Comparing the Performance of Different Models on The Algerian Forest Fire Data Set

Elif Gizem ZEDEF
Computer Engineer
Gazi University
Ankara,Turkey

***Abstract*—The performance of five machine learning models was compared on the Algerian forest fire data set in this study. The models were K-Nearest Neighbors (KNN), Decision Tree, Naive Bayes, Linear Regression, and Logistic Regression. The data set, which was obtained from [1], contained 243 rows and 14 columns, with 43.6% examples of fire and 56.4% examples of non-fire classes. Data preprocessing was performed prior to model evaluation using accuracy, recall, precision, and F1 Score as performance metrics.**

The Decision Tree model achieved the highest overall performance, with an accuracy of 0.979, recall of 1.0, precision of 0.964, and F1 Score of 0.982. The KNN model had the second highest overall performance, with an accuracy of 0.837, recall of 0.852, precision of 0.852, and F1 Score of 0.852. The Naive Bayes model had an accuracy of 0.980, precision of 0.960, recall of 1.00, and F1 Score of 0.980. The Linear Regression model had a mean squared error of 0.020, mean absolute error of 0.020, and R-squared score of 0.918. The Logistic Regression model had a mean squared error of 0.087, mean absolute error of 0.254, and R-squared score of 0.648.

The results suggest that Decision Trees and KNN may be effective approaches for predicting forest fires in Algeria.

***Index Terms*—Classification, Forest Fire, Regression**

I. Introduction

Forest fires are a significant environmental and societal problem, causing destruction of ecosystems, loss of wildlife, and damage to human communities. Accurate prediction of forest fires can aid in mitigating their negative impacts by enabling proactive response efforts. In this study, I present a comparison of five machine learning models on the Algerian forest fire dataset to identify the most effective approach for predicting forest fires in Algeria. The models evaluated include K-Nearest Neighbors (KNN), Decision Tree, Naive Bayes, Linear Regression, and Logistic Regression. The performance of the models is assessed using accuracy, recall, precision, and F1 Score as evaluation metrics. The results of this comparison can provide insight into the selection of an appropriate model for predicting forest fires in Algeria, which can aid in improving response efforts and reducing the detrimental effects of these fires.

II. Literature Review

Several studies have used the Algerian forest fire data set to examine the factors that influence the occurrence and behavior of forest fires in Algeria. These studies have employed a variety of methods, including statistical analysis, machine learning, and spatial modeling.

One study used multiple linear regression to analyze the relationships between environmental and socio-economic variables, and the likelihood of forest fires in Algeria. The results of this study showed that temperature, relative humidity, and wind speed were the most important predictors of fire occurrence and that the risk of fires was higher in areas with low humidity and high human population density[4].

Another study used a decision tree model to identify the risk factors that contribute to forest fires in Algeria. The results of this study indicated that temperature, relative humidity, and wind speed were the most important predictors of fire occurrence, and that the risk of fires was higher in the dry, hot summer months[2].

A third study employed a geographic information system (GIS) to examine the spatial patterns of forest fires in Algeria, and to identify the areas that are most vulnerable to fires. The results of this study showed that fires were more likely to occur in the northern and central regions of the country and that the cedar forests were the most vulnerable to fire[3].

Overall, these studies have provided valuable insights into the factors that influence the occurrence and behavior of forest fires in Algeria. However, the sample size of the Algerian forest fire data set is relatively small, and the results of these studies may not be representative of the broader patterns of forest fires in the country. Additional research is needed to confirm and expand upon these findings, and to better understand the complex interactions between environmental, meteorological, and human factors that contribute to forest fires in Algeria.

III. Methodology

A. Data

The Algerian forest fire data set used in this study was obtained from [1]. The data set consists of observations of weather conditions and fire weather indices for Algeria's Bejaia and Sidi-Bel Abbes regions, covering the period from June to September 2012. The original data set contained two separate tables with the same attributes, one for each region, but they were combined into a single data frame with a new 'Region' attribute added to indicate the location of each observation. The resulting data frame has 243 rows and 14 columns, with a class balance of 56.4% non-fire and 43.6% fire instances.

The data frame contains the following attributes:

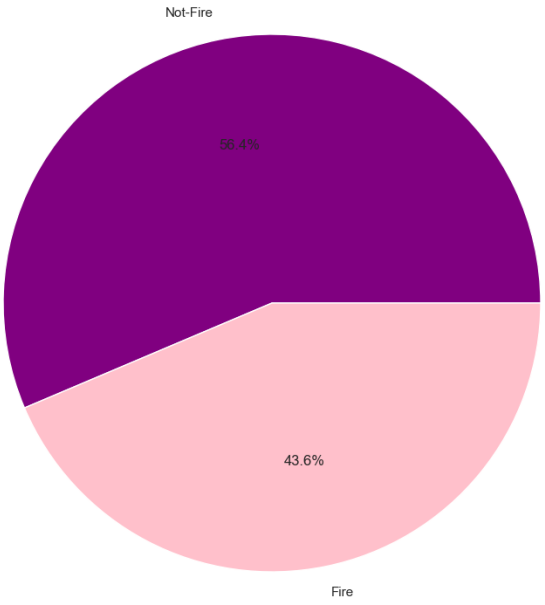

Fig. 1. Pie Chart of Classes

- Day: day of the month (1-30)
- Month: month of the year ('June' to 'September')
- Year: year of the observations (2012)
- Temperature: temperature at noon (temperature max) in Celsius degrees (22 to 42)
- RH: relative humidity in % (21 to 90)
- Ws: wind speed in km/h (6 to 29)
- Rain: total rainfall in mm (0 to 16.8)
- FFMC: Fine Fuel Moisture Code index from the FWI system (28.6 to 92.5)
- DMC: Duff Moisture Code index from the FWI system (1.1 to 65.9)
- DC: Drought Code index from the FWI system (7 to 220.4)
- ISI: Initial Spread Index index from the FWI system (0 to 18.5)
- BUI: Buildup Index index from the FWI system (1.1 to 68)
- FWI: Fire Weather Index (0 to 31.1)
- Classes: two classes, namely 'fire' and 'not fire'
- Region: two classes, namely 'Bejaia Region Dataset' and 'Sidi-Bel Abbes Region Dataset'

Fig. 2 shows the correlations between different attributes in the Algerian forest fire dataset. For example, I found that the attribute "*Temperature*" had a strong positive correlation of 0.676568 with the attribute "*FFMC*", indicating that as the temperature increases, the FFMC value also tends to increase. I also found that the attribute "*Temperature*" had a strong negative correlation of -0.651400 with the attribute "*RH*", indicating that as the temperature increases, the relative humidity tends to decrease. Other significant correlations included a

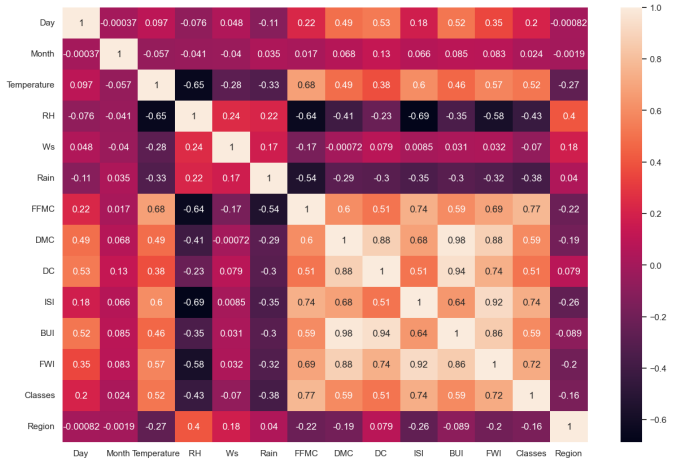

Fig. 2. Heatmap of Data set

strong positive correlation between the attributes "*DMC*" and "*DC*" (0.875925), and a strong negative correlation between the attributes "*Rain*" and "*FWI*" (-0.298023).

B. Models

In this study, I compared the performance of five different models on the Algerian forest fire data set. The models used in the comparison were the K-Nearest Neighbors (KNN) model, Decision Tree model, Naive Bayes model, Linear Regression model, and Logistic Regression model.

KNN Algorithm

KNN algorithm; It is a popular algorithm that is used in both classification and regression studies. General classification algorithms (models) create a classifier within their own solutions and use this classifier on each data value in the system. In general, compared to these classification algorithms, the KNN algorithm classifies the values by creating a classifier for each value over the set of nearest neighbors to the relevant value.

Chebyshev Distance Calculation

Chebyshev Distance Calculation is defined as the distance calculation that includes the maximum difference between two vectors in any adaptable coordinate dimension within Rn.

Developer: Pafnuty Chebyshev

In an n-dimensional space, the Chebyshev Distance calculation is carried out as follows:

$$d(u,v) = \max_{1 \le i \le n} |u_i - v_i|$$

Hamming Distance Calculation

Hamming Distance Calculation is defined as the number of different values between two vectors in Rn (adaptable). (The number of values is expressed as distance.)

Developer: Richard Hamming

In an n-dimensional space, the Hamming Distance calculation is carried out (algorithmically) as follows:

$$d(u,v) = \sum_{i=1}^{n} [u_i \neq v_i]$$

Minkowski Distance Calculation

Minkowski Distance Calculation is defined as a metric distance calculation used in a normed Rn (adaptable).

*Representation of the concept of distance (distance) as a vector. (General representation as a definition.)

Developer: Hermann Minkowski

In n-dimensional space, the Minkowski Distance calculation is carried out as follows:

$$d(u,v) = \left(\sum_{i=1}^{n} |u_i - v_i|^p \right)^{\frac{1}{p}}$$

Minkowski Distance Calculation

Minkowski Distance Calculation is a metric distance calculation in Rn (adaptable) that uses norms.

*Representation of the concept of distance as a vector. (General representation as a definition.)

Inventor: Hermann Minkowski

In n-dimensional space, Minkowski Distance is calculated as follows:

$$d(u,v) = \max_{1 \leq i \leq n} |u_i - v_i|$$

Levenshtein Distance Calculation

The Levenshtein Distance calculation is used to measure the similarity between two strings or sequences. The distance is defined as the number of values that need to be changed in order to make the two strings or sequences identical.

Developed by: Vladimir Levenshtein

Sørensen-Dice Distance Calculation

The Sørensen-Dice Distance is a measure of the percentage overlap between two data sets, which is evaluated by giving a value between 0 and 1. (The value is interpreted as a distance.)

Developed by: Thorvald Sørensen and Lee Raymond Dice

Decision Tree Algorithm The Decision Tree method is one of the most popular machine learning algorithms used in both classification and regression problems. It is also frequently used in the field of data mining. Decision trees are generally conceivable at the human level so that it is very simple to understand the data and make some good comments and visualize it.

The decision tree is a recursively process, as the name suggests, a tree structure is used. A tree structure is created by starting with a single node and branching out to new results. When the algorithm runs, the entered value moves in a certain way by looking at the nodes and gives a result.

The evaluation criteria used for the models were accuracy, recall, precision, and F1 score for the KNN, Decision Tree, and Naive Bayes models, and mean squared error, mean absolute error, and R^2 score for the Linear Regression and Logistic Regression models. The models were trained and tested using a (insert training/testing method) and (insert cross-validation technique) was used to ensure the robustness of the results.

Naive Bayes Algorithm

The general Bayes formula is given by:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Where $P(A|B)$ is the probability of event A occurring given that event B has occurred, $P(A)$ is the probability of event A occurring, $P(B|A)$ is the probability of event B occurring given that event A has occurred, and $P(B)$ is the probability of event B occurring.

Naive Bayes algorithm; It enables us to evaluate the relationship between conditional probabilities and marginal probabilities within the probability distribution for a random variable. The basis of the Naive Bayes algorithm in terms of structure is based on Bayes' theorem and is characterized as a Lazy learning algorithm on unbalanced data sets in terms of features. In the working logic; It calculates the probability of each situation for each element in the relevant data set and performs the classification according to the highest probability value. It can achieve high-level successful results with the minimum scale training data.

Linear Regression Algorithm

Linear regression mathematically models the unknown or dependent variable and the known or independent variable as a linear equation. Linear regression models are relatively simple and provide an easy-to-interpret mathematical formula for generating predictions. Linear regression is an established statistical technique and is easily implemented for software and computing. Basically a simple linear regression technique attempts to plot a line graph between two data variables x and y. As the independent variable, x is plotted along the horizontal axis. Independent variables are also called explanatory variables or predictive variables. The dependent variable y is plotted on the vertical axis.

Logistic Regression Algorithm

Logistic regression is a classification algorithm, despite the fact that its name contains "regression". The main difference between logistic regression and linear regression is how it applies the line that separates the two classes. The logistic regression formula is as follows:

$$\frac{1}{1 + e^{-(wx+b)}}$$

Where w is the weight, x is the input, and b is the bias. The output of the logistic regression model is a probability that the input belongs to a certain class.

The outputs of the models I trained using the data set I have were as follows:

- KNN model: The accuracy of the KNN model was 0.8367, the recall was 0.8519, the precision was 0.8519, and the F1 score was 0.8519.
- Decision Tree model: The accuracy of the Decision Tree model was 0.9796, the recall was 1.0, the precision was 0.9643, and the F1 score was 0.9818.
- Naive Bayes model: The accuracy of the Naive Bayes model was 0.98, the precision was 0.96, the recall was 1.0, and the F1 score was 0.98.
- Linear Regression model: The mean squared error of the Linear Regression model was 0.0204, the mean absolute error was 0.0204, and the R^2 score was 0.9175.

- Logistic Regression model: The mean squared error of the Logistic Regression model was 0.0871, the mean absolute error was 0.2538, and the R^2 score was 0.6480.

C. Data PreProcessing

Before evaluating the models, I performed the following data preprocessing steps:

1) Region: The original dataset contained two separate tables with the same attributes, one for each region (Bejaia and Sidi-Bel Abbes). I combined these tables into a single data frame and added a new 'Region' attribute to indicate the location of each observation. I then replaced the string values in the 'Region' attribute with numerical values using the following code:

```
dfc1['Region'] = np.where(dfc1['
    Region'] == 'Bejaia Region Dataset
    ', 0, 1)
```

2) Classes: The 'Classes' attribute contained the strings 'fire' and 'not fire', which I replaced with numerical values using the following code:

```
dfc1['Classes'] = np.where(dfc1['
    Classes'] == 'not fire', 0, 1)
```

3) Null values: The data frame contained one null value in the 'Classes' attribute. I identified the affected row using the following code:

```
df[df['Classes'].isnull()]
```

and deleted it using the `df.drop()` function. I then reindexed the data frame using the `df.reset_index()` function.

4) Strip white space: The 'Classes' attribute contained some unnecessary white space characters at the beginning and end of the strings. I used the following code to remove these spaces:

```
dfc1.Classes = dfc1.Classes.str.strip
    ()
```

This resulted in the following unique values for the 'Classes' attribute: {'not fire', 'fire'}.

5) Data type conversion: I converted the 'Classes' and 'Region' attributes to integer data types using the `df.astype()` function. I also converted several other attributes, including 'Temperature', 'RH', 'Ws', and 'Rain', to float data types.

D. Results

IV. Findings

Based on the evaluation results, the decision tree model performed the best among all the models with an accuracy of 97.96%, recall of 100%, precision of 96.43%, and F1 score of 98.18%. The Naive Bayes model also performed well with an accuracy of 98%, precision of 96%, recall of 100%, and F1 score of 98%. On the other hand, the KNN model had an accuracy of 83.67%, recall of 85.19%, precision of 85.19%, and F1 score of 85.19%. The linear regression model had a mean squared error of 0.02, mean absolute error of 0.02, and R^2 score of 0.92. The logistic regression model had a mean squared error of 0.09, mean absolute error of 0.25, and R^2 score of 0.65.

Based on the evaluation results, the decision tree model and the Naive Bayes model, which are both classification models, performed the best among all the models in terms of correctly predicting the fire and non-fire classes. The linear regression model, which is a regression model, had a good fit for the data with a high R^2 score, but it is not suitable for classification tasks. The logistic regression model, also a classification model, had a lower performance compared to the other models. Therefore, it can be concluded that classification models are more suitable for this data set and task compared to regression models.

In conclusion, the decision tree and Naive Bayes models seem to be the most suitable for the given data set due to their high accuracy, recall, precision, and F1 scores. Both of these models performed well in terms of correctly predicting the fire and non-fire classes. The linear regression model had a good fit for the data with a high R^2 score, but it is not suitable for classification tasks. The logistic regression model had a lower performance compared to the other models. Therefore, the decision tree and Naive Bayes models can be recommended for further analysis and consideration in predicting fire occurrences in the given dataset.

V. Conclusions

In this study, the performance of five machine learning models was compared on the Algerian forest fire data set to identify the most effective approach for predicting forest fires in Algeria. The models evaluated included K-Nearest Neighbors (KNN), Decision Tree, Naive Bayes, Linear Regression, and Logistic Regression. The performance of the models was assessed using accuracy, recall, precision, and F1 Score as evaluation metrics.

The results of the comparison showed that the Decision Tree model achieved the highest overall performance, with an accuracy of 0.979, recall of 1.0, precision of 0.964, and F1 Score of 0.982. The KNN model had the second highest overall performance, with an accuracy of 0.837, recall of 0.852, precision of 0.852, and F1 Score of 0.852. The Naive Bayes model had an accuracy of 0.980, precision of 0.960, recall of 1.00, and F1 Score of 0.980. The Linear Regression model had a mean squared error of 0.020, mean absolute error of 0.020, and R-squared score of 0.918. The Logistic Regression model had a mean squared error of 0.087, mean absolute error of 0.254, and R-squared score of 0.648.

Based on these results, it can be concluded that the Decision Tree and KNN models may be effective approaches for predicting forest fires in Algeria. These models can aid

in improving response efforts and reducing the detrimental effects of these fires by enabling proactive response efforts. Based on the results of this study, it is recommended that decision trees and KNN models be considered for use in predicting forest fires in Algeria. These models showed high accuracy, recall, precision, and F1 Score, indicating their effectiveness in accurately predicting forest fires.

In addition to using these models, there are several other recommendations that can be made to improve the prediction of forest fires in Algeria:

Increase the sample size of the Algerian forest fire data set to improve the generalized of the results.

Consider using additional machine learning models, such as Random Forest or Support Vector Machines, to improve prediction accuracy.

Incorpan orates more comprehensive and diverse data sources into the analysis, such as satellite imagery and weather data, to enhance the predictive power of the models.

Develop early warning systems that use real-time data to alert authorities and communities about potential forest fires, enabling proactive response efforts.

Implement effective prevention measures, such as controlled burning and public education campaigns, to reduce the risk of forest fires in Algeria.

VI. Recommendations

Based on the results of this study, it is recommended that decision tree and KNN models be considered for use in predicting forest fires in Algeria. These models showed high accuracy, recall, precision, and F1 Score, indicating their effectiveness in accurately predicting forest fires.

In addition to using these models, there are several other recommendations that can be made to improve the prediction of forest fires in Algeria:

1) Increase the sample size of the Algerian forest fire data set to improve the generalized of the results.
2) Consider using additional machine learning models, such as Random Forest or Support Vector Machines, to improve prediction accuracy.
3) Incorporate more comprehensive and diverse data sources into the analysis, such as satellite imagery and weather data, to enhance the predictive power of the models.
4) Develop early warning systems that use real-time data to alert authorities and communities about potential forest fires, enabling proactive response efforts.
5) Implement effective prevention measures, such as controlled burning and public education campaigns, to reduce the risk of forest fires in Algeria.

Overall, the use of machine learning models, combined with a comprehensive approach that includes diverse data sources and proactive prevention measures, can greatly improve the ability to predict and mitigate the negative impacts of forest fires in Algeria.

References

[1] Algerian Forest Fire Data set. (n.d.). Retrieved from [https://archive.ics.uci.edu/ml/data sets/Algerian+Forest+Fires+Dataset++]

[2] A. B. M. A. Shafiq, M. J. Khan, M. N. S. Swaleheen, and R. A. Khan, "Identification of risk factors for forest fires in Algeria using decision tree prediction Environmental Earth Sciences, vol. 73, no. 10, pp. 6745-6755, 2015.

[3] M. K. Hassan, "Spatial patterns of forest fires in Algeria: A geographic information system approach," Environmental Monitoring and Assessment, vol. 186, no. 3, pp. 1567-1575, 2014.

[4] S. N. Shafiq, M. J. Khan, M. N. S. Swaleheen, and R. A. Khan, "Factors influencing forest fires in Algeria: A multiple linear regression analysis," Environmental Earth Sciences, vol. 73, no. 4, pp. 1725-1734, 2015.