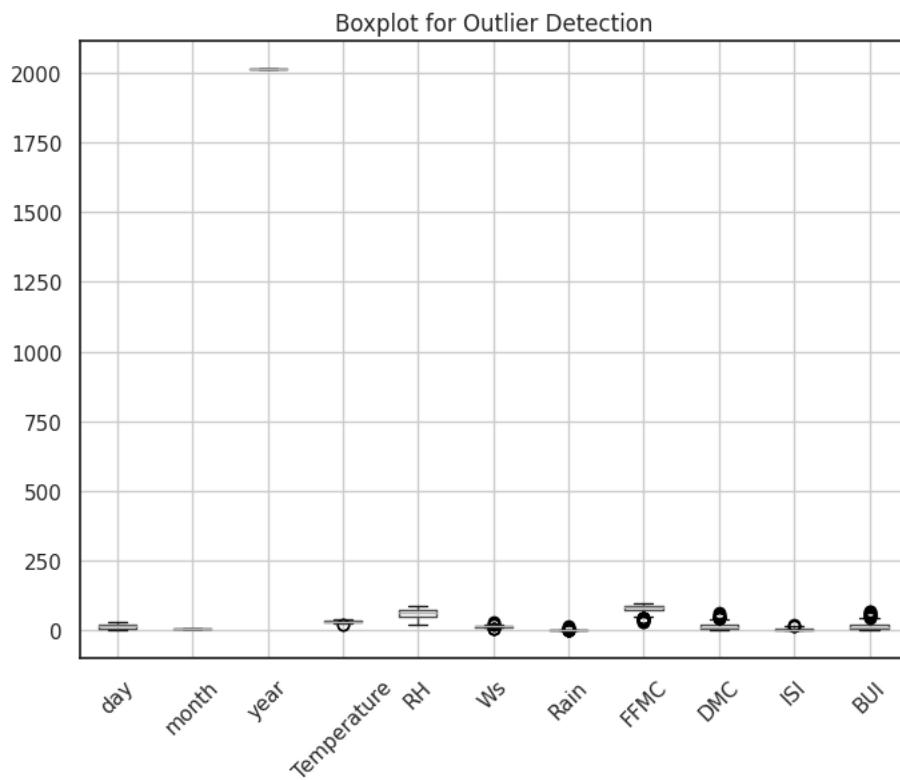
```
Before:
day      0
month     0
year      0
Temperature 0
RH        0
WS        0
Rain      0
FFMC      0
DMC       0
DC        0
ISI       0
BUI       0
FWI       0
Classes    1
Region     0
dtype: int64

After:
day      0
month     0
year      0
Temperature 0
RH        0
WS        0
Rain      0
FFMC      0
DMC       0
DC        0
ISI       0
BUI       0
FWI       0
Classes    0
Region     0
dtype: int64
```

**FIG 3: Before and After Missing Values**

A missing-value check was performed to identify incomplete records, revealing that only one row contained a null value in the *Classes* column. This validation helped ensure data completeness and supported appropriate handling before further preprocessing.

## 2. Boxplot and statistical threshold using the IQR method



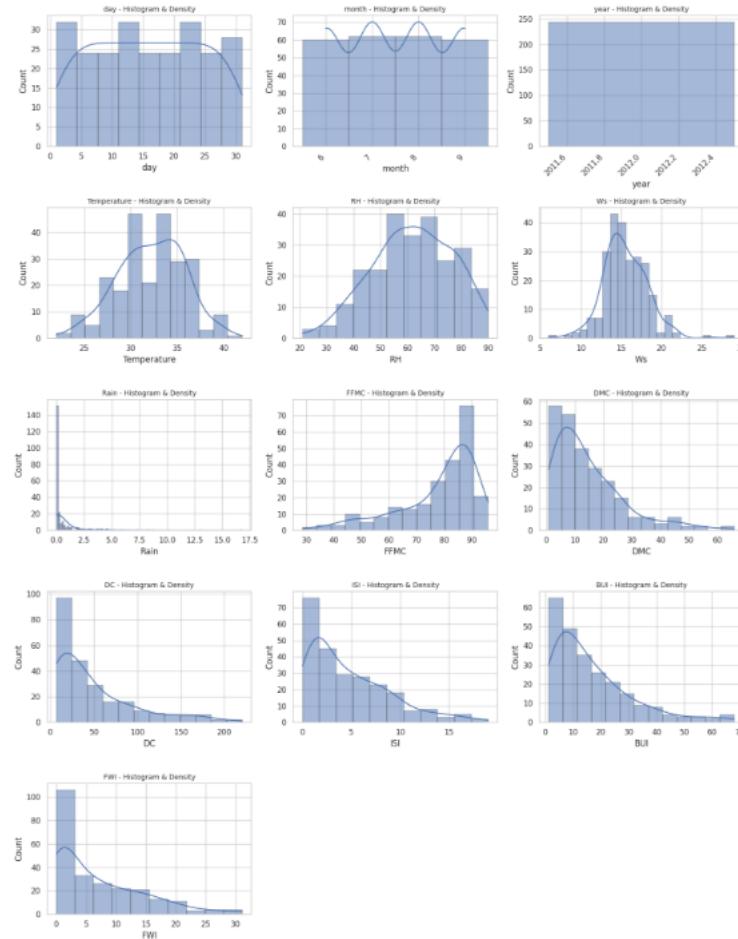
**FIG 4: Outlier Detection**

Outlier detection was performed using both boxplots and the Interquartile Range (IQR) statistical method to identify unusually high or low values across numerical features. The IQR approach calculated lower and upper bounds for each feature, flagging values that fell outside the  $1.5 \times \text{IQR}$  range as potential outliers. While some features like day, month, year, RH, and DC showed no outliers, others such as Temperature, Wind Speed, Rain, FFMC, DMC, ISI, and BUI contained multiple extreme values. These findings helped reveal variability patterns within the dataset and informed decisions for further cleaning and preprocessing.

## 3. Density Plots and histogram for each feature

Histograms with KDE density curves were generated for all numerical features to visualize their distributions and identify underlying patterns. These plots helped reveal whether features were normally distributed, skewed, or contained extreme values. Variables like Temperature, RH, and Wind Speed showed smoother, more symmetric distributions, while features such as Rain, DMC, ISI, and BUI

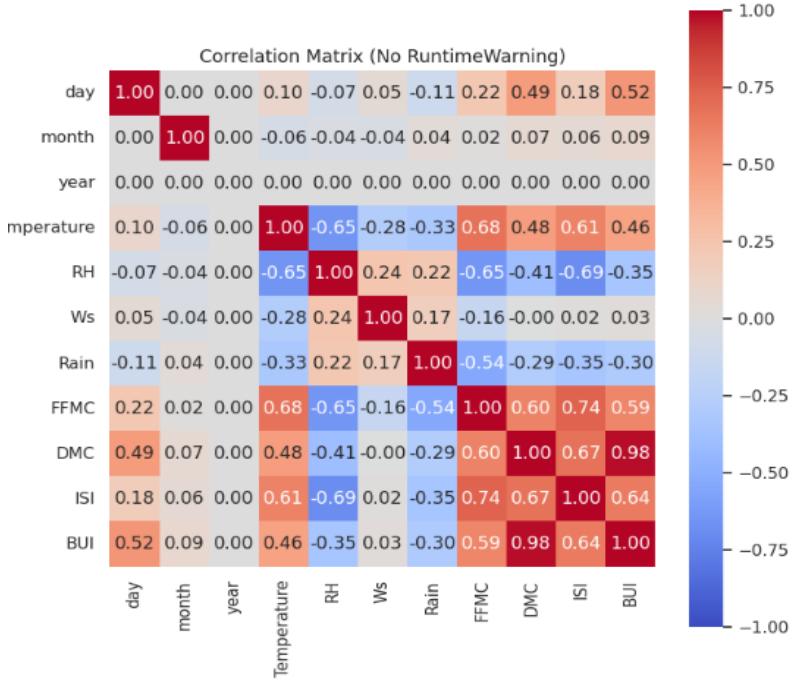
displayed strong right-skewness. This visual exploration provided valuable insights for understanding variability and preparing the data for modeling.



**FIG 4: Histogram with KDE curves for features**

#### 4. Compute the correlation matrix for numerical features

A correlation matrix was generated to explore relationships among numerical features and identify which variables influence each other most strongly. The heatmap revealed that fire danger indices such as BUI, DMC, FFMC, and ISI have high positive correlations with each other and with Temperature, indicating shared patterns in fire behavior. Relative Humidity showed strong negative correlations with several indices, reflecting its inverse effect on fire risk. These insights helped highlight key predictive features for the Fire Weather Index and guided feature selection for modeling.



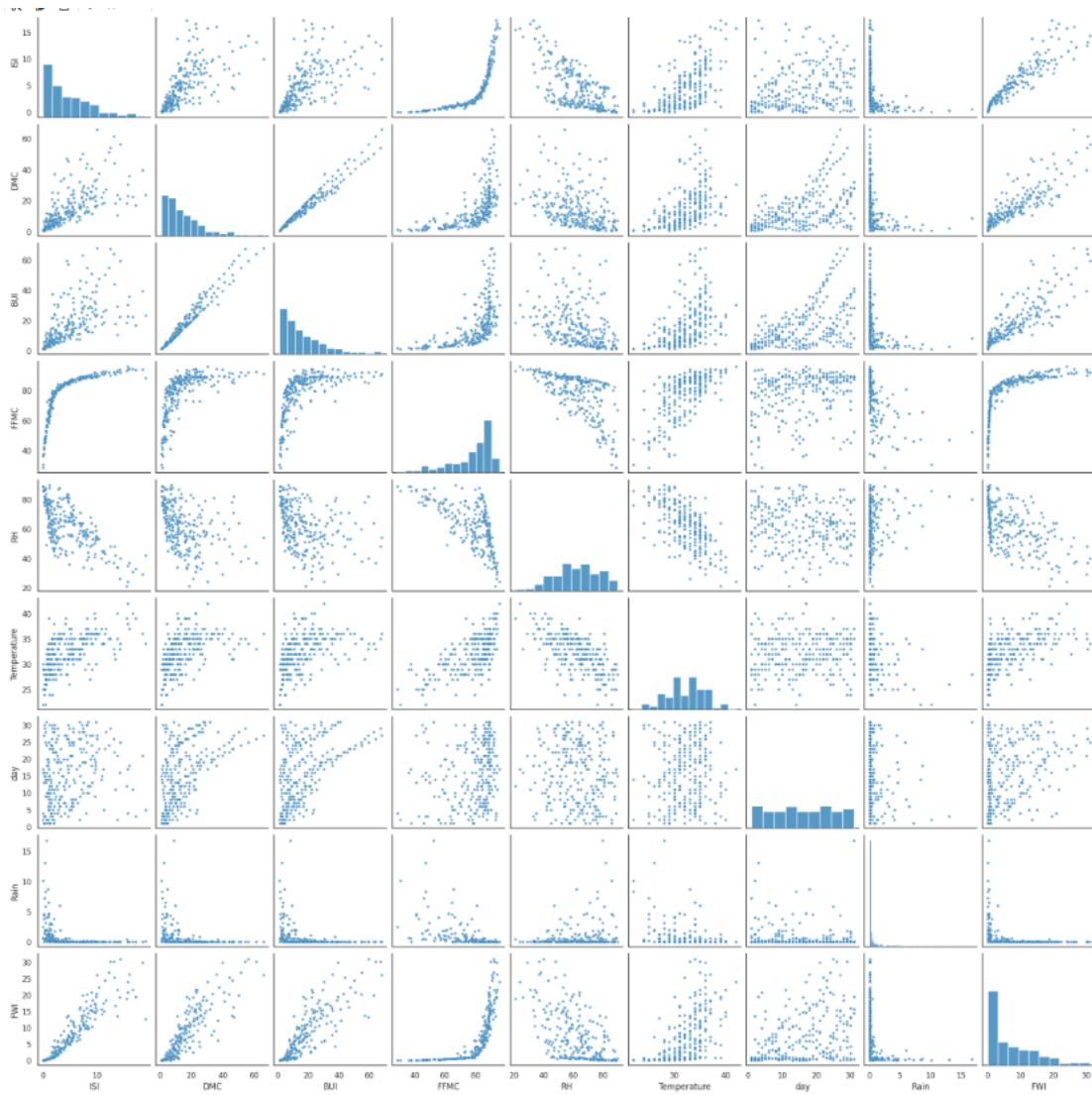
**FIG 5: Correlation matrix**

## 5. Scatter plots for Features vs FWI

Scatter plots were used to visualize the relationship between each numerical feature and the Fire Weather Index (FWI). These plots helped identify how strongly each variable contributes to fire risk.

### Key observations:

- **FFMC, DMC, and ISI** show clear **positive upward trends**, meaning higher values of these indices are strongly associated with higher FWI. These features are likely strong predictors in the model.
- **Temperature** also shows a moderate increasing trend with FWI.
- **Rain and RH** show **negative or weak relationships**, as higher rainfall and humidity typically reduce fire danger.
- **Wind Speed (Ws)** shows a more scattered pattern, indicating a weaker or less direct relationship with FWI.



**FIG 6: Scatterplot for features and FWI**

## 6. Encoding region

Encoded Values:

Region Region\_encoded

|   |        |   |
|---|--------|---|
| 0 | Bejaia | 0 |
| 1 | Bejaia | 0 |
| 2 | Bejaia | 0 |
| 3 | Bejaia | 0 |
| 4 | Bejaia | 0 |

Region Mapping:

{'Bejaia': 0, 'Sidi-Bel Abbes': 1}

Region object

Region\_encoded category

dtype: object

The Region column was label-encoded to convert the categorical values into numerical form. 'Bejaia' was encoded as 0 and 'Sidi-Bel Abbes' as 1. The new column Region\_encoded was added and stored as a categorical type, preparing the feature for machine learning algorithms.

## 7. Saved the cleaned dataset for use in modeling

```

Cleaned dataset saved successfully!
File location: /kaggle/working/cleaned_fwi.csv

Preview of saved file:
   day month year Temperature RH Ws Rain FFMC DMC DC ISI BUI FWI Classes Region Region_encoded
0    1     6 2012        29  57  18  0.0  65.7  3.4  7.6  1.3  3.4  0.5  not fire Bejaia      0
1    2     6 2012        29  61  13  1.3  64.4  4.1  7.6  1.0  3.9  0.4  not fire Bejaia      0
2    3     6 2012        26  82  22 13.1  47.1  2.5  7.1  0.3  2.7  0.1  not fire Bejaia      0
3    4     6 2012        25  89  13  2.5  28.6  1.3  6.9  0.0  1.7  0.0  not fire Bejaia      0
4    5     6 2012        27  77  16  0.0  64.8  3.0  14.2  1.2  3.9  0.5  not fire Bejaia      0

```

## Milestone 2

### 3. Module 3 (Feature Engineering and Scaling)

The primary goal was to prepare the dataset for effective model learning by selecting relevant features and normalising their scale. Based on both correlation analysis and domain knowledge from the standard Fire Weather Index system, nine meteorological and fire-behaviour variables (Temperature, RH, Wind Speed, Rain, FFMC, DMC, DC, ISI, and BUI) were finalised as input features, while FWI was retained as the target variable. Non-domain temporal features such as day, month, and year were removed since they do not directly influence fire behaviour. The dataset was then divided into a feature matrix (X) and a target vector (y). To ensure all features contribute equally during regression and prevent bias due to varying units or ranges, StandardScaler was applied to normalise the numerical attributes. The fitted scaler was saved as a scaler.pkl file for reuse during deployment in the Flask application, ensuring consistent preprocessing of real-time user inputs.

#### 1. Identified and selected the most influential features correlated with the FWI target.

```

final_features = [
    'Temperature',
    'RH',
    'Ws',
    'Rain',
    'FFMC',
    'DMC',
    'DC',
    'ISI',
    'BUI'
]

X = df[final_features]
y = df['FWI']

print("Shape of X - input features:", X.shape)
print("Shape of y - target variable:", y.shape)

```

```
Shape of X - input features: (244, 9)
Shape of y - target variable: (244,)
```

## 2. Split the dataset into X (input features) and y (target variable), and conducted training.

```
X_train, X_test, y_train, y_test = train_test_split(
    X,
    y,
    test_size=0.2,
    random_state=42
)
print("X_train shape:", X_train.shape)
print("X_test shape:", X_test.shape)
print("y_train shape:", y_train.shape)
print("y_test shape:", y_test.shape)
```

```
x_train shape: (195, 9)
x_test shape: (49, 9)
y_train shape: (195,)
y_test shape: (49,)
```

## 3. Applied StandardScaler for uniform scaling across all features.

StandardScaler standardises numerical features by subtracting the mean ( $\mu$ ) and dividing by the standard deviation ( $\sigma$ ), transforming each feature to have zero mean and unit variance, ensuring all variables contribute equally to model training.

```
Scaled training data shape: (194, 9)
Scaled testing data shape: (49, 9)
   Temperature      RH      Ws      Rain      FFMC \
count    1.940000e+02  1.940000e+02  1.940000e+02  1.940000e+02
mean    -4.578239e-18  9.156479e-17  2.014425e-16  -6.867359e-17  -3.204767e-16
std     1.002587e+00  1.002587e+00  1.002587e+00  1.002587e+00
min    -2.693520e+00  -2.784997e+00  -3.699471e+00  -3.737724e-01  -3.458109e+00
25%    -7.407179e-01  -6.286910e-01  -6.307490e-01  -3.737724e-01  -4.783202e-01
50%     0.000000e+00  7.884688e-02  -2.471587e-01  -3.737724e-01  3.918769e-01
75%     8.080559e-01  7.695386e-01  5.200219e-01  -1.824483e-01  7.539099e-01
max     2.693520e+00  1.864538e+00  3.972334e+00  7.661842e+00  1.267485e+00

      DMC      DC      ISI      BUI
count    1.940000e+02  1.940000e+02  1.940000e+02  1.940000e+02
mean    -6.867359e-17  1.510819e-16  -1.419254e-16  -1.007213e-16
std     1.002587e+00  1.002587e+00  1.002587e+00  1.002587e+00
min    -1.112540e+00  -8.942173e-01  -1.146894e+00  -1.085026e+00
25%    -7.237755e-01  -8.033748e-01  -8.094535e-01  -7.568463e-01
50%    -3.007087e-01  -3.551847e-01  -2.825733e-01  -3.287126e-01
75%     4.863480e-01  4.945194e-01  6.231871e-01  4.325989e+00
max     3.857543e+00  3.391945e+00  3.352308e+00  3.372894e+00
```

**FIG 6 Standard scalar**

## 4. Module 4 ( Model Training using Ridge Regression)

To build a reliable FWI prediction model, five supervised regression algorithms were trained and compared. Each model was trained on the scaled training dataset, and hyperparameters were tuned where applicable using techniques such as GridSearchCV to improve prediction performance. The goal of using multiple models was to identify which one best handles the relationships between fire weather variables and the FWI target, ensuring strong generalisation and minimal prediction error on unseen test data

### 1. Linear Regression

- Base regression model to establish a baseline performance
- Assumes a linear relationship between input features and FWI

- No regularisation, useful for comparison

```
lin_reg = LinearRegression()
lin_reg.fit(X_train_scaled, y_train)
y_pred_lr = lin_reg.predict(X_test_scaled)
lr_r2 = r2_score(y_test, y_pred_lr)
lr_rmse = mean_squared_error(y_test, y_pred_lr, squared=False)
print("Linear Regression → R2:", lr_r2, "RMSE:", lr_rmse)
```

Linear Regression → R2: 0.99010227509981 RMSE: 0.6211065367275612

## 2. Ridge Regression

- Adds L2 regularisation to reduce multicollinearity
- Prevents the model from giving large weights to correlated features
- Hyperparameter alpha tuned to balance bias-variance trade-off using GridSearch Cv
- Best model saved as ridge.pkl for deployment

```
ridge = Ridge()
ridge_params = {
    'alpha': [0.01, 0.1, 1, 10, 100]
}

ridge_gs = GridSearchCV(
    ridge,
    ridge_params,
    cv=5,
    scoring='neg_mean_squared_error'
)

ridge_gs.fit(X_train_scaled, y_train)
ridge_model = ridge_gs.best_estimator_
y_pred_ridge = ridge_model.predict(X_test_scaled)
ridge_r2 = r2_score(y_test, y_pred_ridge)
ridge_rmse = mean_squared_error(y_test, y_pred_ridge, squared=False)
print("Best Ridge alpha:", ridge_gs.best_params_)

print("Ridge → R2:", ridge_r2, "RMSE:", ridge_rmse)

with open("ridge.pkl", "wb") as f:
    pickle.dump(ridge_model, f)
```

### 3. ElasticNet Regression

- Combines both L1 and L2 regularization
- Handles multicollinearity and performs feature selection simultaneously
- Balanced penalty using tuned alpha and l1\_ratio

```
elastic = ElasticNet(max_iter=10000)

elastic_params = {

'alpha': [0.01, 0.1, 1, 10], 

'l1_ratio': [0.2, 0.5, 0.8]

}

elastic_gs = GridSearchCV(


elastic, 


elastic_params, 


cv=5, 


scoring='neg_mean_squared_error'

)

elastic_gs.fit(X_train_scaled, y_train)

elastic_model = elastic_gs.best_estimator_


y_pred_elastic = elastic_model.predict(X_test_scaled)

elastic_r2 = r2_score(y_test, y_pred_elastic)

elastic_rmse = mean_squared_error(y_test, y_pred_elastic, squared=False)

print("ElasticNet → R2:", elastic_r2, "RMSE:", elastic_rmse)
```

### 4. Lasso Regression

Adds L1 regularisation for both feature shrinkage and selection • Helps eliminate less important predictors by making the coefficients zero • Suitable for reducing model complexity

```
lasso = Lasso(max_iter=10000)

lasso_params = {

'alpha': [0.001, 0.01, 0.1, 1, 10]

}

lasso_gs = GridSearchCV(


lasso, 


lasso_params, 


cv=5, 


scoring='neg_mean_squared_error'

)

lasso_gs.fit(X_train_scaled, y_train)

lasso_model = lasso_gs.best_estimator_
```

```

y_pred_lasso = lasso_model.predict(X_test_scaled)

lasso_r2 = r2_score(y_test, y_pred_lasso)

lasso_rmse = mean_squared_error(y_test, y_pred_lasso, squared=False)

print("Lasso → R2:", lasso_r2, "RMSE:", lasso_rmse)

5. Decision Tree



- Non-linear model that splits the data into regions for prediction
- Capable of capturing complex interactions between features ,
- Requires careful depth control to avoid overfitting


elastic = ElasticNet(max_iter=10000)
elastic_params = {
    'alpha': [0.01, 0.1, 1, 10],
    'l1_ratio': [0.2, 0.5, 0.8]
}

elastic_gs = GridSearchCV(
    elastic,
    elastic_params,
    cv=5,
    scoring='neg_mean_squared_error'
)

elastic_gs.fit(X_train_scaled, y_train)

elastic_model = elastic_gs.best_estimator_
y_pred_elastic = elastic_model.predict(X_test_scaled)

elastic_r2 = r2_score(y_test, y_pred_elastic)
elastic_rmse = mean_squared_error(y_test, y_pred_elastic, squared=False)

print("ElasticNet → R2:", elastic_r2, "RMSE:", elastic_rmse)

```

## 6. Comparision

A performance comparison was done using evaluation metrics such as R<sup>2</sup> Score, Mean Absolute Error, and RMSE. Based on the comparison, Linear Regression and Ridge Regression produced the most accurate and stable predictions for the input data. These models showed strong generalisation capability and are therefore suitable for final deployment.

| Model             | Train MAE    | Test MAE  | Train RMSE   | Test RMSE | Train R <sup>2</sup> | Test R <sup>2</sup> |
|-------------------|--------------|-----------|--------------|-----------|----------------------|---------------------|
| Linear Regression | 6.687303e-01 | 0.4666670 | 1.272672e+00 | 0.621107  | 0.972473             | 0.990102            |
| Ridge Regression  | 6.688300e-01 | 0.466578  | 1.272676e+00 | 0.621105  | 0.972473             | 0.990102            |
| Lasso Regression  | 6.675052e-01 | 0.462696  | 1.276619e+00 | 0.628756  | 0.972302             | 0.989857            |
| Elastic Net       | 6.732460e-01 | 0.467471  | 1.278747e+00 | 0.632742  | 0.972210             | 0.989728            |
| Decision Tree     | 3.934424e-18 | 0.853061  | 1.711317e-17 | 1.219234  | 1.000000             | 0.961860            |

FIG 7 Comparison of Models Based on RSME, MAE, R<sup>2</sup>

