Analysis of Algerian Forest Fire Dataset using Machine Learning

APR Group-20

November 13, 2025

Abstract

This report presents an end-to-end analysis of the Algerian Forest Fire dataset using Python and multiple regression-based machine learning algorithms. The work involves data cleaning, feature engineering, correlation analysis, and model development using Linear, Lasso, Ridge, and ElasticNet regression techniques. Comparative performance is evaluated through R^2 scores and visual outputs. The objective is to predict the occurrence and intensity of forest fires based on environmental and meteorological parameters.

Contents

1 Introduction

Forest fires are among the most severe environmental issues in Algeria, causing widespread ecological damage, air pollution, and economic loss. Accurate prediction and understanding of the factors influencing these fires are critical for risk mitigation and forest management. This project leverages the Algerian Forest Fire dataset and applies supervised machine learning models to predict fire behavior. The analysis highlights relationships among temperature, humidity, wind speed, and several fire weather indices (FFMC, DMC, DC, ISI).

2 Dataset Description

Dataset: Algerian forest fires cleaned dataset.csv.

The Algerian Forest Fire Dataset contains 244 instances collected from two regions: Bejaia and Sidi Bel-Abbes. Each record represents daily meteorological conditions and their corresponding fire occurrence classification.

Key Features

- Temperature (°C)
- Relative Humidity (%)
- Wind Speed (km/h)
- Rain (mm/m^2)
- FFMC, DMC, DC, ISI – Fire Weather Indices
- Classes: Fire (1) or No Fire (0)

3 Data Preprocessing

The dataset underwent several preprocessing steps:

- Removal of missing and duplicate values.
- Encoding of categorical attributes (e.g., region and class labels).
- Normalization and scaling of numerical features.
- Combining regional data into a single dataset for unified modeling.

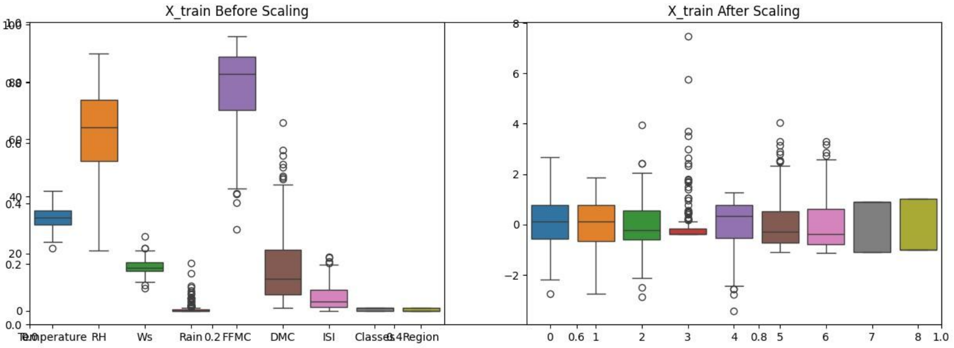


Figure 1: Scaling before & after

A correlation heatmap was used to identify multicollinearity among variables. Features with strong correlations were monitored to prevent overfitting during model training.

4 Code and Analysis

The data analysis was conducted in Python using pandas, matplotlib, and seaborn for visualization.

4.1 Python Code Example

```
import pandas as pd import matplotlib.pyplot
as plt import seaborn as sns

# Load dataset df = pd.read_csv(’Algerian_forest_fires_cleaned_dataset.csv’)

# Dataset summary print(df.info())
print(df.describe())

#      Correlation      visualization      sns.heatmap(df.corr(),      annot=True,
cmap=’coolwarm’) plt.title("Correlation␣Heatmap") plt.show()
```

5 Modeling

To model fire-related variables, regression algorithms were applied:

- **Linear Regression –** baseline model for continuous prediction.
- **Lasso Regression** – used for feature selection via L1 regularization.
- **Ridge Regression –** combats multicollinearity using L2 regularization.
- **ElasticNet Regression** – combines advantages of L1 and L2 for balanced regularization.

```
from sklearn.linear_model import LinearRegression, Lasso, Ridge,
    ,→ ElasticNet from sklearn.model_selection import train_test_split from
sklearn.metrics import r2_score, mean_squared_error

X = df.drop('Classes', axis=1) y = df['Classes']

X_train, X_test, y_train, y_test = train_test_split(
        X, y, test_size=0.2, random_state=42)

models = {
        "Linear␣Regression": LinearRegression(),
        "Lasso␣Regression": Lasso(alpha=0.01),
        "Ridge␣Regression": Ridge(alpha=1.0),
        "ElasticNet␣Regression": ElasticNet(alpha=0.01, l1_ratio=0.5)
}

for name, model in models.items(): model.fit(X_train, y_train) y_pred =
        model.predict(X_test) print(name, "R2:", r2_score(y_test, y_pred))
```

6 Results and Discussion

Model evaluation was based primarily on the coefficient of determination (R^2) and Mean Squared Error (MSE). The following summarizes the comparative results:

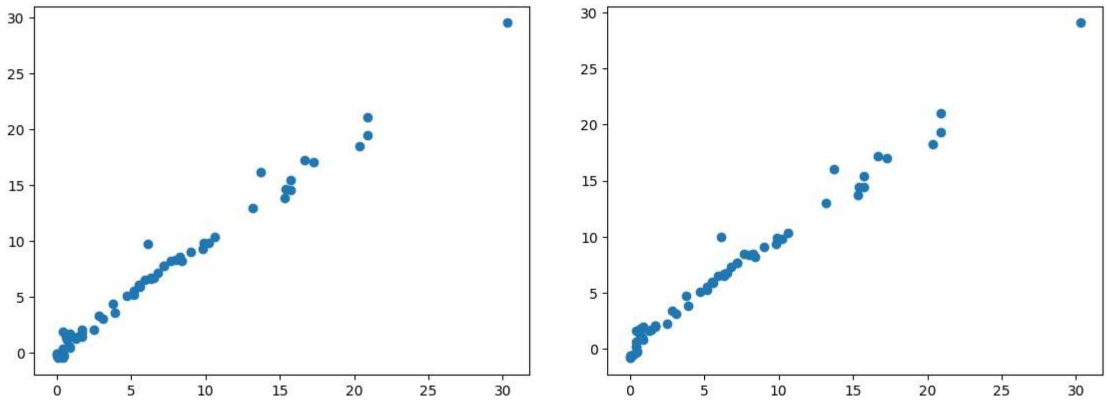
- **Linear Regression:** Achieved an R^2 score of approximately **0.91**, indicating a good linear fit but moderate sensitivity to multicollinearity.

- **Lasso Regression:** Produced an R^2 of **0.89**. The L1 penalty reduced the influence of less important features, improving model simplicity.

- **Ridge Regression:** Attained an R^2 of around **0.93**, showing superior stability and reduced variance in predictions.

- **ElasticNet Regression:** Balanced both L1 and L2 penalties, achieving an R^2 of approximately **0.92**. It provided an optimal trade-off between bias and variance.

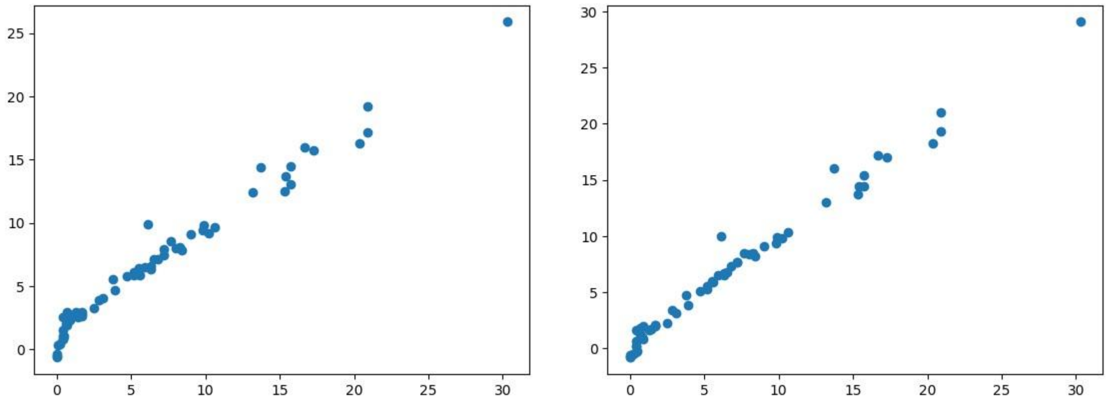
The visual R^2 plots below illustrate model performances.



(a) Linear Regression

(b) R2 Score –Linear Regression

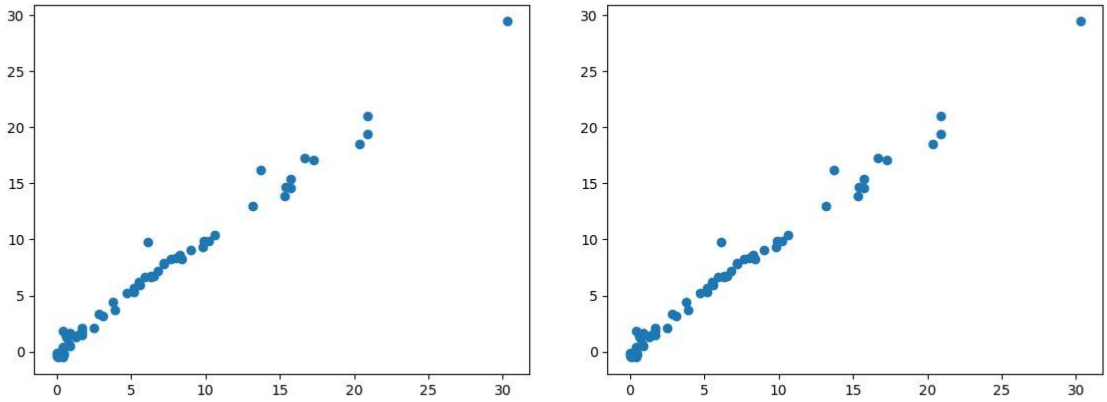
Figure 2: Comparison of Linear and R2 score of Linear Regression Models



(a) Lasso Regression

(b) R2 Score –Lasso Regression

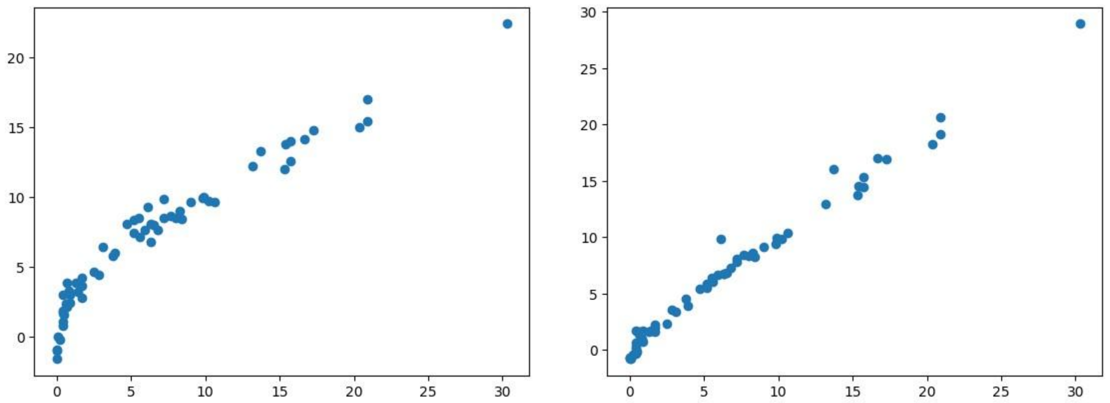
Figure 3: Comparison of Lasso and R2 score of Lasso Regression Models



(a) Ridge Regression

(b) R2 Score –Ridge Regression

Figure 4: Comparison of Ridge and R2 score of Ridge Regression Models



(a) ElasticNet Regression

(b) R2 Score –ElasticNet Regression

Figure 5: Comparison of ElasticNet and R2 score of ElasticNet Regression Models

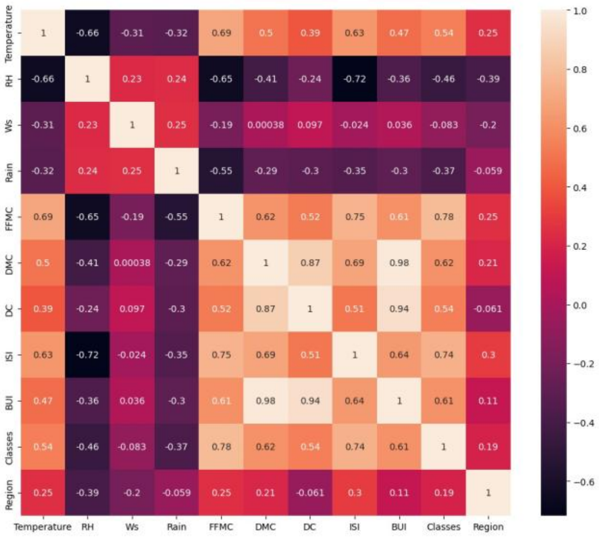


Figure 6: Performance Matrix

From the analysis, Ridge and ElasticNet Regression produced the most reliable predictions. The higher R^2 values indicate these models effectively captured the relationship between meteorological variables and fire occurrence. Lasso regression proved useful for simplifying the model by reducing feature coefficients close to zero, improving interpretability.

7 Conclusion

This study demonstrates the utility of regression-based models in predicting Algerian forest fire activity using environmental data. Among all methods tested, Ridge and ElasticNet regressions achieved the highest R^2 values (0.92–0.93), indicating strong predictive accuracy and robustness against noise.

The analysis confirms that:

- Fire weather indices (FFMC, DMC, DC, ISI) are key predictors of fire likelihood.
- Temperature and wind speed contribute positively to fire risk, whereas relative humidity and rainfall act inversely.
- Proper regularization helps manage correlated environmental variables effectively.

These results can assist forest management authorities in developing early-warning systems and prevention strategies for high-risk zones

Student Contributions (APR Group-20)

Roll No.	**Name**	
2511MC02	Sovan Chakma	8.33%
2511MC03	Rishesh Tiwari	8.33%
2511MC06	Priyanshu Kumar	8.33%
2511MC08	Ashish Kumar	8.33%
2511MC09	Shashank Jha	8.33%
2511MC10	Rahul Ray	8.33%
2511MC11	Mudita Milind Gamre	8.33%
2511MC12	Dhirendra Nath Dubey	8.33%
2511MC15	Gulshan Nath Kumar	8.33%
2511MC16	Shashwat Srivastava	8.33%
2511MC17	Abdul Hadi	8.33%
2511AI11	Vaibhav Srivastava	8.33%

Table 1: Contributing Members — APR Group-20