Infosys Springboard Virtual Internship 6.0 Completion Report

Project Title

Tempest FWI Predictor - A ML Model to Predict Fire Weather Index

Batch Number: 6
Start date: 10.11.2025
Name: Shruti Vilas Wani
Internship Duration: 8 Weeks

Table of Contents

ABSTRACT

In the context of increasing climate variability and rising forest fire incidents worldwide, accurate and timely assessment of fire risk has become a critical requirement for effective disaster management and prevention. This project presents a machine learning–based Fire Weather Index (FWI) Prediction System designed to estimate fire intensity and danger levels using meteorological and fire weather indicators. The system leverages the Algerian Forest Fires Dataset, which contains real-world observations from fire-prone regions, to model the relationship between environmental conditions and fire behavior.

A comprehensive data preprocessing and statistical analysis pipeline was employed, including feature selection, normalization, correlation analysis, and visualization techniques such as histograms and box plots. To address multicollinearity among correlated fire weather indices, a Ridge Regression model was adopted. The trained model demonstrated strong predictive performance, achieving a Mean Absolute Error (MAE) of 0.75, Mean Squared Error (MSE) of 1.24, Root Mean Squared Error (RMSE) of 1.11, and an R^2 score of 0.97, indicating high accuracy and robustness.

To enable practical real-world usage, the predictive model was deployed as an interactive Flask-based web application. The platform supports single-instance prediction, batch prediction through CSV uploads, and RESTful API access for system integration. Predicted FWI values are further classified into qualitative fire danger levels—Low, Moderate, High, and Extreme—to enhance interpretability for decision-makers and non-technical users. By integrating data-driven modeling with a scalable web interface, this system provides an end-to-end solution for forest fire risk assessment, supporting early warning, preparedness, and informed decision-making.

CHAPTER 1
DATASET DESCRIPTION

1.1 Overview

The **Algerian Forest Fires Dataset** is a real-world environmental dataset collected from two forest regions of Algeria, namely Bejaia and Sidi Bel-Abbes. The data was recorded during the period from June to September in 2012, which corresponds to the peak forest fire season in the region due to high temperatures and dry weather conditions. The dataset was developed to support the analysis and prediction of forest fire occurrence and fire intensity using weather-based indicators. It combines daily meteorological observations with fire danger indices derived from the internationally recognized Canadian Fire Weather Index (FWI) system, making it a reliable and widely used dataset in forest fire research and machine learning studies.

1.2 Temporal Features

Day: Represents the day of the month on which the observation was recorded.

Month: Indicates the month of observation, capturing seasonal variations in fire activity.

Year: Denotes the year of data collection (2012), providing chronological context.

These features help identify seasonal and short-term temporal patterns in forest fire occurrence.

1.3 Meteorological Features

Temperature (°C): Daily average air temperature. Higher temperatures contribute to drying of forest fuels, increasing the likelihood of fire ignition.

Relative Humidity (RH %): Measures the moisture content in the air. Lower humidity levels reduce fuel moisture and increase fire risk.

Wind Speed (Ws km/h): Indicates the average wind speed. Wind plays a critical role in accelerating fire spread and increasing fire intensity.

Rain (mm): Represents daily rainfall. Rainfall increases fuel moisture and generally reduces the probability of fire occurrence.

Meteorological variables directly influence fuel dryness and fire behavior, making them essential predictors.

1.4 Fire Weather Index (FWI) System Components

The Fire Weather Index system is a widely adopted standard for estimating fire danger based on weather conditions.

FFMC (Fine Fuel Moisture Code): Measures the moisture content of surface litter and fine fuels. Higher values indicate easier ignition.

DMC (Duff Moisture Code): Represents moisture content of moderately deep organic layers. High values indicate prolonged drying conditions.

DC (Drought Code): Reflects long-term dryness of deep organic layers and drought effects.

ISI (Initial Spread Index): Combines wind speed and FFMC to estimate the potential rate of fire spread.

BUI (Buildup Index): Combines DMC and DC to represent the total amount of fuel available for combustion.

FWI (Fire Weather Index): The final index indicating overall fire intensity and danger level. Higher FWI values correspond to more severe and uncontrollable fires.

These indices provide domain-specific insights into fire ignition potential, spread rate, and intensity.

1.5 Target Variables

Classes: A categorical variable indicating fire occurrence, labeled as *fire* or *not fire*. This variable is used for classification tasks.

FWI: A numerical variable representing fire intensity, commonly used as the target for regression-based prediction models.

Region: Identifies the geographical region of data collection (Bejaia or Sidi Bel-Abbes), capturing regional climatic differences.

CHAPTER 2

DATA PREPROCESSING

2.1 Dataset Loading and Initial Inspection

The cleaned Algerian Forest Fires dataset was loaded using the Pandas library. An initial inspection was performed to understand the dataset structure, including the number of records, feature types, and overall data consistency. This step ensured that the dataset did not contain missing values or invalid entries and was suitable for further preprocessing and analysis.

2.2 Feature Selection

To align the dataset with the regression objective, only relevant numerical features were retained. The categorical variable **Classes** was removed, as the focus of this phase was on predicting the continuous **Fire Weather Index (FWI)** rather than performing classification. Additionally, temporal features such as **day, month, and year**, along with the **Region** variable, were excluded to prevent the model from learning time- or location-specific biases and to ensure that predictions were driven primarily by meteorological conditions and fire weather indices.

2.3 Target Variable Identification

The Fire Weather Index (FWI) was selected as the target variable for prediction, as it represents overall fire intensity and danger. All remaining features were treated as input variables. This clear separation of input and output variables was necessary for supervised learning and regression model training.

2.4 Train–Test Split

The dataset was divided into training and testing sets to allow objective evaluation of model performance. The training set was used to learn model parameters, while the testing set was reserved exclusively for performance evaluation. This step ensured that the model's generalization capability could be assessed on unseen data.

2.5 Feature Scaling (StandardScaler)

Since the dataset contains features with varying numerical ranges such as temperature, wind speed, and drought indices—feature scaling was applied using StandardScaler. This technique standardizes features to have zero mean and unit variance. Scaling was performed to prevent features with larger magnitudes from dominating the learning process and to improve the convergence and stability of regression models.

2.6 Exploratory Data Analysis (EDA)

Exploratory Data Analysis was conducted to understand feature distributions and relationships. Histograms were used to visualize the spread and skewness of individual variables, while correlation analysis was performed to examine relationships between input features and the target variable (FWI). The analysis showed that indices such as **ISI and DMC** are positively correlated with FWI, whereas **Relative Humidity (RH) and Rain** exhibit negative correlation, confirming their inverse influence on fire intensity.

2.7 Preprocessing Outcome

After completing feature selection, scaling, and exploratory analysis, the dataset was transformed into a clean, normalized, and well-structured form suitable for regression modeling. These preprocessing steps reduced noise, improved feature interpretability, and enhanced the model's ability to learn meaningful patterns related to forest fire behavior.

CHAPTER 3

DATASET STATISTICS

3.1 Descriptive Statistical Analysis

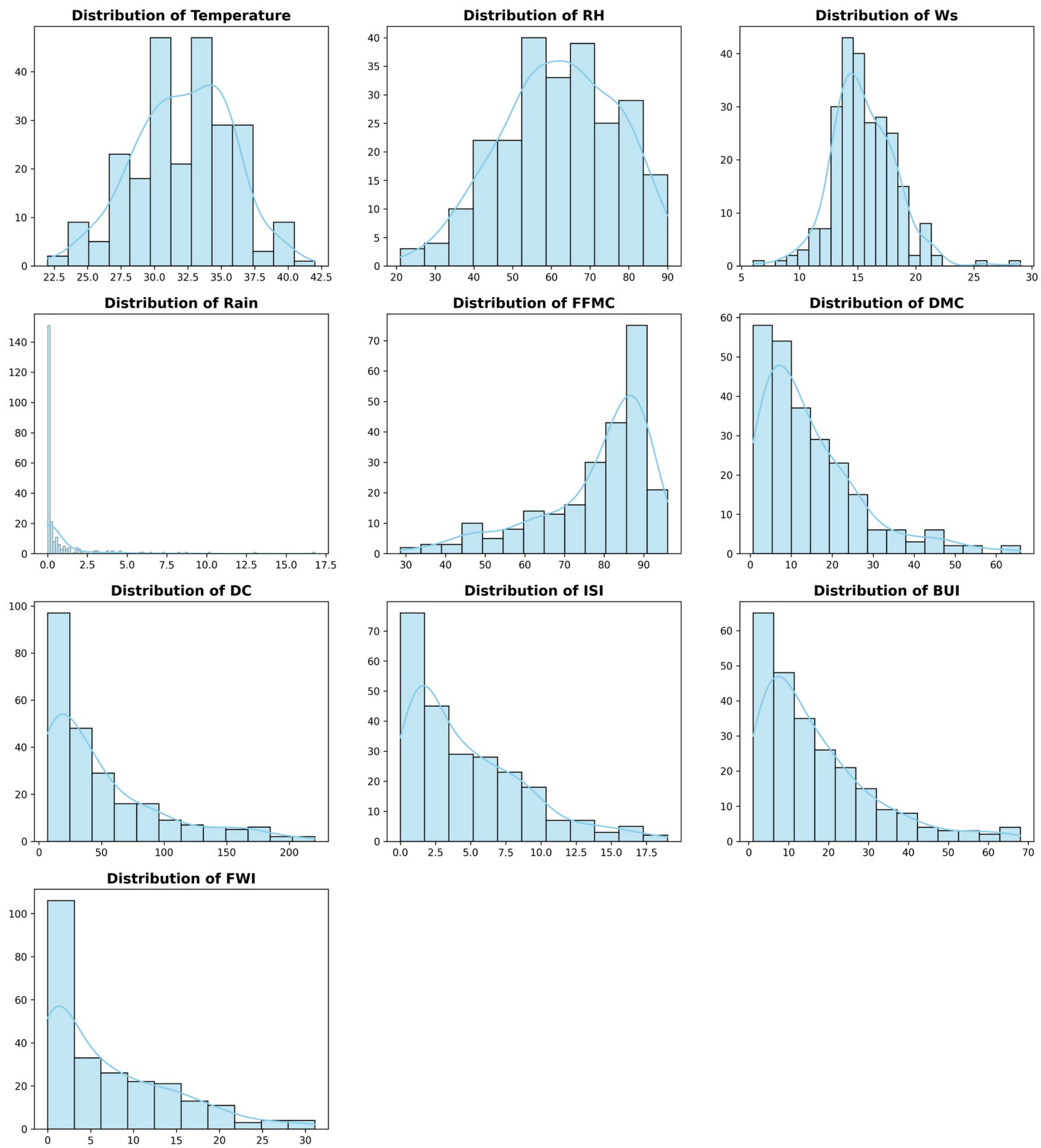
A descriptive statistical analysis was performed to summarize the central tendency and dispersion of the dataset features. Key statistical measures such as **mean, median, standard deviation, minimum, and maximum values** were computed for all numerical variables. This analysis provided an overall understanding of the range and variability of meteorological variables and fire weather indices. Features such as temperature, wind speed, and fire indices showed noticeable variation, reflecting changing environmental and fire conditions across different days.

	mean	**std**	**min**	**25%**	**50%**	**75%**	**max**
day	15.761317	8.842552	1.0	8.00	16.0	23.00	31.0
month	7.502058	1.114793	6.0	7.00	8.0	8.00	9.0
year	2012.000000	0.000000	2012.0	2012.00	2012.0	2012.00	2012.0
Temperature	32.152263	3.628039	22.0	30.00	32.0	35.00	42.0
RH	62.041152	14.828160	21.0	52.50	63.0	73.50	90.0
Ws	15.493827	2.811385	6.0	14.00	15.0	17.00	29.0
Rain	0.762963	2.003207	0.0	0.00	0.0	0.50	16.8
FFMC	77.842387	14.349641	28.6	71.85	83.3	88.30	96.0
DMC	14.680658	12.393040	0.7	5.80	11.3	20.80	65.9
DC	49.430864	47.665606	6.9	12.35	33.1	69.10	220.4
ISI	4.742387	4.154234	0.0	1.40	3.5	7.25	19.0
BUI	16.690535	14.228421	1.1	6.00	12.4	22.65	68.0
FWI	7.035391	7.440568	0.0	0.70	4.2	11.45	31.1
Region	0.497942	0.501028	0.0	0.00	0.0	1.00	1.0

3.2 Distribution Analysis Using Histograms

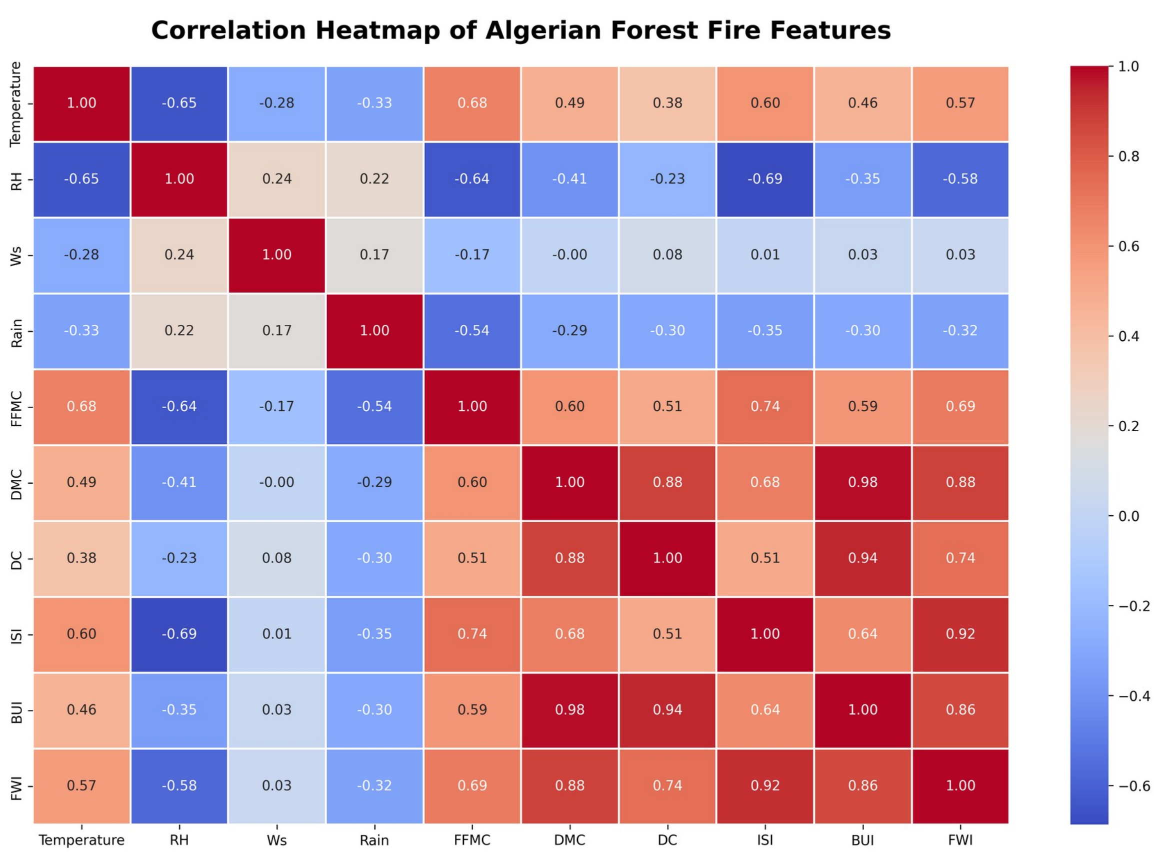
Histograms were used to analyze the distribution of individual features in the dataset. This visualization technique helped identify the shape, spread, and skewness of the data. Meteorological variables such as temperature and wind speed showed relatively smooth distributions, while fire weather indices like

DMC, DC, and FWI exhibited right-skewed distributions, indicating that higher fire danger values occur less frequently. Distribution analysis is important for understanding data behavior and for identifying potential transformations if required for modeling.



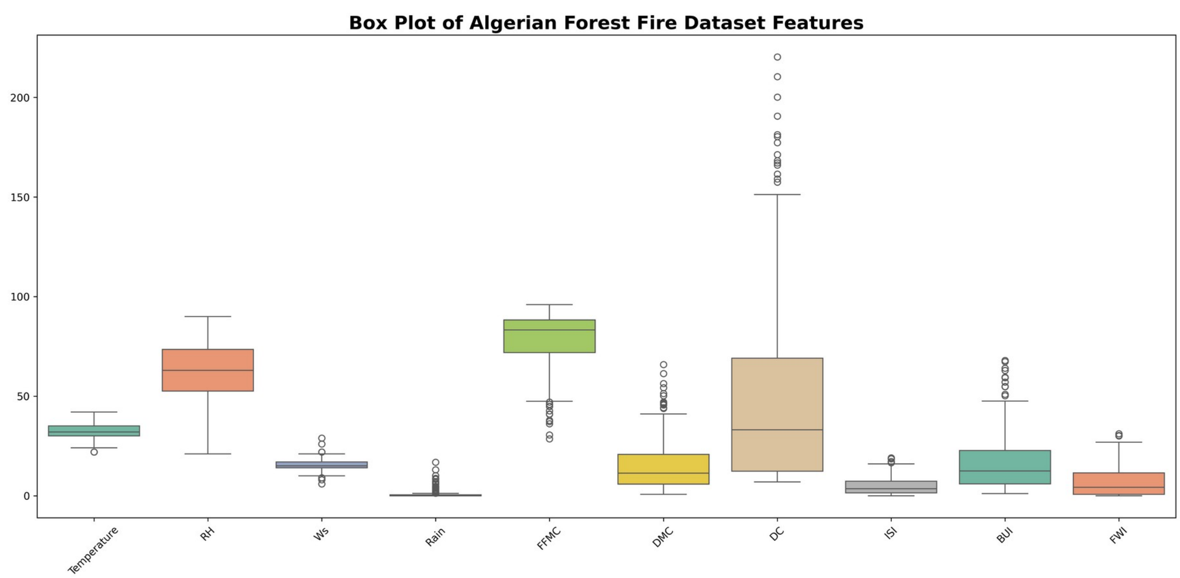
3.3 Correlation Analysis

Correlation analysis was conducted to examine the linear relationships between input features and the target variable **FWI**. A correlation matrix was computed to quantify the strength and direction of relationships between variables. The analysis revealed that **ISI, DMC, DC, and BUI** have strong positive correlations with FWI, indicating that increases in these indices are associated with higher fire intensity. In contrast, **Relative Humidity (RH)** and **Rain** showed negative correlations with FWI, confirming that higher moisture levels reduce fire danger. This correlation analysis helped identify the most influential features contributing to forest fire intensity.



3.4 Box Plot Analysis

Box plot analysis was conducted to examine the distribution, variability, and presence of outliers among the dataset features. The visualization highlighted that meteorological variables such as **Temperature, Relative Humidity (RH), and Wind Speed (Ws)** have relatively compact interquartile ranges, indicating stable conditions across most observations, while **Rain** shows a highly skewed distribution with a median value of zero and several upper outliers due to infrequent but intense rainfall events. Fire Weather Index components including **DMC, DC, ISI, BUI, and FWI** exhibit wider interquartile ranges and prominent upper outliers, reflecting significant variation in fire danger levels. In particular, **DC and FWI** display extreme values beyond the upper quartile, suggesting that although most days experience low to moderate fire risk, a smaller number of days correspond to very high fire intensity. This box plot analysis confirms the presence of variability and outliers in the dataset and justifies the use of feature scaling and robust regression models for forest fire prediction.



3.5 Interpretation of Statistical Findings

The statistical analysis confirmed that both meteorological variables and fire weather indices play significant roles in determining forest fire risk. Positively correlated features contribute to fuel dryness and fire spread, while negatively correlated features suppress fire activity. These insights guided feature selection and justified the use of regression models, as the relationships between predictors and the target variable were found to be meaningful and consistent with domain knowledge.

CHAPTER 4

MODELS

4.1 Ridge Regression

Ridge Regression is a regularized linear regression technique used to address the problem of overfitting and multicollinearity in datasets with correlated input features. It extends ordinary least squares (OLS) regression by introducing an additional penalty term to the loss function. This penalty term is proportional to the square of the magnitude of the regression coefficients, which discourages the model from assigning excessively large weights to any single feature. As a result, Ridge Regression produces more stable and generalized models, especially when predictor variables exhibit strong correlations, as observed in fire weather indices such as DMC, DC, ISI, and BUI.

4.2 Alpha Tuning

Mathematically, Ridge Regression minimizes the sum of squared errors along with an L2 regularization term controlled by a hyperparameter known as **alpha (α)**. The alpha parameter determines the strength of regularization applied to the model. When alpha is set to zero, Ridge Regression behaves like standard linear regression. As the value of alpha increases, the penalty on large coefficients becomes stronger, leading to smaller and more evenly distributed feature weights. In this study, alpha tuning was performed to identify an optimal balance between bias and variance. Lower alpha values allow the model to fit the training data closely, while higher values reduce overfitting by constraining coefficient magnitudes.Tuned alpha value helped improve generalization performance on unseen test data without significantly increasing prediction error

4.3 Mean Absolute Error (MAE)

Mean Absolute Error measures the average absolute difference between the predicted and actual values of the Fire Weather Index (FWI). It provides a clear and intuitive measure of prediction accuracy by treating all errors equally, without giving extra weight to larger errors. In this study, the MAE value obtained was **0.7534**, indicating that the model's predictions deviate from the true FWI values by less than one unit on average, reflecting high prediction accuracy.

4.4 Mean Squared Error (MSE)

Mean Squared Error calculates the average of the squared differences between the predicted and actual values. By squaring the errors, MSE penalizes larger prediction errors more heavily, making it sensitive to outliers. The Ridge Regression model achieved an MSE value of **1.2477**, which suggests that large deviations between predicted and actual FWI values are minimal and well controlled.

4.5 Root Mean Squared Error (RMSE)

Root Mean Squared Error is the square root of the Mean Squared Error and expresses the error in the same units as the target variable. RMSE provides an interpretable measure of the typical magnitude of prediction errors. The RMSE value obtained was **1.1170**, indicating that, on average, the predicted FWI values differ from the actual values by approximately 1.1 units.

4.6 R-squared (R^2 Score)

The R-squared score measures the proportion of variance in the target variable that is explained by the model. It ranges from 0 to 1, with higher values indicating better model performance. In this study, the Ridge Regression model achieved an R^2 score of **0.9765**, meaning that **97.65%** of the variability in the Fire Weather Index is explained by the input features. This high value demonstrates the strong explanatory power and effectiveness of the model.

4.7 Overall Interpretation

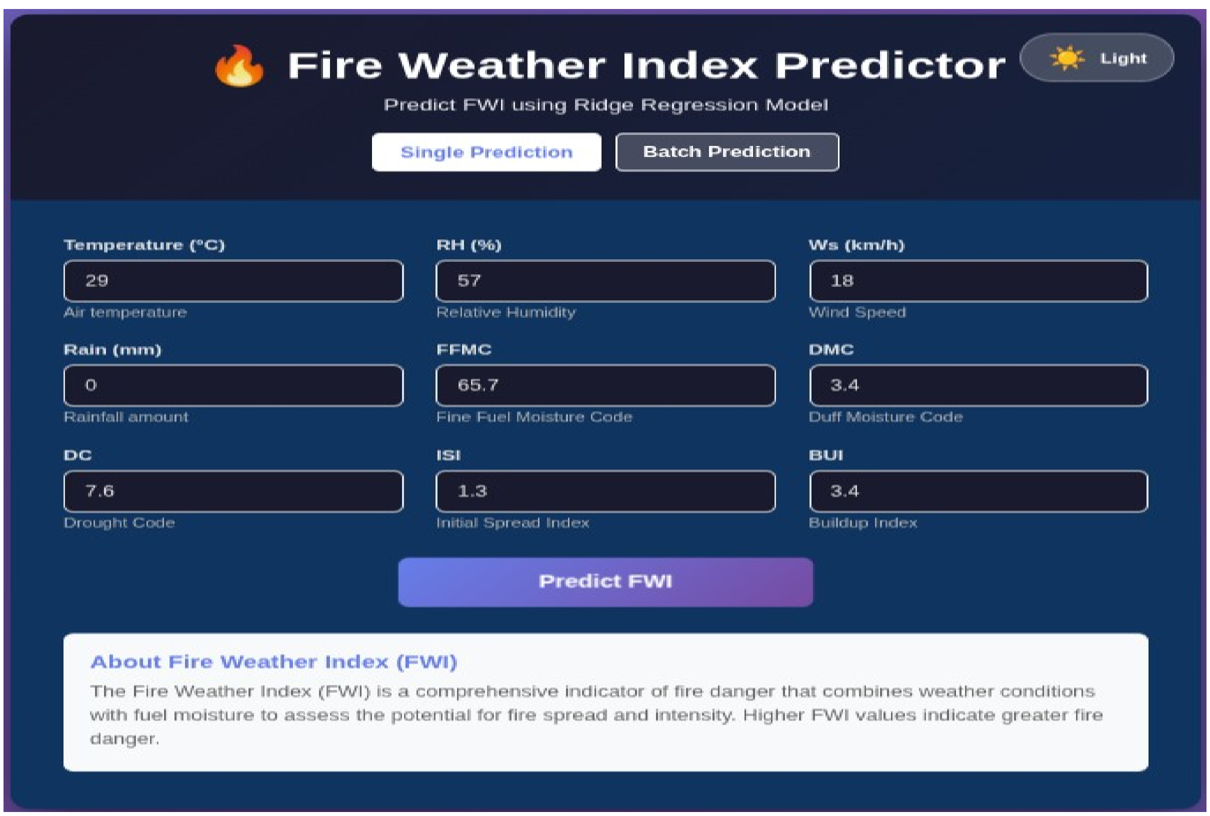
The low error values (MAE, MSE, and RMSE) combined with a high R^2 score indicate that the Ridge Regression model performs exceptionally well in predicting fire intensity. These metrics collectively confirm the robustness, accuracy, and generalization capability of the model.

CHAPTER 5

FLASK APPLICATION DESIGN AND IMPLEMENTATION

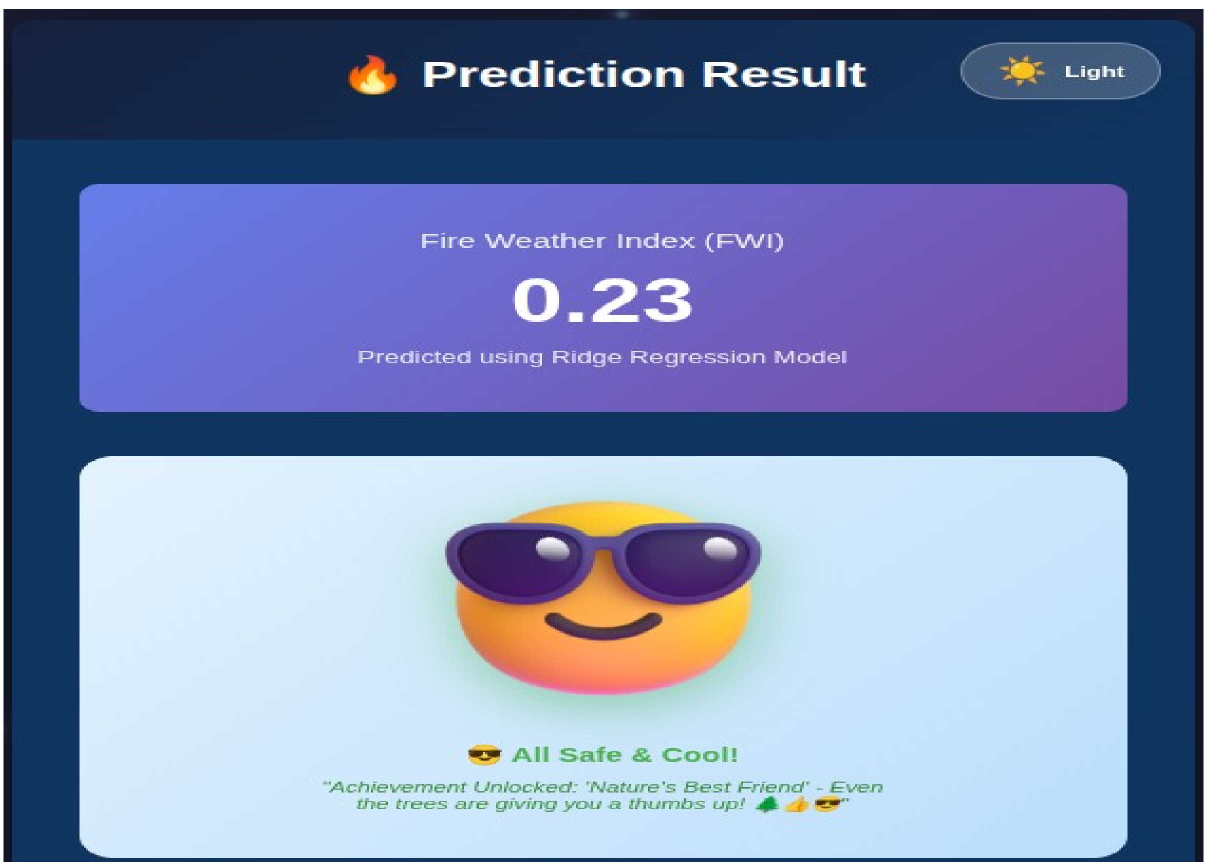
5.1 Home Route (Single Prediction Interface)

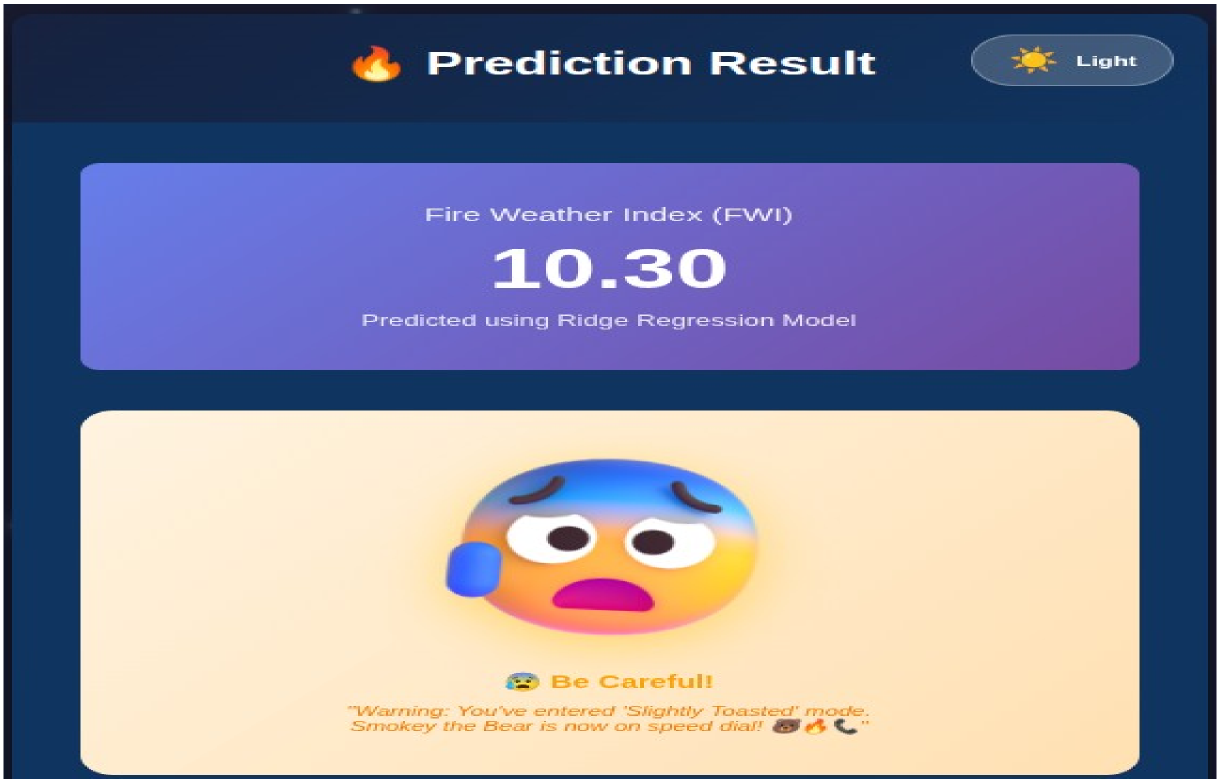
The root route (/) serves as the home page of the application and renders the main user interface for single FWI prediction. This page displays an input form where users manually enter meteorological and fire index parameters such as temperature, relative humidity, wind speed, rainfall, and fire weather indices. The route does not perform any computation but prepares the feature list required for prediction and passes it to the frontend template for rendering. This route acts as the entry point of the system and ensures a user-friendly interaction with the deployed model.

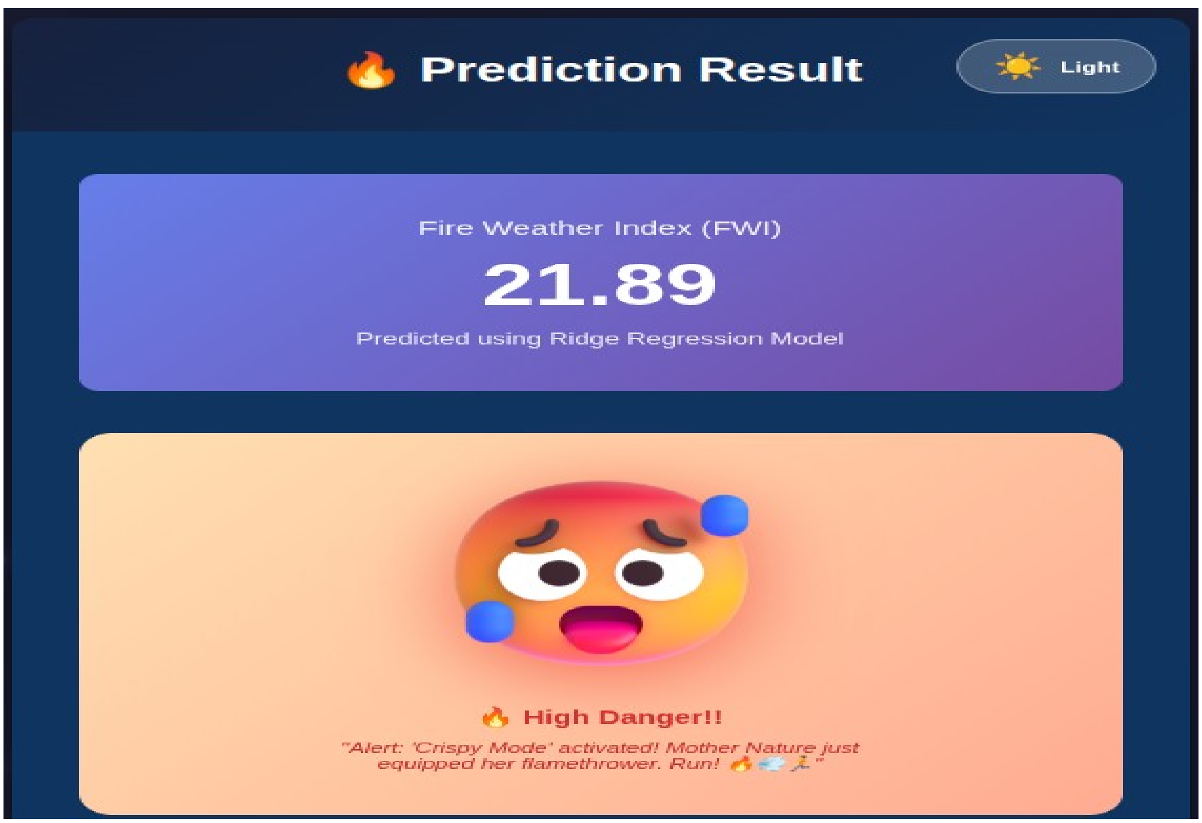


5.2 Single Prediction Route

The /predict route handles single-instance FWI prediction requests submitted through the web form or as JSON input. Upon receiving the request, the application extracts the input features in a predefined order to match the model's training configuration. These values are converted into numerical format and reshaped into a suitable array for model input. The features are then scaled using the preloaded StandardScaler to maintain consistency with training preprocessing. The scaled input is passed to the trained Ridge Regression model, which generates the predicted FWI value. The result is displayed to the user along with the input feature summary, while errors are handled gracefully using a dedicated error page .

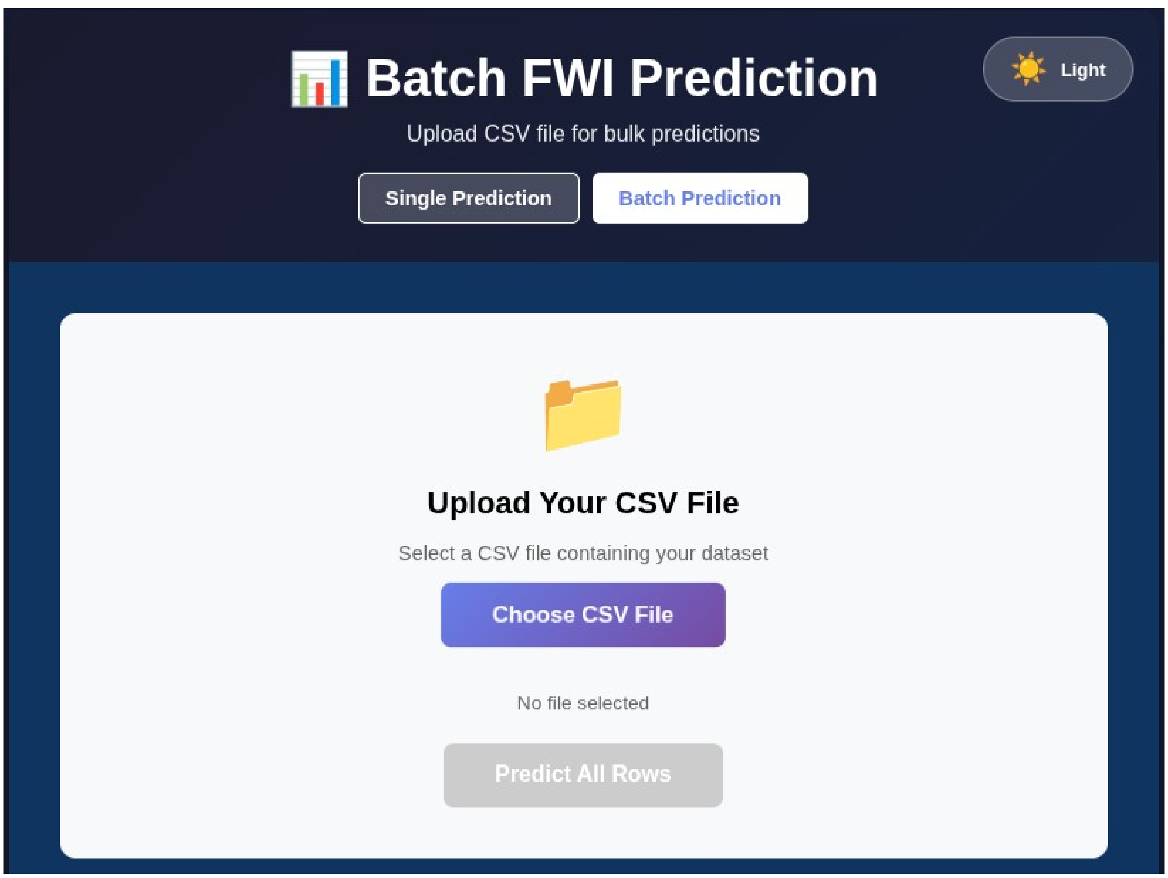


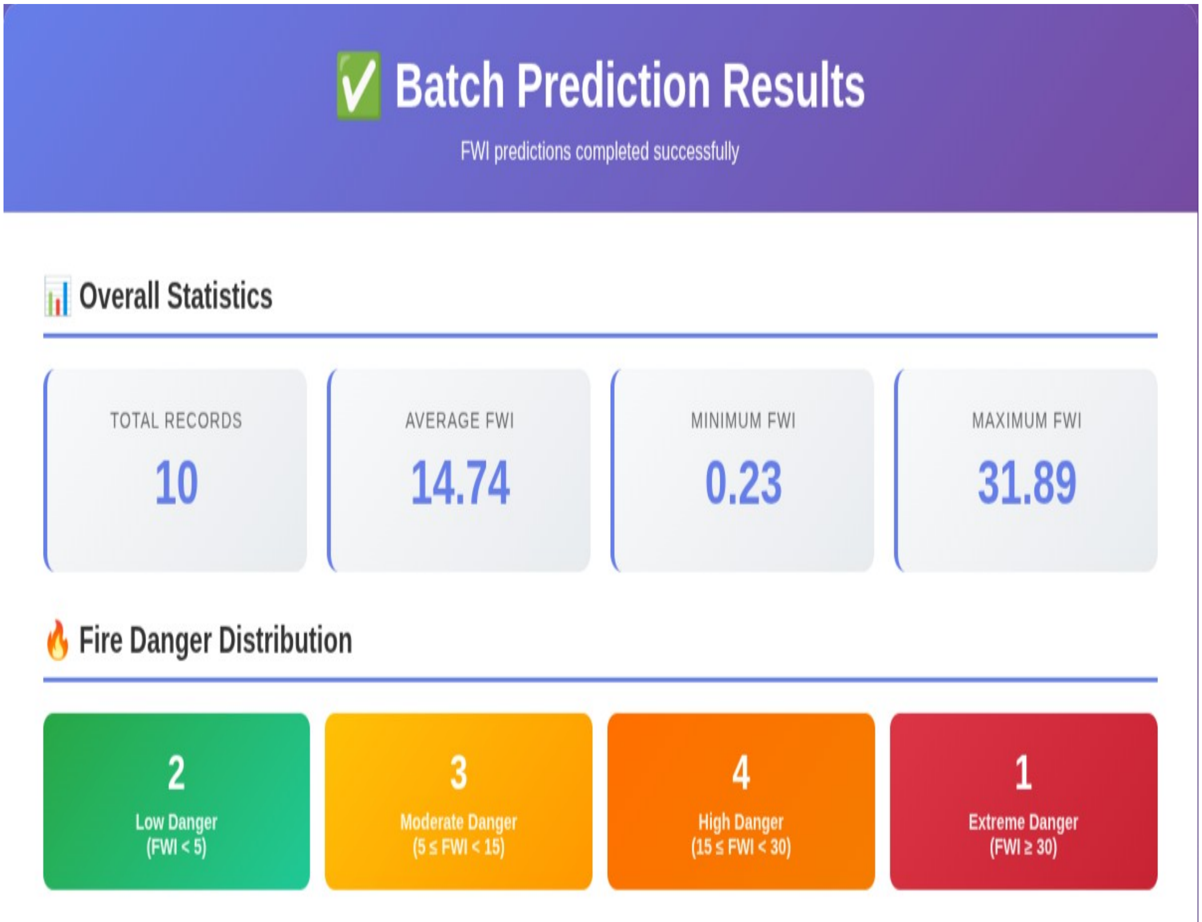

🔥 Prediction Result
☀️ Light
Fire Weather Index (FWI)
10.30
Predicted using Ridge Regression Model
😰 Be Careful!
"Warning: You've entered 'Slightly Toasted' mode. Smokey the Bear is now on speed dial! 🐻🔥📞"


🔥 Prediction Result
☀️ Light
Fire Weather Index (FWI)
21.89
Predicted using Ridge Regression Model
🔥 High Danger!!
"Alert: 'Crispy Mode' activated! Mother Nature just equipped her flamethrower. Run! 🔥💨🏃"

5.3 Batch Prediction Route

The /batch_predict route enables bulk prediction by allowing users to upload a CSV file containing multiple records. When accessed via a GET request, it renders a file upload interface. Upon receiving a POST request, the application validates the uploaded file format, checks for missing required columns, and ensures that no null values are present in the dataset. The required features are extracted and scaled using the same StandardScaler used during training. The Ridge Regression model then generates predictions for all records in a single operation. The predicted FWI values are appended to the dataset along with a categorical fire danger level, enabling large-scale analysis in one step .





5.4 Download Results Route

The /download/<filename> route allows users to download the batch prediction results as a CSV file. After batch processing, the predicted results are temporarily stored on the server. This route retrieves the file securely and sends it to the user as a downloadable attachment. Error handling is implemented to manage cases where the requested file does not exist or cannot be accessed.

5.5 Error Handling and Validation

The application includes a centralized error-handling mechanism to manage invalid inputs, missing files, unsupported formats, and runtime exceptions. Whenever an error occurs during prediction or file processing, a dedicated error page is rendered with a descriptive message. This ensures application robustness and improves the overall user experience by preventing abrupt failures