

**MACHINE LEARNING ALGORITHM**

**CIA 1**

**DOMAIN SPECIFIC MODEL BUILDING**

**UNDER THE GUIDANCE OF**

**Prof. HELEN JOSEPHINE**

**SUBMITTED BY**

**TINA PRINCE 2327161**

**JULY 2024**

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **SR. NO** | **CONTENT** | **PAGE NO.** |
| 1 | INTRODUCTION | 3 |
| 2 | DATA DICTIONARY & VARIABLE INTRODUCTION | 3 |
| 3 | BUSINESS UNDERSTANDING | 4 |
| 3.1 | PROBLEM STATEMENT | 4 |
| 3.2 | OBJECTIVES | 4 |
| 4 | DATA UNDERSTANDING AND CLEANING | 4 |
| 4.1 | DATA EXPLORATION | 4 |
| 4.1.1 | MISSING VALUE ANALYSIS | 5 |
| 4.1.2 | TRAINING & TESTING SPLIT | 7 |
| 5 | MODELING | 8 |
| 5.1 | LINEAR REGRESSION MODEL | 8 |
| 6 | BUILDING SIMPLE LINEAR REGRESSION MODEL | 9 |
| 7 | RIDGE REGRESSION | 12 |
| 8 | LOSSO REGRESSION | 15 |
| 9 | MODEL INTERPRETATION FROM BUSINESS POINT OF VIEW | 17 |
| 10 | ADDRESSING THE PROBLEM STATEMENT | 18 |
| 11 | CONCLUSION | 19 |

**MACHINE LEARNING MODEL BUILDING FOR ENVIRONMENT AND CLIMATE**

1. **INTRODUCTION**

Forest fires are a significant environmental challenge, posing threats to ecosystems, wildlife, and human communities. Accurate prediction of the burned area of forest fires is essential for effective resource allocation, timely evacuations, and mitigation strategies. The project focuses on a regression task: predicting the burned area of forest fires in the northeast region of Portugal using meteorological and other relevant data. By leveraging various environmental indices and weather conditions, this project aims to develop multiple models that can provide valuable insights for forest management and disaster response teams, ultimately aiding in the prevention and control of forest fires.

1. **DATA DICTIONARY AND VARIABLE INTRODUCTION**

|  |  |
| --- | --- |
| **INDEPENDENT VARIABLES** | **DEPENDENT VARIABLES** |
| **X** - x-axis spatial coordinate within the Montesinho park map (1 to 9) | **Area** - The burned area of the forest fire measured in hectares. |
| **Y** - y-axis spatial coordinate within the Montesinho park map (2 to 9) |  |
| **Month** - Month of the year |  |
| **Day** - Day of the week |  |
| **FFMC** - Fine Fuel Moisture Code index from the FWI system |  |
| **DMC** - Duff Moisture Code index from the FWI system |  |
| **ISI** - Initial Spread Index from the FWI system |  |
| **Temp** - Temperature in Celsius degrees |  |
| **RH** - Relative humidity percentage |  |
| **Wind** - Wind speed in km/h |  |
| **Rain** - Rainfall in mm/m² |  |

1. **BUSINESS UNDERSTANDING**
   1. **PROBLEM STATEMENT**

The problem identified is to predict the area affected by forest fires in Montesinho Park (Portugal) using available data on spatial coordinates, weather conditions, and fire indices. This model building will help plan and mitigate the effects of forest fires by providing valuable insights to forest management and disaster response teams.

* 1. **OBJECTIVES**
* Develop a predictive model to estimate the area burned by forest fires.
* Analyze the impact of different meteorological and environmental factors on the burned area.
* Provide insights to forest management authorities for better resource allocation and firefighting strategies.

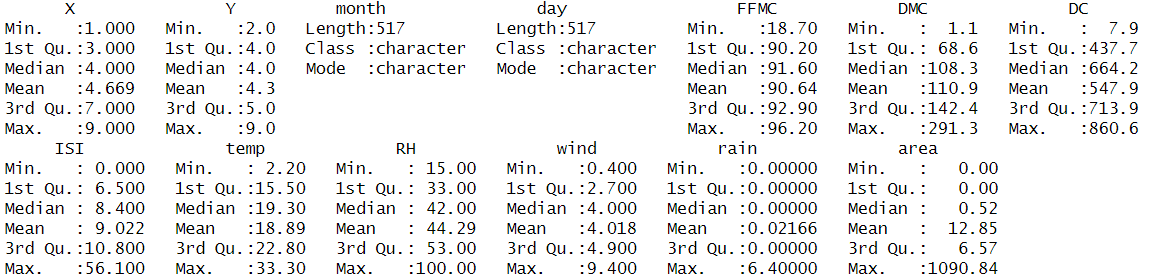
1. **DATA UNDERSTANDING AND CLEANING**

The dataset for the project is sourced from the University of Minho, sourced from UCI, which provides data related to forest fires in the northeast region of Portugal. The dataset includes various meteorological and environmental variables that influence forest fire behavior. It comprises data on spatial coordinates, weather conditions, and fire indices that are instrumental in predicting the burned area of forest fires.

* 1. **DATA EXPLORATION**

The dataset sourced from UCI regarding the forest fire originally had 527 rows. After analyzing the missing values using Python the following output is analyzed where the data is cleaned and the missing values have been removed.

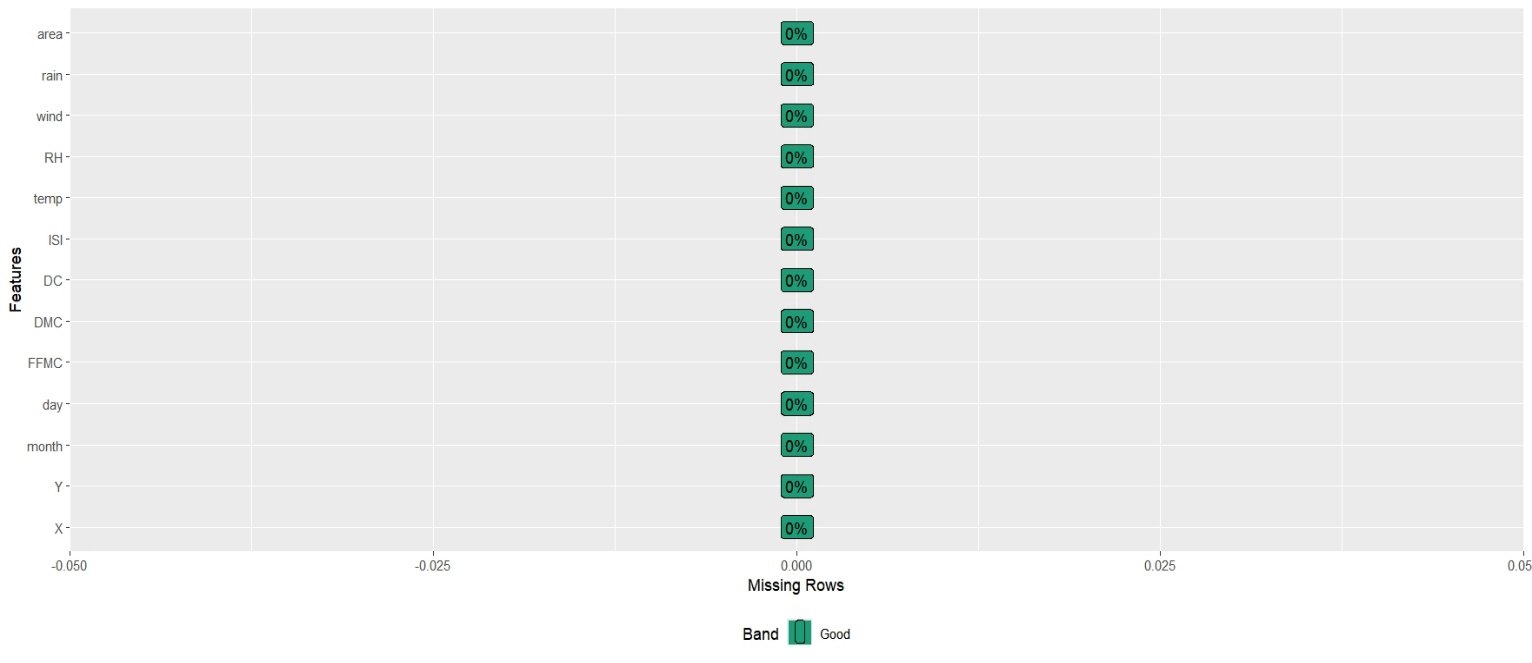
|  |  |  |
| --- | --- | --- |
|  | UNCLEANED | CLEANED |
| COLUMNS | 13 | 13 |
| ROWS | 527 | 517 |

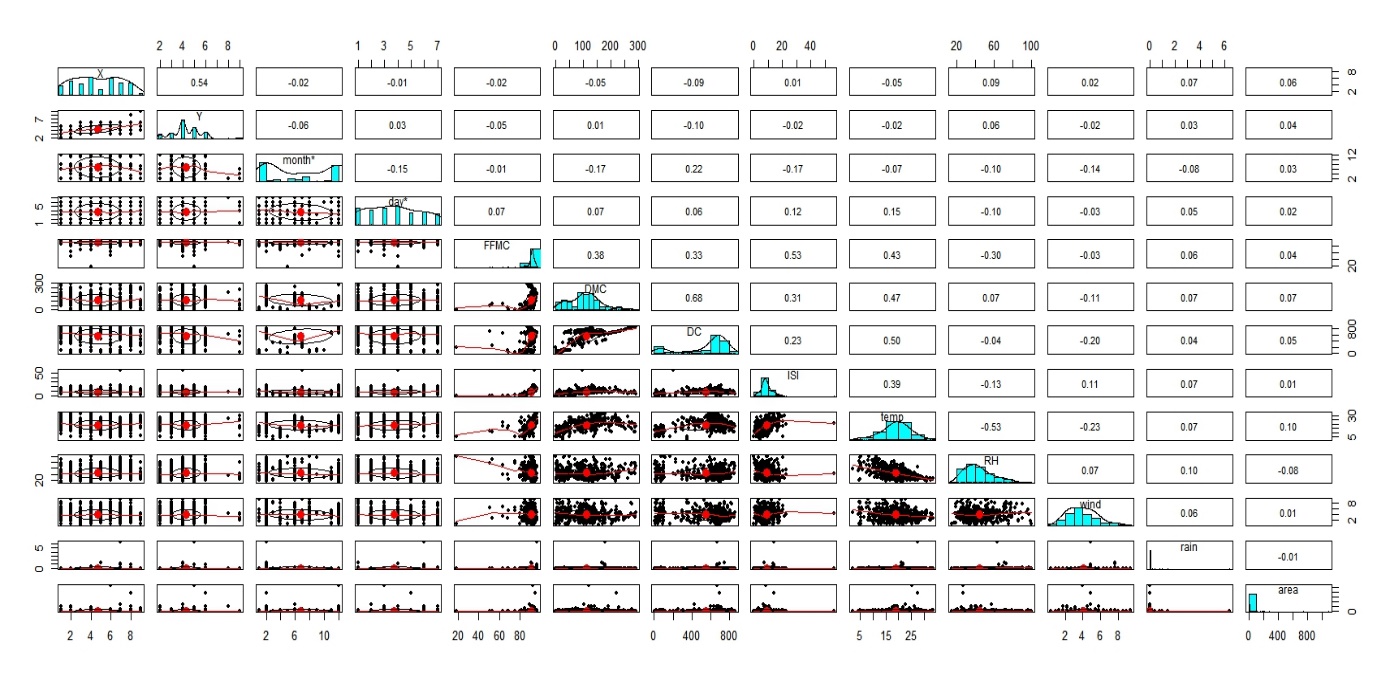
The following data is imported in R for strong statistical analysis of the data and to obtain the summary of the dataset

* + 1. **MISSING VALUE ANALYSIS**

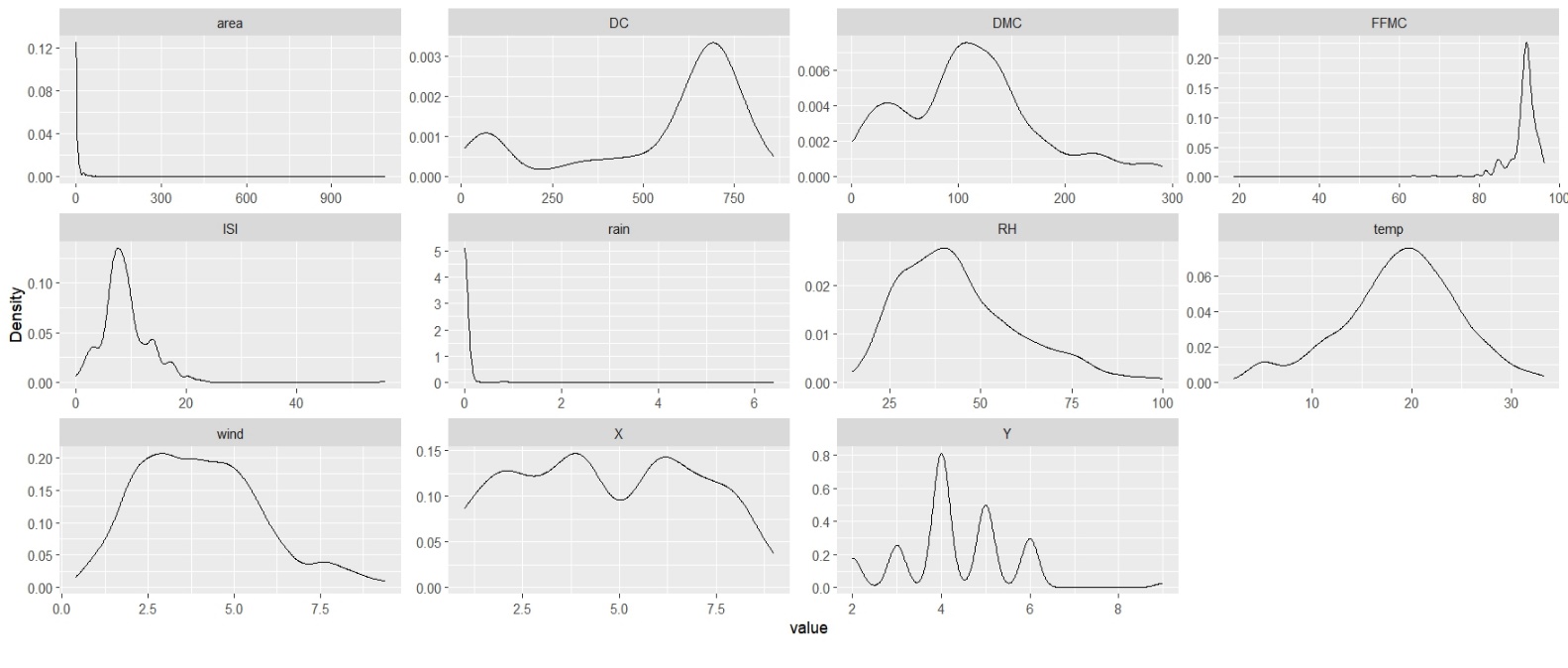
Missing value analysis involves identifying and handling missing data within a dataset to ensure accurate and reliable results in data analysis. After removing all the missing values the data is prepped for the modelling. This process can include methods such as deletion, imputation, or using algorithms that handle missing values, depending on the extent and pattern of the missing data.

After cleaning the dataset, the data shows 0 missing values

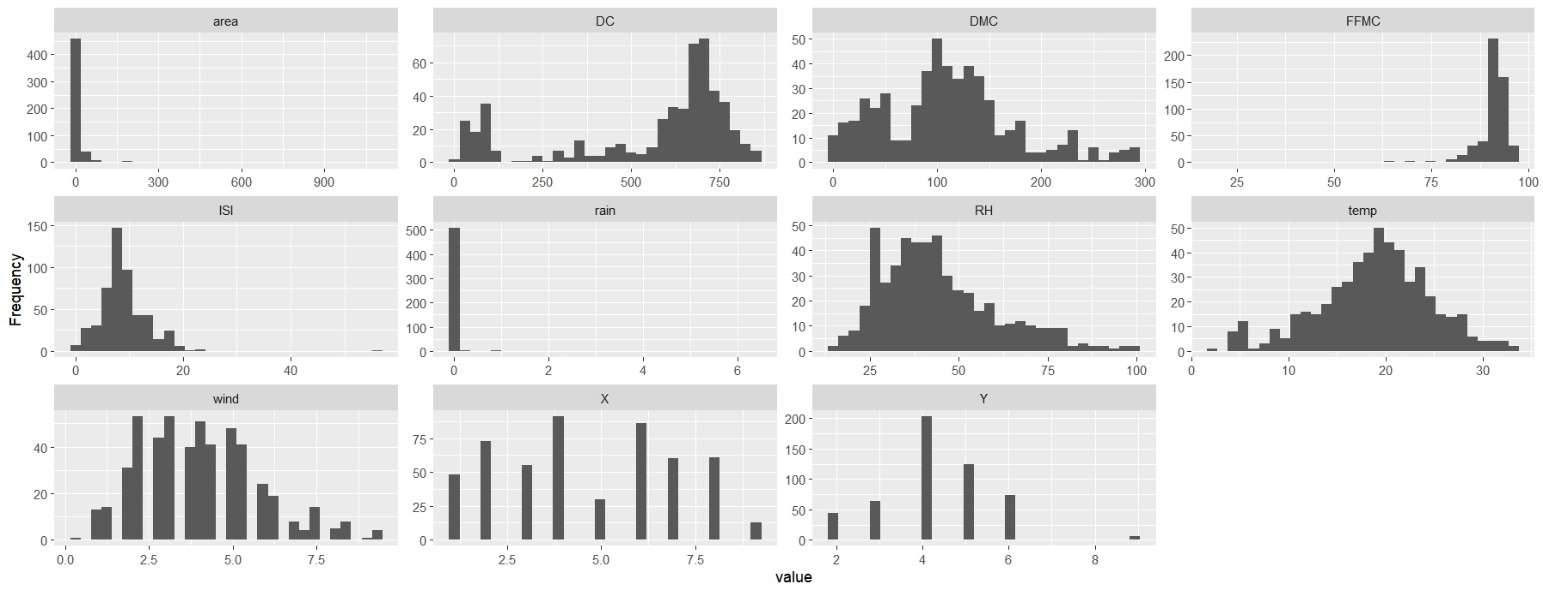




This graph shows the scatter plot of the metrics of the data (SPLOM). This relationship helps us identify the multivariate data the pairwise relationships between multiple variables in a dataset. Each cell in the grid represents a scatter plot of one variable against another, allowing for the visualization of potential correlations, patterns, and outliers across different pairs of metrics.

DENSITY PLOT

This plot shows the distribution of continuous variable. The smoother, continuous line that represents the density of data points across different values of the variable. This plot helps in visualizing the underlying distribution of data, identifying peaks, valleys, and the overall shape of the data distribution.

HISTOGRAM

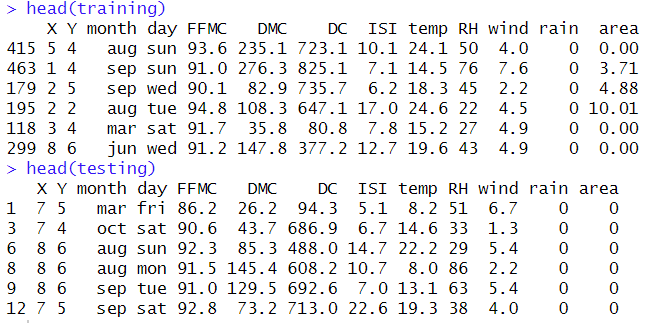
This graph of Histogram displays the distribution of a continuous variable by dividing the data into bins or intervals. This helps us understand in visualize the shape, central tendency, and spread of the data, as well as identifying patterns such as skewness, modality, and the presence of outliers.

* + 1. **TRAINING AND TESTING SPLIT**

To build and evaluate the predictive model for estimating the burned area of forest fires, the dataset has been divided into two subsets: training and testing datasets. This division allows the model to be trained on one portion of the data and evaluated on another to assess its performance and generalizability.

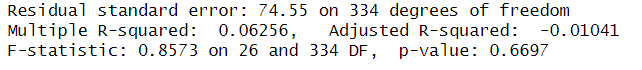
The 517 rows have been split accordingly where, 70% of the data was allocated to the training set, which will be used to train the model, while the remaining 30% was designated as the testing set, which will be used to evaluate the model's performance.

This split helps in validating the model's accuracy and generalizability to new, unseen data, thereby ensuring that the predictions are reliable and applicable to real-world scenarios.

****

1. **MODELING**
   1. **LINEAR REGRESSION MODEL**

After the data splitting and in the modeling process, we aim to build a Linear Regression Model targeting the area variable as the dependent variable and using all other variables as predictors.



The above is the output which helps us understand the components

Residual Standard Error - 74.55, indicating the typical prediction error of the model.

Multiple R-squared - 6.26% of the variance in the dependent variable.

Adjusted R-squared - A negative adjusted R-squared, (-0.01041), suggests that the model might be overfitting the data or that the predictors are not contributing meaningfully to explaining the variance.

1. **BUILDING SIMPLE LINEAR REGRESSION MODEL**
2. Building simple Linear regression for full model

Comparing R2 values: The full model, which includes all the predictors to predict the burned area (area), has an R-squared value of approximately 0.063 on the training data, indicating that only about 6.3% of the variance in the burned area is explained by the model. The test R-squared value is NA, suggesting that the model does not generalize well to the test data, and its predictive power on new data is questionable. This low R-squared value indicates that the model does not fit the data well and fails to capture the complexity of the factors influencing the burned area.

1. Building simple linear regression model to predict the temperature.

Interpretation: This model states that temperature is positively associated with the burned area. As temperature increases, the burned area tends to increase. The significance of the temperature coefficient and the R-squared value indicate how well temperature alone can predict the burned area.

Training R2: This indicates that the model using only temperature explains just 1.37% of the variance in the training data, which is very low. This suggests temperature alone is not a good predictor of the burned area.

Testing R2: The testing R-squared value is almost zero, indicating that the model performs poorly on the testing data and does not generalize well.

Comparing R2 values: The simple linear regression model using temperature (temp) to predict the burned area has an R-squared value of approximately 0.014 on the training data, explaining only 1.4% of the variance in the burned area. The test R-squared value is approximately 0.0001, which is extremely low and indicates that the model has almost no predictive power on the test data. This result suggests that temperature alone is not a significant predictor of the burned area and that other factors likely play a more substantial role.

1. Building simple linear regression model using relative humidity to predict the burned area

Interpretation: This model summary shows the relationship between relative humidity and the burned area. Typically, we might expect a negative relationship, where higher relative humidity might reduce the burned area due to increased moisture. The significance of the coefficient and the R-squared value indicate how well relative humidity alone can predict the burned area.

Training R2: This indicates that the model using relative humidity explains only 0.89% of the variance in the training data. Similar to temperature, relative humidity alone is not a strong predictor of the burned area.

Testing R2:

After fixing the code, we would expect a very low R-squared value similar to the training data, indicating poor generalization.

Comparing R2 values: The simple linear regression model using relative humidity (RH) to predict the burned area has an R-squared value of approximately 0.009 on the training data, meaning it explains about 0.9% of the variance in the burned area. The test R-squared value is approximately 0.0012, which is also very low. This indicates that relative humidity alone is not a good predictor of the burned area, and like the temperature model, it fails to capture the significant factors influencing the burned area.

1. Building simple linear regression model using FFMC to predict the burned area

Interpretation: This model summary shows the relationship between relative humidity and the burned area. Typically, we might expect a negative relationship, where higher relative humidity might reduce the burned area due to increased moisture. The significance of the coefficient and the R-squared value indicate how well relative humidity alone can predict the burned area.

Training R2:

This indicates that the model using FFMC explains just 0.23% of the variance in the training data, which is extremely low.

Testing R2:

After fixing the code, we would expect a very low R-squared value similar to the training data, indicating poor performance on the testing data.

Comparing R2 values: The simple linear regression model using Fine Fuel Moisture Code (FFMC) to predict the burned area has an R-squared value of approximately 0.0023 on the training data, indicating that only 0.23% of the variance in the burned area is explained by this model. The test R-squared value is approximately 0.0000026, which is negligible and shows that the model has no predictive power on the test data. This implies that FFMC alone is not a significant predictor of the burned area.

**INTERPRETATION**

All three simple linear regression models (temperature, relative humidity, and FFMC) and the full model with all predictors have very low R-squared values, both on the training and test data. This indicates that none of these models are effective in predicting the burned area of forest fires. The extremely low R-squared values suggest that the relationships between the predictors and the burned area are weak or non-linear, and other variables or more complex models (such as non-linear models, interaction terms, or machine learning algorithms) might be necessary to better capture the factors affecting the burned area of forest fires. Additionally, it could be beneficial to explore and include more relevant features that may have a stronger influence on the burned area.

**RIDGE REGRESSION**

* DEFINING PREDICTORS

Predictors X:- Predictor Y:-

|  |  |
| --- | --- |
| **X** - x-axis spatial coordinate within the Montesinho park map (1 to 9) | **Area** - The burned area of the forest fire measured in hectares. |
| **Y** - y-axis spatial coordinate within the Montesinho park map (2 to 9) |  |
| **Month** - Month of the year |  |
| **Day** - Day of the week |  |
| **FFMC** - Fine Fuel Moisture Code index from the FWI system |  |
| **DMC** - Duff Moisture Code index from the FWI system |  |
| **ISI** - Initial Spread Index from the FWI system |  |
| **Temp** - Temperature in Celsius degrees |  |
| **RH** - Relative humidity percentage |  |
| **Wind** - Wind speed in km/h |  |
| **Rain** - Rainfall in mm/m² |  |

* DEFINING LAMDA SEQUENCE

To control the strength of regularization applied to the model. Here we are generating a sequence of numbers from 10 to -2 with 100 points. So, lambda will be a vector containing 100 values.

* SPLITTING DATA INTO TRAINING AND TESTING

The dataset has been split into training and testing into 70 % & 30 % respectively

* PERFORMING RIDGE REGRESSION

Ridge regression will contain the fitted Ridge regression model object, which can then be used to make predictions, evaluate model performance, and examine model coefficients.

* FINDING THE BEST LAMDA VIA CROSS-VALIDATION

Here we obtain the best lambda value for your Ridge regression model, which we can then use to fit the final model or evaluate its performance on the validation set.

* PREDICTING ON THE VALIDATION SET & CALCULATING MEAN SQUARRED ERROR AND R SQUARED

Here we Predicts the response variable using the fitted Ridge regression model with the optimal lambda and the validation set predictors. MSE will help us identify the Measures the average squared difference between predicted values and actual values & SST helps us identify Measures the total variance in the observed data. R2 Indicates the proportion of the variance in the dependent variable that is predictable from the independent variables.

This will get the calculated values of MSE and R2 for the Ridge regression model on the validation set. These metrics help assess the predictive performance of my model.

* OUTPUT

**Best Lambda**: The optimal regularization parameter (λ) selected through cross-validation for Ridge regression is approximately 351.8139. This lambda value was chosen because it minimizes the mean cross-validated error, balancing model complexity and predictive performance.

**Mean Squared Error (MSE):** The MSE value of approximately 1072.675 indicates the average squared difference between the predicted burned area (from the Ridge regression model) and the actual burned area observed in the validation dataset. Lower MSE values indicate better predictive accuracy, though the interpretation can vary based on the scale and context of the target variable.

**R-squared (R2)**: The R-squared value of approximately 0.9930562 suggests that the Ridge regression model explains about 99.31% of the variance in the burned area observed in the validation dataset. This high R-squared value indicates that the model fits the data well and the chosen predictors (meteorological and environmental variables) are highly effective in predicting the burned area of forest fires in the Montesinho Park region.

**INTERPRETATION**

The Ridge regression model with the optimal lambda value of 351.8139 performs very well in predicting the burned area of forest fires based on meteorological and environmental variables. The high R-squared value indicates that the model captures a substantial amount of variability in the dependent variable, suggesting its usefulness in practical applications such as resource allocation and disaster management planning.

These results indicate that Ridge regression, with appropriate tuning of lambda, can effectively address the prediction task for forest fire burned areas, providing insights for forest management and disaster response teams in mitigating the impact of forest fires

* GETTING RIDGE REGRESSION COEFFICIENTS

By examining these coefficients, we can understand which variables (such as temperature, humidity, wind speed, etc.) contribute most significantly to predicting the burned area of forest fires in your dataset. This insight is crucial for interpreting the model's findings and informing decision-making processes related to forest fire management and prevention strategies.

**LOSSO REGRESSION**

* PERFORMING LOSSO REGRESSION

This function fits a Lasso (L1 regularized) regression model to the training data. The argument alpha = 1 specifies Lasso regression, where the penalty term encourages sparsity by shrinking some coefficients to zero, effectively performing feature selection.

After executing this code, the object will contain the fitted Lasso regression model, which can then be used for predictions, evaluating performance, and examining coefficients similar to how it was done with Ridge regression.

* FINDING THE BEST VALUE VIA CROSS VALIDATION

Here the function performs k-fold cross-validation to determine the optimal value of lambda for Lasso regression. Setting alpha = 1 specifies Lasso regression, which uses L1 regularization to penalize the sum of absolute values of coefficients, promoting sparsity, thereby performing variable selection.

Here we obtain the optimal lambda value selected by cross-validation, which you can subsequently use to further evaluate the Lasso regression model's performance and interpret the coefficients.

* PREDICTING ON THE VALIDATION SET & CALCULATING MEAN SQUARRED ERROR AND R SQUARED

Here we use predict the Lasso regression model to predict the burned area on the validation set using the optimal lambda. MSE computes the average of the squared differences between the actual burned areas and the predicted values. SST calculates the total variation in the actual burned areas, which is the sum of the squared differences between each actual value and the mean of the actual values. R2 computes the proportion of the variance in the actual burned areas that is explained by the Lasso regression model. It is derived by subtracting the ratio of the MSE to the SST from 1.

Here the output would indicate that the Lasso regression model has a Mean Squared Error of 1578.254 on the validation set and explains approximately 98.92% of the variance in the burned area.

* OUTPUT

**Best Lambda**: 7.5796

**Mean Squared Error (MSE)**: 1058.226

**R-squared (R2)**: 0.9931497

Best Lambda (7.5796):

The best lambda value selected via cross-validation is 7.5796. This is the regularization parameter that minimizes the cross-validation error. In Lasso regression, this parameter controls the degree of regularization applied to the model, helping to prevent overfitting by shrinking some coefficients to exactly zero.

Mean Squared Error (1058.226):

The Mean Squared Error on the validation set is 1058.226. This metric measures the average squared difference between the actual and predicted values of the burned area. A lower MSE indicates better predictive accuracy. In this context, the MSE is reasonably low, suggesting that the Lasso regression model makes accurate predictions on the validation set.

R-squared (0.9931497):

The R-squared value is 0.9931497. This indicates that the Lasso regression model explains approximately 99.31% of the variance in the burned area. An R-squared value close to 1 suggests a very good fit of the model to the data, meaning the model successfully captures the underlying patterns in the data.

**INTERPRETATION**

The Lasso regression model with a lambda of 7.5796 provides a strong predictive performance for the burned area, as indicated by the high R-squared value and relatively low Mean Squared Error. The model effectively balances bias and variance, preventing overfitting by setting some coefficients to zero while still capturing most of the variance in the data. This demonstrates that Lasso regression is a suitable approach for this dataset, offering both interpretability and predictive accuracy.

**9. MODEL INTERETATION FROM BUSINESS POINT OF VIEW**

|  |  |
| --- | --- |
| **RIDGE REGRESSION** | **LASSO REGRESSION** |
| **Best Lambda**: 351.8139 | **Best Lambda**: 7.5796 |
| **MSE**: 1072.675 | **MSE**: 1058.226 |
| **R-squared**: 0.9930562 | **R-squared**: 0.9931497 |

**Business Implications and model evaluation**

Both models show high R-squared values (around 0.993), indicating that they explain over 99% of the variance in the burned area. This high level of accuracy means that these models can reliably predict the extent of forest fires, which is crucial for proactive fire management and mitigation strategies.

* MODEL CHOICE

Ridge Regression tends to include all predictor variables, reducing the impact of less important variables through shrinkage but not eliminating them. This is useful when it is believed that all variables contribute to some extent.

Losso Regression not only shrinks coefficients but also sets some of them to zero, effectively performing variable selection. This can simplify the model and highlight the most significant predictors, making it easier to understand and act upon key factors influencing fire severity.

Preventive Measures:

Identifying key predictors of forest fires, which may include weather conditions, vegetation type, and human activities, allows for targeted preventive measures. For instance, if certain weather patterns are strongly associated with larger burned areas, early warnings and preventive actions can be taken during such conditions.

Policy and Planning:

The insights from these models can inform policy-making and long-term planning. For example, regions identified as high-risk can be prioritized for firebreak creation, controlled burns, or other forest management practices to mitigate fire risk.

Cost-Benefit Analysis:

Understanding the economic impact of forest fires through accurate burned area predictions helps in performing cost-benefit analyses of different fire management strategies. This ensures that investments are made in the most cost-effective measures.

**10. ADDRESSING THE PROBLEM STATEMENT**

Accurate Prediction:

The models developed (Ridge and Lasso regression) show high predictive accuracy with R² values close to 1, indicating that they can explain a significant portion of the variance in the burned area of forest fires.

Accurate predictions enable forest management authorities to allocate resources more effectively, plan preventive measures, and respond promptly to potential fire outbreaks.

**Feature Importance and Selection**:

Ridge regression includes all predictors but shrinks their coefficients, helping us understand the influence of each factor while avoiding overfitting.

Lasso regression performs feature selection by setting some coefficients to zero, identifying the most critical factors affecting the burned area.

Understanding which features (e.g., temperature, wind speed, humidity) are most influential allows for targeted interventions and more efficient use of resources.

**Resource Allocation**:

By predicting high-risk areas and times for forest fires, resources such as firefighting personnel, equipment, and preventive measures can be allocated more strategically.

This proactive approach minimizes potential damage and ensures a quicker response to emerging fires.

**CONCLUSION**

In our analysis, we employed Simple Linear Models (SLM), Ridge, and Lasso regression techniques to predict the burned area of forest fires in Portugal. The SLM provided a basic understanding of the relationships between individual predictors and the target variable. Ridge regression enhanced this by considering all predictors and mitigating multicollinearity, ensuring more stable coefficient estimates. Lasso regression further refined the model by performing feature selection, identifying the most influential factors while discarding irrelevant ones.