TITLE: Fire Weather Index Predictor

(A Machine Learning Model To predict Fire Weather Index)

Infosys SpringBoard Virtual Internship Program 6.0

Submitted By

Manyatha N

Under the guidance of Mentor **Praveen**

Project Statement

The increasing frequency and destructiveness of wildfires puts ecosystems, human populations, and financial resources at danger. The Fire Weather Index (FWI), a widely used predictor of wildfire threat, is based on meteorological elements such as temperature, relative humidity, wind speed, rainfall, FFMC, DMC, ISI, and region. However, the substantial reliance of conventional FWI estimates on human procedures and static models limits the speed and accuracy of early-warning systems. The project's objective is to create a machine learning-based FWI prediction model that uses a whole pipeline that includes feature engineering, data preprocessing, StandardScaler-based normalization, and Ridge Regression modeling to assess real-time environmental data. In order to facilitate rapid user input and automatic wildfire-risk prediction, the system is further integrated into a Flask web application, assisting emergency planners, forest departments, and climate researchers in proactive wildfire management.

Expected Outcomes

- A machine learning model based on Ridge Regression that can correctly forecast the Fire Weather Index (FWI).
- A complete preparation pipeline that prepares and normalizes environmental data using StandardScaler.
- A functional Flask web application that lets users enter values and get FWI forecasts in real time.
- An intelligent, data-driven wildfire risk assessment tool for researchers and authorities.

Modules to be Implemented

1. Data Collection
2. Data Exploration & Data Preprocessing
3. Feature Engineering & Scaling
4. Model Training using Ridge Regression
5. Evaluation & Optimization
6. Deployment via Flask App
7. Presentation & Documentation

System Requirements

1. Python
2. Pandas
3. Scikit-learn (Ridge Regression, StandardScaler)
4. Matplotlib / Visualization tools (used in EDA)
5. Flask (for deployment)
6. Pickle (.pkl files) (ridge.pkl, scaler.pkl)
7. Git / GitHub (for final submission)

Milestone 1

Module 1

1. Data Collection

The primary goal of is to gather and prepare the fundamental dataset needed to develop the Fire Weather Index (FWI) prediction model. In this phase, a structured dataset is collected that includes the FWI target variable and important environmental characteristics like temperature, relative humidity, wind speed, rainfall, FFMC, DMC, ISI, and region. After that, the data is examined to make sure the formatting, data types, and consistency are correct. Checking sure each column is appropriately represented and prepared for analysis is part of this. The cleaned dataset is next put into a Pandas DataFrame, which serves as the foundation for all further preprocessing, model construction, and exploration procedures.

Loading and Inspecting the FWI Dataset (Head, Info, and Summary Statistics)

```
import pandas as pd

file_path = 'FWI_Dataset.csv'
df = pd.read_csv(file_path)

print(df.head())
print(df.info())
print(df.describe())
```

Output

- Check for the attributes in raw dataset
- Drop unwanted / unnecessary attributes (class, DC, Region)

```
   day  month  year  Temperature   RH  Ws  Rain   FFMC  DMC    DC  ISI  BUI  \
0    1      6  2012           29   57  18   0.0  65.7  3.4   7.6  1.3  3.4
1    2      6  2012           29   61  13   1.3  64.4  4.1   7.6  1.0  3.9
2    3      6  2012           26   82  22  13.1  47.1  2.5   7.1  0.3  2.7
3    4      6  2012           25   89  13   2.5  28.6  1.3   6.9  0.0  1.7
4    5      6  2012           27   77  16   0.0  64.8  3.0  14.2  1.2  3.9

   FWI    Classes    Region
0  0.5  not fire     Bejaia
1  0.4  not fire     Bejaia
2  0.1  not fire     Bejaia
3    0  not fire     Bejaia
4  0.5  not fire     Bejaia
<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
```

FIG – 1: Basic information of a dataset

Cleaned_FWI_Dataset:

- Cleaned FWI dataset generated.

	day	month	year	Temperature	RH	Ws	\
count	244.000000	244.000000	244.0	244.000000	244.000000	244.000000	
mean	15.754098	7.500000	2012.0	32.172131	61.938525	15.504098	
std	8.825059	1.112961	0.0	3.633843	14.884200	2.810178	
min	1.000000	6.000000	2012.0	22.000000	21.000000	6.000000	
25%	8.000000	7.000000	2012.0	30.000000	52.000000	14.000000	
50%	16.000000	7.500000	2012.0	32.000000	63.000000	15.000000	
75%	23.000000	8.000000	2012.0	35.000000	73.250000	17.000000	
max	31.000000	9.000000	2012.0	42.000000	90.000000	29.000000	

	Rain	FFMC	DMC	ISI	BUI
count	244.000000	244.000000	244.000000	244.000000	244.000000
mean	0.760656	77.887705	14.673361	4.774180	16.664754
std	1.999406	14.337571	12.368039	4.175318	14.204824
min	0.000000	28.600000	0.700000	0.000000	1.100000
25%	0.000000	72.075000	5.800000	1.400000	6.000000
50%	0.000000	83.500000	11.300000	3.500000	12.250000
75%	0.500000	88.300000	20.750000	7.300000	22.525000
max	16.800000	96.000000	65.900000	19.000000	68.000000

FIG – 2: Descriptive Statistics of the Collected Environmental Dataset

Module 2

2. Data Preprocessing

Data preprocessing is a crucial step that prepares the collected dataset for reliable model training. At this point, the dataset is carefully examined for null or missing values, which are treated properly to maintain data integrity. To avoid skewed model behavior, boxplots and statistical criteria are used to identify outliers. Histograms and density plots are used to assess the feature distributions in order to spot trends and problems with data quality. Relationships between variables are explored using correlation matrices and scatterplots to understand how environmental features influence the Fire Weather Index (FWI). Additionally, label encoding or mapping techniques are used to encode categorical data like region so that machine learning algorithms may use them. After cleaning, transforming, and validating all features, the processed dataset is saved for use in the next stages of feature engineering and modeling.

Data Preprocessing: Missing-value handling, outlier treatment, encoding, and scaling

- **Libraries and modules used from python**

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import LabelEncoder
```

- **Missing-value handling**

```
for col in numerical_cols:
       if col in df.columns:
              df[col] = pd.to_numeric(df[col].astype(str).
              str.strip(), errors='coerce')
              df[col].fillna(df[col].median(), inplace=True)
```

```
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
day                     0
month                   0
year                    0
Temperature             0
Relative Humidity       0
Wind Speed              0
Rain                    0
FFMC                    0
DMC                     0
DC                      0
ISI                     0
BUI                     0
FWI                     0
Classes                 1
Region                  0
dtype: int64
```

FIG-3: Missing Values Summary of the Dataset

- **Histogram statistics**

```
for col in numerical_cols:
    if col in df.columns:
        sns.histplot(df[col], kde=True)
        plt.show()
```

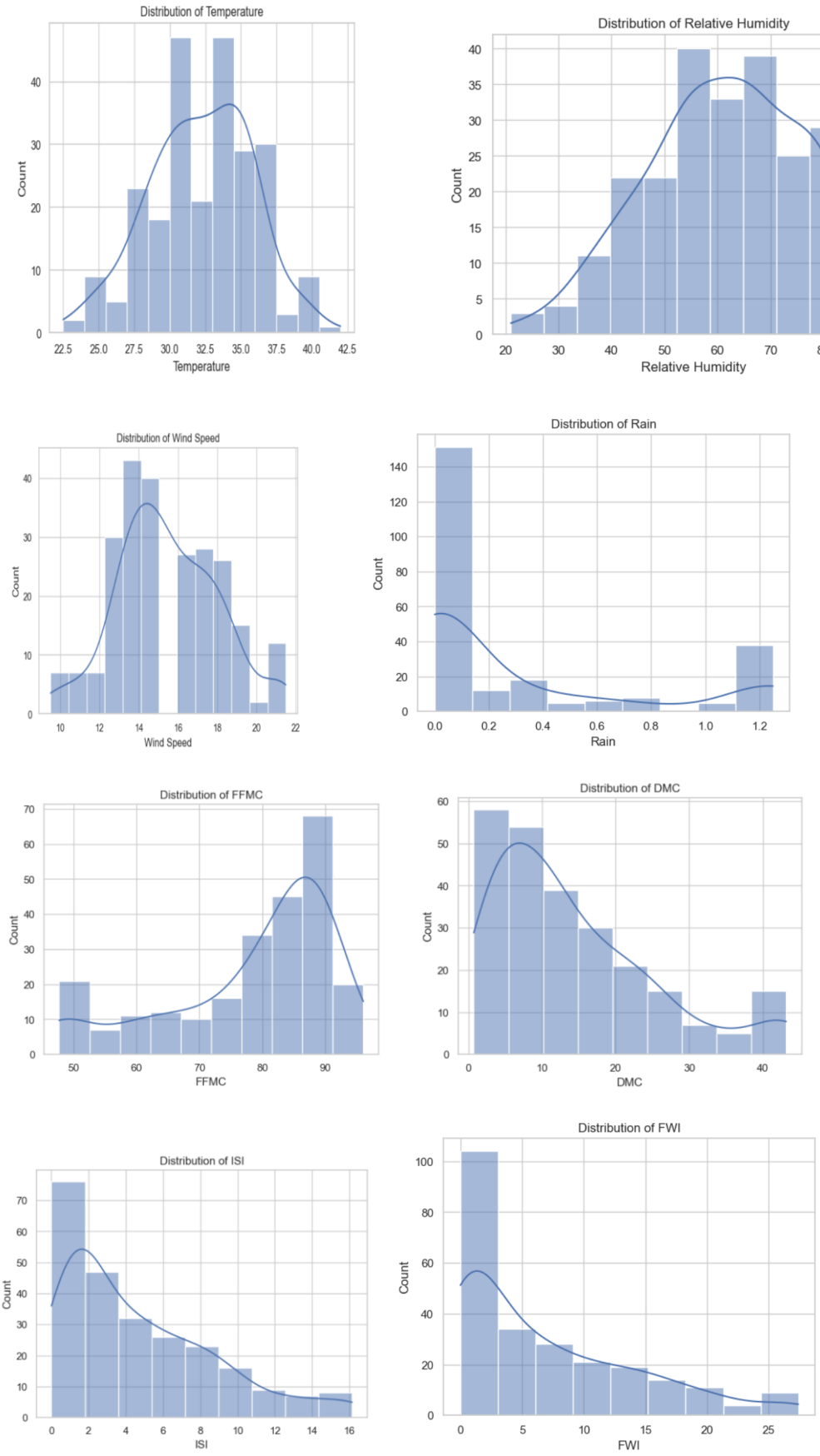


FIG-4: Histograms of statistics

- **Correlation matrix**

```
corr_matrix = df[corr_cols].corr()
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm')
plt.show()
```

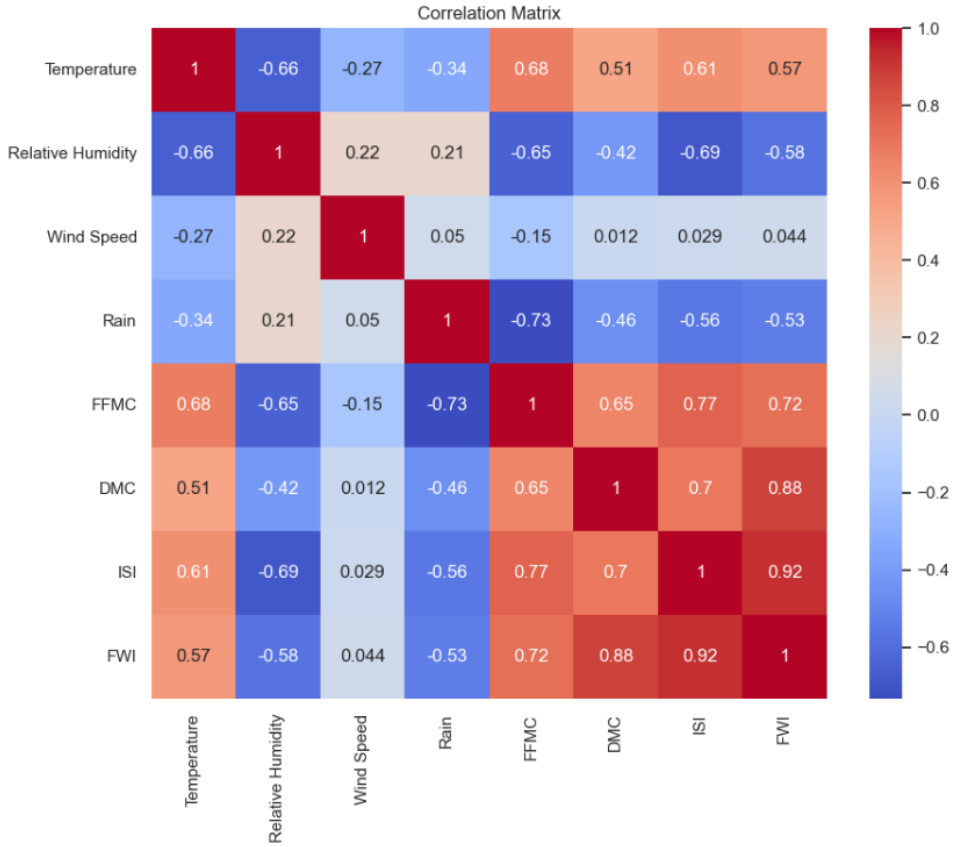


FIG-5: Correlation matrix

- **Scatter plot of FWI**

```
if {'Temperature', 'FWI'}.issubset(df.columns):
    sns.scatterplot(x='Temperature', y='FWI', data=df)
    plt.show()
```

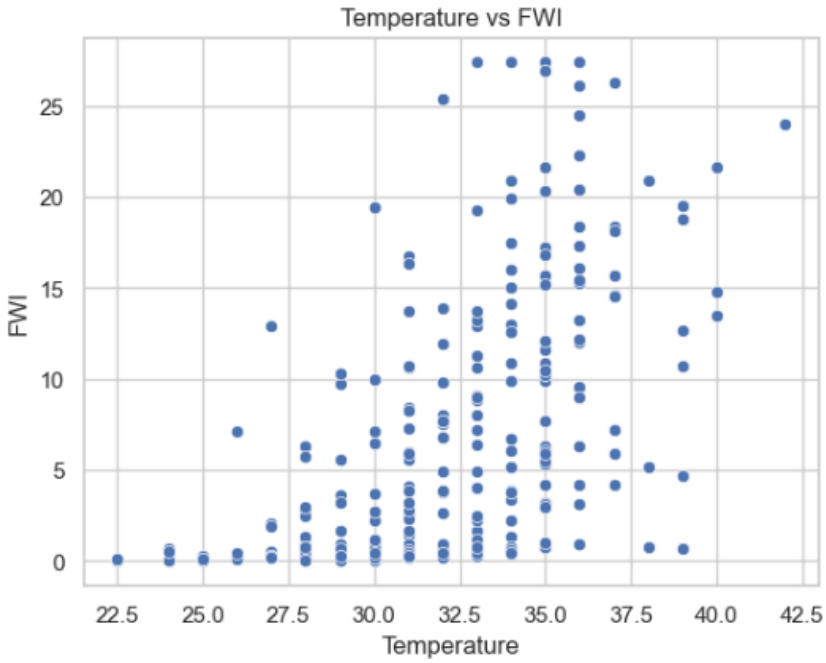


FIG-6: Scatter plot for features of FWI

- **Preview after preprocessing**

```
print(df.info())
print(df.head())
```

```
<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    float64
 4   Relative Humidity  244 non-null    float64
 5   Wind Speed         244 non-null    float64
 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    float64
 13  Classes            243 non-null    object
 14  Region_encoded     244 non-null    int32
dtypes: float64(9), int32(1), int64(3), object(2)
memory usage: 27.8+ KB
None
   day  month  year  Temperature  Relative Humidity  Wind Speed  Rain  \
0    1      6  2012         29.0               57.0        18.0  0.00
1    2      6  2012         29.0               61.0        13.0  1.25
2    3      6  2012         26.0               82.0        21.5  1.25
3    4      6  2012         25.0               89.0        13.0  1.25
4    5      6  2012         27.0               77.0        16.0  0.00

      FFMC  DMC    DC  ISI  BUI  FWI      Classes  Region_encoded
0  65.7000  3.4   7.6  1.3  3.4  0.5  not fire                 0
1  64.4000  4.1   7.6  1.0  3.9  0.4  not fire                 0
2  47.7375  2.5   7.1  0.3  2.7  0.1  not fire                 0
3  47.7375  1.3   6.9  0.0  1.7  0.0  not fire                 0
4  64.8000  3.0  14.2  1.2  3.9  0.5  not fire                 0
```

FIG-7: Dataset Information and Preview After Preprocessing

Milestone 2

Module 3

3. Feature Engineering and Scaling

Feature Engineering and Scaling involves selecting the key input features that are most closely related to the Fire Weather Index (FWI) target variable to improve the performance of the model. Selecting relevant features helps reduce noise and ensures that the model focuses on important information. StandardScaler is used to normalize the dataset's numerical features, reducing all values to a consistent scale and preventing any one characteristic from influencing the learning process. After scaling, the dataset is divided into input features (X) and the target variable (y). The data is then split into training and testing sets using train_test_split to allow proper evaluation of the model. Finally, the trained scaler is saved as a .pkl file to maintain consistency between training and deployment stages.

- **Feature Selection Based on Correlation with FWI**

```
target = 'FWI'
numeric_df = df.apply(pd.to_numeric, errors='coerce')
correlations=numeric_df.corr()[target].abs().sort_va
lues(ascending=False)
top_features = correlations.index[1:6]
```

Output

- ISI, DMC, BUI, DC, and FFMC show strong correlation with FWI.
- Other features show weak correlation or contain NaN values.

```
FWI                 1.000000
ISI                 0.922941
DMC                 0.877739
BUI                 0.855623
DC                  0.739009
FFMC                0.722423
Relative Humidity   0.575836
Temperature         0.570230
Rain                0.533121
day                 0.350713
Region_encoded      0.195407
month               0.083244
Wind Speed          0.044176
year                     NaN
Classes                  NaN
```

- **Separation of Input Features (X) and Target Variable (y)**

```
X = numeric_df[top_features]
y = numeric_df[target]
```

Output

- Input features (X): ISI, DMC, BUI, DC, FFMC
- Target variable (y): FWI

- **Normalization of Numerical Features**

```
scaler_params = {}
X_scaled = X.copy()
for col in X.columns:
    mean_val = X[col].mean()
    std_val = X[col].std()
    if std_val == 0 or np.isnan(std_val):
        std_val = 1.0
    scaler_params[col] = {'mean': mean_val, 'std': std_val}
    X_scaled[col] = (X[col] - mean_val) / std_val
```

Output

- Numerical features normalized to a common scale.
- Scaling parameters stored for each feature.

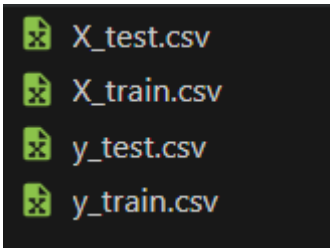
```
{
 'ISI': {
    'mean': np.float64(4.74672131147541),
    'std':  np.float64(4.0919057796664156)
 },
 'DMC': {
    'mean': np.float64(14.286065573770491),
    'std':  np.float64(11.194393460586191)
 },
 'BUI': {
    'mean': np.float64(16.664754098360657),
    'std':  np.float64(14.204823977055081)
 },
 'DC': {
    'mean': np.float64(49.430864197530866),
    'std':  np.float64(47.66560598458993)
 },
 'FFMC': {
    'mean': np.float64(78.28811475409836),
    'std':  np.float64(13.283492963829604)
 }
}
```

- **Train–Test Split of the Dataset**

```
n_test = int(n_samples * test_size)
indices = np.arange(n_samples)
np.random.shuffle(indices)
X_train = X_scaled.iloc[train_indices]
X_test = X_scaled.iloc[test_indices]
y_train = y.iloc[train_indices]
y_test = y.iloc[test_indices]
print(len(X_train), len(X_test))
```

Output

```
196 48
```



- **Ensured the scaler was saved as a .pkl file**

```
with open('scaler.pkl', 'wb') as f:
    pickle.dump(scaler_params, f)
```

Output

Module 4

4. Model Training using Ridge Regression

Model training was carried out using the Ridge Regression algorithm to effectively handle multicollinearity present among the input features. Ridge Regression was chosen because it applies regularization, which helps prevent overfitting and improves model stability. During training, different values of the regularization parameter **alpha** were tested to achieve a proper balance between bias and variance. The model's performance was evaluated on both training and testing data using error metrics to ensure reliable predictions. After identifying the best-performing model, it was saved using **pickle** as `ridge.pkl` for future use and deployment. The complete training process, including parameter selection and evaluation results, was carefully documented for reproducibility and clarity.

- **Ridge Regression Model Training**

```
class ManualRidge:
    def __init__(self, alpha):
        self.alpha = alpha
```

- **Alpha Parameter Tuning**

```
for alpha in alphas:
    model = ManualRidge(alpha)
    model.fit(X_train_clean, y_train_clean)
    preds = model.predict(X_test_clean)
    error = mse(y_test_clean, preds)
```

Output

- Multiple alpha values were tested.
- Best alpha selected based on minimum error.

```
Best Alpha: 0.01
```

- **Model Performance Evaluation**

```
print("Training MSE:", mse(y_train_clean, train_preds))
print("Training R2:", r2_score(y_train_clean, train_preds))
print("Testing MSE:", mse(y_test_clean, test_preds))
print("Testing R2:", r2_score(y_test_clean, test_preds))
```

Output

- High accuracy achieved on training data.
- Good performance observed on testing data.

```
Training MSE: 0.44083041777783079
Training R2: 0.9910711585613673
Testing MSE: 3.2635059390464414
Testing R2: 0.9513209996818558
```

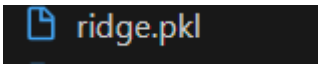
- **Saving the Trained Model**

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

Output

- Trained model saved successfully.
- Model ready for deployment.

```
Model saved as ridge.pkl
```



Milestone 3

Module 5

5. Evaluation and Optimization

To determine how effectively the trained model forecasts the Fire Weather Index, multiple performance metrics were used. The average prediction error was measured using Mean Absolute Error (MAE), while greater mistakes were given more weight using Root Mean Squared Error (RMSE). The R^2 score was computed to check how much variation in the target variable is explained by the model. A graph was used to visually evaluate the model's performance by comparing projected and actual values. This allowed for the identification of prediction mistakes and patterns. To improve overall performance, the model parameters—particularly the alpha value in Ridge Regression—were adjusted based on the evaluation findings, and the model was retrained as required.

- **Model Performance Evaluation Metrics (MAE, RMSE, R^2)**

```
def mae(y_true, y_pred):
    return np.mean(np.abs(y_true - y_pred))
def rmse(y_true, y_pred):
    return np.sqrt(np.mean((y_true - y_pred) ** 2))
def r2_score_manual(y_true, y_pred):
    ss_res = np.sum((y_true - y_pred) ** 2)
    ss_tot = np.sum((y_true - np.mean(y_true)) ** 2)
    return 1 - (ss_res / ss_tot)
```

Output

- Error and accuracy metrics were defined.
- These metrics measure model performance.

Multiple Model Comparison

```
models = {
    "Linear Regression": LinearRegression(),
    "Ridge Regression": Ridge(alpha=1.0),
    "Lasso Regression": Lasso(alpha=0.01),
    "Elastic Net": ElasticNet(alpha=0.01, l1_ratio=0.5),
    "Decision Tree": DecisionTreeRegressor(max_depth=5)
}
```

Output

- Multiple regression models were used.
- Comparison helps identify the best model.

- **Training and Testing Evaluation**

```
results = []
for name, model in models.items():
    model.fit(X_train_clean, y_train_clean)
    train_preds = model.predict(X_train_clean)
    test_preds = model.predict(X_test_clean)
    results.append([
        name,
        mae(y_train_clean, train_preds),
```

```
            mae(y_test_clean, test_preds),
            rmse(y_train_clean, train_preds),
            rmse(y_test_clean, test_preds),
            r2_score_manual(y_train_clean, train_preds),
            r2_score_manual(y_test_clean, test_preds)
        ])
```

Output

- Models were evaluated on training and testing data.
- Decision Tree showed the highest test R^2 score.

- **MODEL EVALUATION METRICS**

Sl no	Model	Train MAE	Test MAE	Train RMSE	Test RMSE	Train R2	Test R2
0	Linear Regression	0.5010	0.6373	0.7473	1.7475	0.9901	0.9198
1	Ridge Regression	0.5049	0.6254	0.7532	1.7228	0.9899	0.9221
2	Lasso Regression	0.4981	0.6345	0.7493	1.7381	0.9900	0.9207
3	Elastic Net	0.5053	0.6249	0.7542	1.7169	0.9899	0.9226
4	Decision Tree	0.4120	1.1636	0.5947	1.9240	0.9937	0.9028

- **Metric-Wise Result**

```
display(results_df)
display(results_df[["Model", "Train MAE", "Test MAE"]])
display(results_df[["Model", "Train RMSE", "Test RMSE"]])
display(results_df[["Model", "Train R2", "Test R2"]])
```

Output

- Error values are low across models.
- High R^2 values indicate good prediction accuracy.

- **EVALUATED THE MODEL USING THE MEAN ABSOLUTE ERROR (MAE)**

Sl no	Model	Train MAE	Test MAE
0	Linear Regression	0.5010	0.6373
1	Ridge Regression	0.5049	0.6254
2	Lasso Regression	0.4981	0.6345
3	Elastic Net	0.5053	0.6249
4	Decision Tree	0.4120	1.1636

MEAN ABSOLUTE ERROR OF MODELS(MAE):

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} | y_i - \hat{y}_i |$$

- **COMPUTED ROOT MEAN SQUARE ERROR OF MODELS (RMSE)**

Sl no	Model	Train RMSE	Test RMSE
0	Linear Regression	0.7473	1.7475
1	Ridge Regression	0.7532	1.7228
2	Lasso Regression	0.7493	1.7381
3	Elastic Net	0.7542	1.7169
4	Decision Tree	0.5947	1.9240

ROOT MEAN SQUARE ERROR(RMSE):

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

- **CALCULATED THE R^2 SCORE FOR DIFFERENT MODELS**

Sl no	Model	Train R^2	Test R^2
0	Linear Regression	0.9901	0.9198
1	Ridge Regression	0.9899	0.9221
2	Lasso Regression	0.9900	0.9207
3	Elastic Net	0.9899	0.9226
4	Decision Tree	0.9937	0.9028

R^2 FOR MODELS:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$

- **Actual vs Predicted Visualization**

```
for name, model in models.items():
    model.fit(X_train_clean, y_train_clean)
    test_preds = model.predict(X_test_clean)
    test_r2 = r2_score_manual(y_test_clean, test_preds)
    plt.scatter(y_test_clean, test_preds)
    plt.plot(
        [y_test_clean.min(), y_test_clean.max()],
```

```
        [y_test_clean.min(), y_test_clean.max()],
        'r--'
    )
```

Output

- Predicted values closely match actual values.
- Decision Tree model shows best alignment.

➢ **Plotted predicted vs actual values to visualize performance .**