FIRE WEATHER INDEX (FWI) PREDICTION MODEL

Milestone 2:

Feature Engineering and Scaling & Model Training using Ridge Regression

Submitted by:

Himanshu Ramole

Infosys Springboard Mentor:

Praveen

Date: 25-12-2025

1. Overview of Milestone 2

Milestone 2 focuses on transforming the cleaned dataset from Milestone 1 into a machine-learning-ready format and training a regression model to predict the Forest Weather Index (FWI). The primary objectives of this milestone were:

- To prepare meaningful input features through feature selection and scaling
- To split the dataset into training and testing subsets
- To train a Ridge Regression model while controlling multicollinearity
- To evaluate model performance using training and testing accuracy
- To save trained artifacts for future deployment

2. Module 3: Feature Engineering and Scaling

2.1 Feature Selection

From the cleaned dataset, all relevant numerical features contributing to fire weather behavior were selected as input variables. The **FWI** column was chosen as the target variable.

- **Input Features (X):**
  Temperature, RH, WS, Rain, FFMC, DMC, DC, ISI, BUI, Region
- **Target Variable (y):**
  FWI

- The selection was guided by correlation analysis performed in Milestone 1 to ensure meaningful relationships with the target variable.

2.2 Train-Test Split

The dataset was split into training and testing sets to evaluate generalization performance.

- Training data: 80%
- Testing data: 20%
- Random state was fixed to ensure reproducibility

This separation ensures that the model is evaluated on unseen data.

2.3 Feature Scaling

Since the input features had different units and ranges, **StandardScaler** was applied to normalize the data.

- Scaling was applied only on training data
- The same scaler was then applied to test data
- This avoids data leakage

The fitted scaler was saved as a .pkl file to ensure consistency during deployment.

Scaler saved at:

```
C:\Infosys Springboard 6.0 Internship\Models\scaler.pkl
```

This completes Module 3 successfully.

3. Module 4: Model Training using Ridge Regression

3.1 Model Selection

Ridge Regression was selected for this project because:

- The dataset contains correlated meteorological features
- Ridge Regression reduces multicollinearity using L2 regularization
- It improves model stability and generalization

3.2 Model Training

The Ridge Regression model was trained using the scaled training data.

- Alpha value was set initially to balance bias-variance tradeoff
- The model was trained on the training dataset only

3.3 Model Evaluation

Model performance was evaluated using **R^2 score** on both training and testing datasets.

- Training accuracy was calculated to check learning capability
- Testing accuracy was calculated to ensure generalization
- Displaying both values helps identify underfitting or overfitting

This satisfies the requirement that **train and test accuracy must be visible in output**.

3.4 Model Persistence

After successful training and evaluation, the trained Ridge Regression model was saved using pickle.

Model saved as:

`ridge.pkl`

Saving the model allows reuse without retraining and supports deployment readiness.

4. Key Outcomes of Milestone 2

- Successfully engineered features suitable for modeling
- Eliminated scale bias using standardization
- Maintained low multicollinearity using Ridge Regression
- Achieved visible training and testing accuracy metrics
- Stored both scaler and model artifacts for future use

5. Conclusion

Milestone 2 was completed successfully by implementing a structured machine learning pipeline. The dataset was transformed into a model-ready format, and a Ridge Regression model was trained and evaluated effectively. All required outputs, including saved scaler and trained model files, were generated as per project guidelines. This milestone prepares the foundation for advanced evaluation, optimization, and deployment in the next phase.