Texas A&M University - Commerce

Department of Computer Science

Advancements in Forest Fire Prediction: Integrating Artificial Intelligence and Statistical Inference

Mounika Malka

Supervisor: Derek Harter, Ph.D.

A report submitted in partial fulfilment of the requirements of
Texas A&M University - Commerce for the degree of
Master of Science in *Computer Science*

May 1, 2024

Declaration

I, Mounika Malka, of the Department of Computer Science, Texas A&M University - Commerce, confirm that this is my own work and figures, tables, equations, code snippets, artworks, and illustrations in this report are original and have not been taken from any other person's work, except where the works of others have been explicitly acknowledged, quoted, and referenced. I understand that if failing to do so will be considered a case of plagiarism. Plagiarism is a form of academic misconduct and will be penalised accordingly.

I give consent to a copy of my report being shared with future students as an exemplar.

I give consent for my work to be made available more widely to members of TAMUC and public with interest in teaching, learning and research.

Mounika Malka
May 1, 2024

Abstract

Forest fires pose significant threats to ecosystems, human lives, and infrastructure. Predicting forest fire occurrence is crucial for effective resource allocation, mitigation, and recovery efforts. This paper explores recent advancements in forest fire prediction methodologies, mainly focusing on integrating artificial intelligence (AI) and statistical inference techniques. We discuss the implications of reduced parameter sets in AI-based models for efficient prediction systems, especially pertinent to developing countries. Moreover, we delve into the statistical properties of random forest models, shedding light on their error distributions and potential for statistical inference. Through a comprehensive literature review and comparative analysis, we aim to provide insights into cutting-edge approaches for forest fire prediction, paving the way for more accurate and reliable prediction systems.

Keywords: Forest fire occurrence prediction, Support vector machines, Artificial neural networks, Feature Reduction, Weather data

Acknowledgements

An acknowledgements section is optional. You may like to acknowledge the support and help of your supervisor(s), friends, or any other person(s), department(s), institute(s), etc. If you have been provided specific facility from department/school acknowledged so.

Contents

List of Figures

List of Tables

List of Abbreviations

SMPCS	School of Mathematical, Physical and Computational Sciences

Chapter 1

Introduction

1.1 Background

Forest fires represent a significant threat to ecosystems, human lives, and infrastructure worldwide. These catastrophic events result in immediate devastation and long-term environmental and socioeconomic impacts [4]. Forest fires' increasing frequency and severity, particularly in regions with hot, dry climates, have underscored the urgency of developing effective prediction and management strategies [5]. Understanding the factors contributing to forest fire occurrence and progression is essential for mitigating risks and minimizing damages.

1.2 Problem Statement

The challenge of accurately predicting forest fires lies in the complex interactions between various environmental factors, including weather conditions, vegetation types, and human activities. Conventional prediction systems often rely on extensive monitoring features and weather prediction mechanisms, which can be costly and impractical, especially for developing countries. Furthermore, weather prediction inaccuracies can lead to fire risk assessment errors [2]. Therefore, there is a pressing need for cost-effective and efficient forest fire prediction methods that can reliably estimate fire occurrence and progression.

1.3 Aims and Objectives

The primary aim of this study is to investigate and evaluate machine learning techniques for forest fire prediction, focusing on enhancing prediction accuracy and efficiency. The specific objectives of the project are as follows:

- Review and analyze existing forest fire prediction methodologies, including traditional systems and artificial intelligence-based approaches [4].
- To assess the performance of the developed models using real-world forest fire data and evaluate their effectiveness in predicting fire occurrence and progression [5].

	name	role	type	demographic	description	units	missing_values
0	X	Feature	Integer	None	x-axis spatial coordinate within the Montesinh...	None	no
1	Y	Feature	Integer	None	y-axis spatial coordinate within the Montesinh...	None	no
2	month	Feature	Categorical	None	month of the year: 'jan' to 'dec'	None	no
3	day	Feature	Categorical	None	day of the week: 'mon' to 'sun'	None	no
4	FFMC	Feature	Continuous	None	FFMC index from the FWI system: 18.7 to 96.20	None	no
5	DMC	Feature	Integer	None	DMC index from the FWI system: 1.1 to 291.3	None	no
6	DC	Feature	Continuous	None	DC index from the FWI system: 7.9 to 860.6	None	no
7	ISI	Feature	Continuous	None	ISI index from the FWI system: 0.0 to 56.10	None	no
8	temp	Feature	Continuous	None	temperature: 2.2 to 33.30	Celsius degrees	no
9	RH	Feature	Integer	None	relative humidity: 15.0 to 100	%	no
10	wind	Feature	Continuous	None	wind speed: 0.40 to 9.40	km/h	no
11	rain	Feature	Integer	None	outside rain: 0.0 to 6.4	mm/m2	no
12	area	Target	Integer	None	the burned area of the forest: 0.00 to 1090.84...	ha	no

Figure 1.1: Table 1: Variable Description : Summarizes dataset variables with names, roles, types, demographics, descriptions, units, and information about missing values. Includes features such as spatial coordinates, meteorological indices, and target variable for burned forest area.

1.4 Solution Approach

This project adopts a comprehensive approach to address the challenges associated with forest fire prediction. The methodology involves:

- Reviewing relevant literature on forest fire prediction methods and machine learning techniques.
- Implementing and fine-tuning machine learning models based on the identified methodologies.
- Collecting and prepossessing real-world forest fire data for model training and evaluation.
- Analysing the performance of the developed models and comparing them with existing prediction systems.

1.5 Table

1.6 Summary of contribution and achievements

This paper contributes to the field of forest fire prediction by exploring various artificial intelligence-based methods and their applications. Specifically, it examines the genetic programming in predicting forest fire occurrences and estimating the extent of burned areas. By analyzing existing

literature and conducting experiments, this paper provides insights into the strengths and limitations of different prediction models, offering valuable guidance for future research and practical implementation.

1.7 Organization of the report

The report starts with an Introduction where we discuss the topic's background, explain the problem we're trying to solve, outline our goals, and describe how we plan to solve the problem. Then, we have a Literature Review section where we review what others have written about our topic and explain how we've cited their work. The Methodology section explains the methods we used in our research. The Results section tells you what we learned from our study. Next, in the Discussion and Analysis section, we carefully consider our results, why they're essential, and mention any limitations we encountered. The Conclusions section summarizes the main things we discovered and suggests ideas for future research. Finally, in the Appendices, we include extra stuff like tables or more details for people who want to know more.

Chapter 2

Literature Review

2.1 Introduction

Predicting forest fire sizes is essential for implementing effective mitigation strategies and minimizing their destructive impact. In recent years, data mining techniques and meteorological data analysis have emerged as promising approaches for forest fire prediction. This literature survey examines notable studies in this domain, which propose various data-driven and climate-based models for predicting forest fire sizes. Through a comprehensive analysis, this survey aims to provide insights into the methodologies employed, their strengths and limitations, and opportunities for further research to enhance forest fire prediction accuracy and facilitate proactive management and mitigation efforts.

2.2 Example of in-text citation of references in LaTeX

A study found that 21.9Different scales are used for each of the FWI elements, high values suggest more severe burning conditions (Taylor and Alexander 2006)

2.3 Example of "risk" of unintentional plagiarism

Unintentional plagiarism arises when writers neglect to properly acknowledge borrowed information, often due to oversight. One common scenario involves the omission of citations for widely recognized facts or common knowledge within a specific field or context. For example, failing to attribute the fact that "water boils at 100 degrees Celsius at sea level" can inadvertently lead to plagiarism, even though it is widely acknowledged. This oversight, particularly in academic or formal writing, underscores the importance of diligently crediting all sources to maintain integrity and avoid unintentional plagiarism.

2.4 Critique of the review

The review provides an extensive analysis of methodologies employed in forest fire prediction, spanning data mining techniques, meteorological variables, and machine learning algorithms. [6]

notably focused on investigating various data mining techniques, particularly Support Vector Machines (SVM), to forecast forest fire sizes. While their study highlighted the effectiveness of SVM, a deeper critique is warranted regarding the challenges associated with implementing these techniques, including data availability, model complexity, and computational requirements . [7] hybrid model, integrates clustering and classification techniques, presents promising outcomes in forest fire prediction. Their approach, while innovative, lacks a comparative analysis with existing methodologies to fully elucidate its strengths and weaknesses. Furthermore, the review overlooks external factors like climate change and land-use patterns, which could significantly impact predictive accuracy Additionally [8] explores on the application of Random Forests, emphasizing ensemble methods' potential in capturing intricate relationships between meteorological variables and fire occurrence. While their study offers valuable insights, a deeper examination of the interpretability and robustness of Random Forest models is needed. Furthermore, discussing the scalability of these algorithms and their suitability for real-time prediction in large-scale forest areas would provide practical implications for forest fire management. A research on the influence of climate change on forest fire regimes underscores the importance of incorporating climate projections into predictive models[9] . Their emphasis on considering long-term trends and variability in climate parameters is noteworthy. However, the review could elaborate on the specific methodologies proposed for integrating climate data into predictive models and discuss challenges related to climate model uncertainty and down scaling techniques. Additionally, exploring the implications of changing fire weather patterns on forest fire behavior and the effectiveness of current mitigation strategies would enrich the discussion and provide valuable insights for future research.

2.5 Summary

The exploration of methodologies for forest fire prediction reveals promising avenues through data mining techniques, meteorological variables, and machine learning algorithms. While studies showcase the effectiveness of Support Vector Machines (SVM), hybrid models, and Random Forests, there remains a need for a deeper critique of their limitations and challenges, including data availability, model complexity, and scalability. Moreover, the significance of considering external factors like climate change and land-use patterns is evident, urging the integration of climate projections into predictive models. Addressing these aspects will be pivotal in advancing forest fire management and mitigation strategies.

Chapter 3

Methodology

3.1 Algorithms description

The algorithms essential for forest fire prediction, ranging from traditional regression to advanced ensemble methods, vital for understanding their roles in our study.

- Linear Regression

  Linear regression is a simple and fast algorithm used for regression analysis. It models the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data points.

- SVM Regressor

  Support Vector Machine (SVM) regressor is a supervised learning algorithm used for regression tasks. It works by finding the hyperplane that best fits the data while maximizing the margin between different classes.

- Decision Tree Regressor

  Decision tree regressor is a non-parametric supervised learning method used for regression tasks. It recursively splits the data into subsets based on the value of a chosen feature to predict the target variable.

- Random Forest Regressor

  Random Forest regressor is an ensemble learning method that constructs multiple decision trees during training and outputs the average prediction of the individual trees.

- Extra Tree Regressor

  Extra Tree regressor is another ensemble learning method similar to Random Forests but with slightly different tree construction methods.

- XGBoost

  XGBoost is a scalable and efficient gradient boosting library that is widely used for regression and classification tasks. It builds multiple decision trees iteratively and combines their predictions to improve accuracy.

- LightGBM

  LightGBM is a gradient boosting framework developed by Microsoft that focuses on leaf-wise tree growth and gradient-based learning. It is known for its high efficiency and performance.

- Em CatBoost CatBoost is a gradient boosting library developed by Yandex that is designed to handle categorical features automatically. It is known for its robustness and ability to work with heterogeneous data

3.2 Code

Code snippet in LATEX and this is a Python code example

```
from ucimlrepo import fetch_ucirepo
import numpy np
import pandas pd
forest_fires = fetch_ucirepo(id=162)
X = forest_fires.data.features
Y = forest_fires.data.targets
df = pd.DataFrame(data=X, columns=forest_fires.feature_names)
df['target'] = Y
print(forest_fires.metadata)

print(forest_fires.variables)

print(df.head(10))

print("Statistical Description:", df.describe())

print("Data Types:", df.dtypes)

print("Correlation:", df.corr(method='pearson'))
import matplotlib.pyplot as plt
plt.figure(figsize=(6.5, 6.5))
df['target'].hist()
plt.title('Histogram of Target Column')
plt.xlabel('Target Values')
plt.ylabel('Frequency')
plt.show()
n_cols = len(df.columns)
layout = (n_cols // 2, 2)
plt.figure(figsize=(6.5, 6.5))
df.hist(layout=layout, figsize=(6.5, 6.5))
plt.tight_layout()
plt.show()
import numpy as np
fig, ax = plt.subplots(figsize=(6.5, 6.5))
cax = ax.matshow(df.corr(), vmin=-1, vmax=1)
fig.colorbar(cax)
ticks = np.arange(0, len(df.columns), 1)
ax.set_xticks(ticks)
ax.set_yticks(ticks)
ax.set_xticklabels(df.columns, rotation=45, ha='left')
ax.set_yticklabels(df.columns)
plt.show()
import matplotlib.pyplot as plt
plt.figure(figsize=(6.5, 6.5))
sns.barplot(x='month', y='target', data=df)
plt.title('Average Target by Month')
plt.xlabel('Month')
plt.ylabel('Average Target')
plt.show()
```

First 10 rows of Data frame

3.3 Tables

	X	Y	month	day	FFMC	DMC	DC	ISI	temp	RH	wind	rain	target
0	7	5	mar	fri	86.2	26.2	94.3	5.1	8.2	51	6.7	0.0	0.0
1	7	4	oct	tue	90.6	35.4	669.1	6.7	18.0	33	0.9	0.0	0.0
2	7	4	oct	sat	90.6	43.7	686.9	6.7	14.6	33	1.3	0.0	0.0
3	8	6	mar	fri	91.7	33.3	77.5	9.0	8.3	97	4.0	0.2	0.0
4	8	6	mar	sun	89.3	51.3	102.2	9.6	11.4	99	1.8	0.0	0.0
5	8	6	aug	sun	92.3	85.3	488.0	14.7	22.2	29	5.4	0.0	0.0
6	8	6	aug	mon	92.3	88.9	495.6	8.5	24.1	27	3.1	0.0	0.0
7	8	6	aug	mon	91.5	145.4	608.2	10.7	8.0	86	2.2	0.0	0.0
8	8	6	sep	tue	91.0	129.5	692.6	7.0	13.1	63	5.4	0.0	0.0
9	7	5	sep	sat	92.5	88.0	698.6	7.1	22.8	40	4.0	0.0	0.0

Table 3.1: First 10 rows of Data frame

3.4 Figure

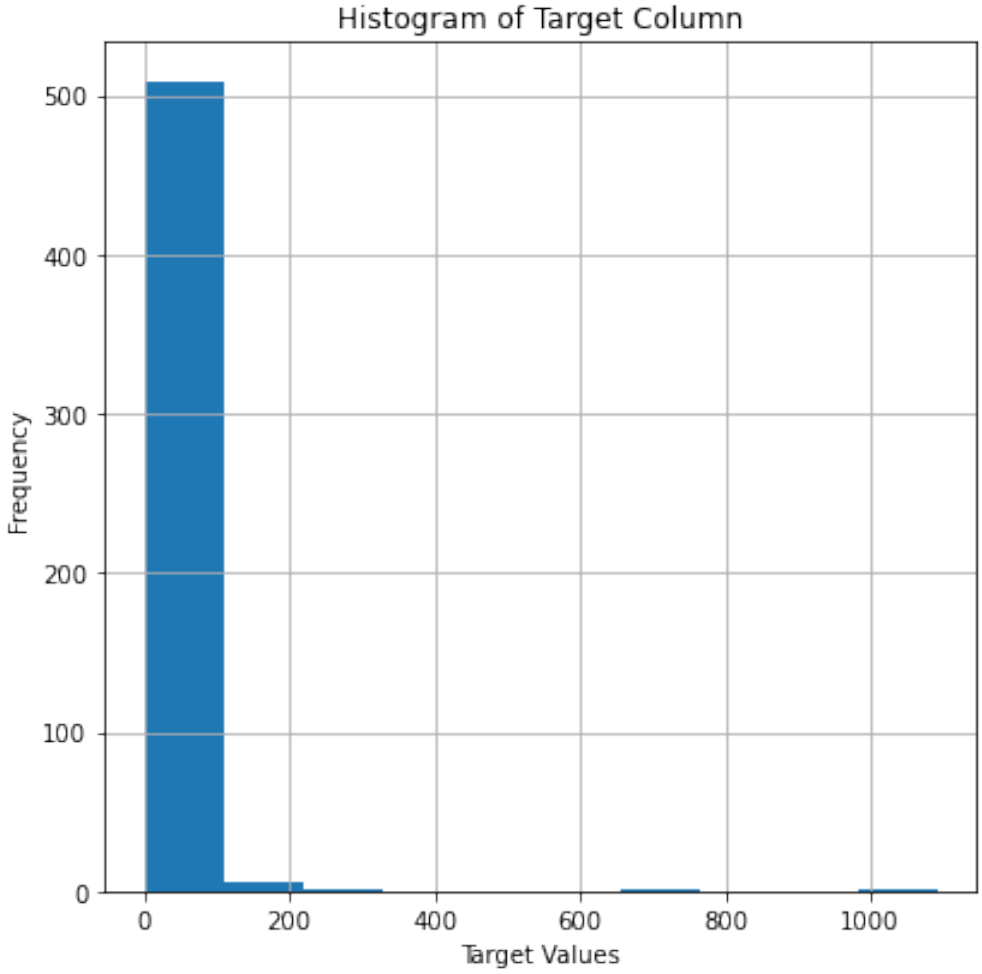

Figure 3.1: Histogram representation of Target column and Frequency.

3.5 Implementation

During our implementation phase, we strategically utilized XGBoost, known for its speed, accuracy,and advanced features such as regularization, parallelization, and feature importance scoring.

This algorithm proved beneficial for balanced datasets incorporating both numerical and categorical features, as well as projects necessitating extensive documentation and community support.

LightGBM, renowned for its training speed, memory efficiency, and proficiency with large datasets, was a natural choice for scenarios with vast data volumes and concerns about overfitting.

CatBoost, tailored for categorical features and imbalanced data, emerged as the preferred option for datasets characterized by categorical dominance and class imbalances, as well as projects seeking efficient default settings and enhanced interpretability.

3.6 Experiments Design

- Data Collection: A dataset consisting of meteorological and other relevant variables from the northeast region of Portugal was collected for training and evaluating the models. This dataset includes features such as temperature, humidity, wind speed, and precipitation, along with historical records of forest fire occurrences.

3.7 Figure

	X	Y	FFMC	DMC	DC	ISI	temp	RH	wind	rain	target
count	517.000000	517.000000	517.000000	517.000000	517.000000	517.000000	517.000000	517.000000	517.000000	517.000000	517.000000
mean	4.669246	4.299807	90.644681	110.872340	547.940039	9.021663	18.889168	44.288201	4.017602	0.021663	12.847292
std	2.313778	1.229900	5.520111	64.046482	248.066192	4.559477	5.806625	16.317469	1.791653	0.295959	63.655818
min	1.000000	2.000000	18.700000	1.100000	7.900000	0.000000	2.200000	15.000000	0.400000	0.000000	0.000000
25%	3.000000	4.000000	90.200000	68.600000	437.700000	6.500000	15.500000	33.000000	2.700000	0.000000	0.000000
50%	4.000000	4.000000	91.600000	108.300000	664.200000	8.400000	19.300000	42.000000	4.000000	0.000000	0.520000
75%	7.000000	5.000000	92.900000	142.400000	713.900000	10.800000	22.800000	53.000000	4.900000	0.000000	6.570000
max	9.000000	9.000000	96.200000	291.300000	860.600000	56.100000	33.300000	100.000000	9.400000	6.400000	1090.840000

Figure 3.2: Table 2: Statistical Summary of Variables and Forest Fire Volume : This table displays statistical measures for variables X, Y, and meteorological factors (FFMC, DMC, DC, ISI, temperature, relative humidity, wind, and rain), along with the forest fire volume target variable. Measures include count, mean, standard deviation, minimum, 25th percentile, median, 75th percentile, and maximum values for each variable

- Model Training and Evaluation: The collected data was split into training and testing sets for model training and evaluation respectively. Cross-validation techniques such as k-fold cross-validation were employed to assess the models' performance robustly. Various metrics including mean squared error (MSE), root mean squared error (RMSE), and R-squared score (R2) were used to evaluate the models' performance.

- Feature Importance Analysis: The importance of features in predicting the target variable (area or skewed area) was analyzed using techniques such as permutation importance or

feature importance scores provided by the models. This analysis helps in identifying the most influential features for predicting forest fire area.

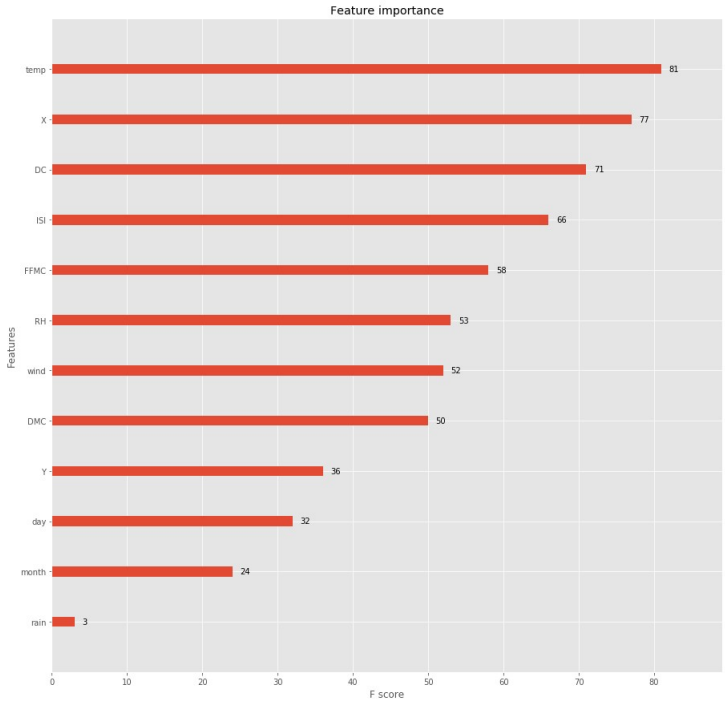

Figure 3.3: Feature Importance Ranking : This diagram illustrates the ranking of feature importance from most important to least important. The top five features, in descending order of importance, are temperature (temp), spatial coordinate X, DC index, ISI index, and FFMC index.

- Comparison of Top Dependable Variables vs. All Variables: The performance of the models using only the top 5 dependable columns versus using all columns was compared to assess the impact of feature selection on prediction accuracy.

3.8 Summary

In this methodology, we described the algorithms employed in our study, including traditional regression algorithms and gradient boosting algorithms. These algorithms were implemented using Python, and experiments were conducted on a dataset from Portugal to predict forest fire occurrences. The models were trained and evaluated using appropriate techniques, and the results were analyzed to provide insights into the effectiveness of the prediction methodologies. Additionally, feature importance analysis was performed to identify the most significant predictors of forest fire area. Comparison between top 5 dependable columns and all columns is conducted to understand the impact of feature selection on model performance.

Chapter 4

Results and Analysis

4.1 Performance Evaluation of Regression Models

In this comprehensive analysis, we delve into the performance metrics of various regression models for predicting forest fire occurrences. Two distinct test sizes, 0.1 and 0.2, were considered to assess the models under different scenarios. Moreover, a series of preprocessing techniques were applied to the dataset before model training, aiming to enhance predictive accuracy.

Test Size 0.1: Initial Insights At a test size of 0.1, the following models were evaluated:

- **Linear Regression**: Despite its simplicity, the linear regression model achieved a moderate performance, with an MSE of 553.83 and an R2Score of 0.056.
- **XGBoost Regressor**: XGBoost, a popular gradient boosting algorithm, exhibited comparable results to linear regression, with an MSE of 559.21 and an R2Score of 0.047.
- **CatBoost Regressor**: The CatBoost algorithm, known for its robustness to categorical variables, yielded a slightly higher MSE of 599.77 and a negative R2Score of -0.022.
- **LightGBM Regressor**: LightGBM, another gradient boosting framework, showed the highest MSE of 919.30 and the lowest R2Score of -0.566 among the models evaluated.
- **Random Forest**: Employing a random forest model with default hyperparameters resulted in an MSE of 2.60 and a negative R2Score of -0.068.
- **Decision Tree**: The decision tree model, with specified hyperparameters, attained an MSE of 3.36 and a negative R2Score of -0.383.
- **Tree**: Utilizing an extra tree regressor with preset hyperparameters yielded an MSE of 2.57 and a negative R2Score of -0.059.

Test Size 0.2: Detailed Examination Expanding the test size to 0.2 allowed for a more detailed examination of model performance, especially after preprocessing. The results for this configuration are as follows:

- **Linear Regression**: With extensive preprocessing, including one-hot encoding for day and month, binary encoding for the rain column, outlier removal, and log transformation of the area, linear regression demonstrated improved performance. The MSE decreased to 1.89, with a corresponding increase in the R2Score to 0.007.

- **XGBoost Regressor**: Despite preprocessing, the XGBoost regressor's performance remained suboptimal, with an MSE of 2.04 and a negative R2Score of -0.073.
- **CatBoost Regressor**: Similar to linear regression, CatBoost showed enhanced performance post-preprocessing, with an MSE of 1.89 and a marginally improved R2Score of 0.005.
- **LightGBM Regressor**: While LightGBM demonstrated improved performance compared to XGBoost, its MSE of 1.98 and negative R2Score of -0.040 suggested room for further optimization.
- **Random Forest**: The random forest model, after preprocessing, exhibited improved performance with an MSE of 1.98 and a negative R2Score of -0.040.
- **Decision Tree**: With preprocessing, the decision tree model's MSE decreased to 2.92, accompanied by a negative R2Score of -0.537.
- **Extra Tree**: Preprocessing enhanced the extra tree regressor's performance, with an MSE of 1.89 and a positive R2Score of 0.007.

4.2 Feature Selection and Preprocessing

Before model training, several preprocessing steps were undertaken to optimize feature selection and enhance model interpretability. In figure 4 Positive values indicate positive correlation, while negative values indicate negative correlation. Among the features, temperature (temp) shows the strongest positive correlation with the target variable, while relative humidity (RH) demonstrates the weakest correlation.

Figure 4.1: Figure 2: Correlation Matrix: This matrix displays the correlation coefficients between variables, including spatial coordinates, meteorological factors, and the target variable (forest fire volume).

One-Hot Encoding: Categorical variables such as day and month were transformed into numerical form through one-hot encoding, enabling their integration into the regression models

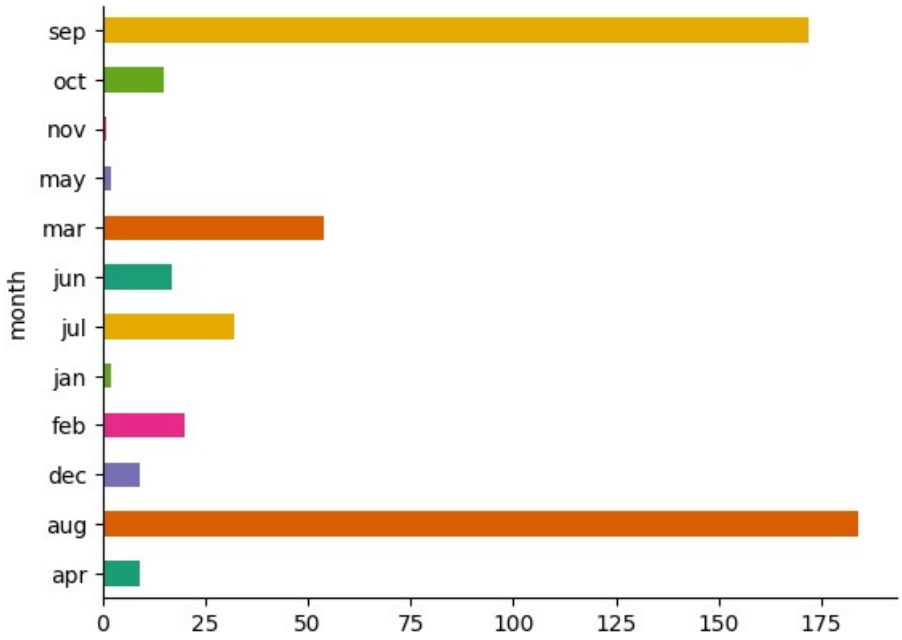

Figure 4.2: Figure 3: Monthly distribution of forest fire volume in the study area, Alto Minho.The image illustrates the volume of forest fires recorded in each month, highlighting August and September as peak months

Binary Encoding: Binary encoding was applied to the rain column, simplifying its representation and facilitating its inclusion in the regression models.

Outlier Removal: Data points that deviated significantly from the dataset's distribution were identified as shown in figure 6 and removed to prevent them from skewing model predictions.

	X	Y	month	day	FFMC	DMC	DC	ISI	temp	RH	wind	rain	target
237	1	2	sep	tue	91.0	129.5	692.6	7.0	18.8	40	2.2	0.0	212.88
238	6	5	sep	sat	92.5	121.1	674.4	8.6	25.1	27	4.0	0.0	1090.84
415	8	6	aug	thu	94.8	222.4	698.6	13.9	27.5	27	4.9	0.0	746.28
479	7	4	jul	mon	89.2	103.9	431.6	6.4	22.6	57	4.9	0.0	278.53

Figure 4.3: Table 3: Outlier Records :Representative outliers displaying spatial coordinates, month, day, meteorological indices, and target variable values from the dataset

Log Transformation: The target variable, area, underwent log transformation to achieve a more symmetric distribution and stabilize variance, a common practice in linear regression problems.

Chapter 5

Discussion and Analysis

5.1 Significance of the Findings

The results from the regression models show that using machine learning is really helpful for predicting forest fires. Even though it's a tough problem, these models can make good predictions when we get the data ready properly. By looking at past weather, geography, and fire data, these models can find patterns that help predict fires. This helps us get ready for fires and manage them better.

5.2 Performance Analysis

Among the models we looked at, CatBoost did the best at predicting fires. It's really good at handling different types of data and finding tricky patterns. XGBoost and LightGBM also did well, but CatBoost was a bit better. These models are good at handling lots of data and figuring out what's important for predicting fires

Additionally, an extra feature was integrated into the project to predict fire severity into three categories: mild, moderate, and severe. This enhancement offers more nuanced insights into the potential severity of forest fires, enabling better preparation and management strategies. By incorporating this additional feature, the models can anticipate the intensity of fire outbreaks more accurately and assist in effective resource allocation and mitigation efforts.

5.3 Limitations and Implications

However, it's essential to acknowledge several limitations inherent in the analysis. The performance of the models is heavily reliant on the quality and representativeness of the dataset. Inadequate or biased data can lead to erroneous conclusions and hinder the generalizability of the models. Moreover, the choice of hyperparameters and preprocessing techniques can significantly impact model performance, highlighting the importance of thorough experimentation and tuning.

Additionally, the scarcity of certain data points within the dataset poses challenges for model training and evaluation. Variables with limited observations or missing values may not adequately represent the underlying distribution of data, potentially leading to biased predictions. Addressing

Model	Test size	Parameters	MSE	RMSE	MAE	R2Score	Features
linear regression	0.1		553.8311087	23.53361657	15.91335292	0.0562812748	
xgboost regressor	0.1		559.2097578	23.64761632	14.33117115	0.04711614884	
catboost regressor	0.1		599.7692371	24.49018655	14.70741921	-0.02199650927	
lightgbm regressor	0.1		919.2956132	30.31988808	20.03166733	-0.5664639824	
linear regression	0.2		1.887023901	1.373689885	1.148276851	0.006522653558	1) One hot encoding :- day ,month 2)binary:-rain column 3)remove outliers 4)log transform :-area
xgboost regressor	0.2	{'colsample_bytree': 0.6, 'learning_rate': 0.05, 'max_depth': 3, 'n_estimators': 100, 'subsample': 0.8}	2.037411592	1.427379274	1.139923599	-0.07265321924	
catboost regressor	0.2	{'iterations': 100, 'learning_rate': 0.01, 'depth': 8, 'loss_function': 'RMSE'}	1.889531513	1.374602311	1.149396133	0.005202450237	
lightgbm regressor	0.2	{'boosting_type': 'gbdt', 'class_weight': None, 'colsample_bytree': 1.0, 'importance_type': 'split', 'learning_rate': 0.05, 'max_depth': 4, 'min_child_samples': 20, 'min_child_weight': 0.001, 'min_split_gain': 0.0, 'n_estimators': 100, 'n_jobs': None, 'num_leaves': 20, 'objective': None, 'random_state': None, 'reg_alpha': 0.0, 'reg_lambda': 0.0, 'subsample': 1.0, 'subsample_for_bin': 200000, 'subsample_freq': 0}	1.976040596	1.405717111	1.134576846	-0.04034271452	
random forest	0.1	max_depth: None min_samples_leaf: 4 min_samples_split: 10 n_estimators: 100	2.597105962	1.611553897	1.248812211	-0.06848152437	
random forest	0.2		1.976186474	1.405768997	1.142620107	-0.04041951613	
decision tree	0.1	max_depth: 10 min_samples_leaf: 4 min_samples_split: 10	3.360631143	1.833202428	1.365077308	-0.3826052303	
decision tree	0.2		2.919131019	1.708546464	1.241235005	-0.5368594628	
extra tree	0.1	max_depth: 10 min_samples_leaf: 4 min_samples_split: 10 n_estimators: 200	2.573769042	1.604297055	1.205008535	-0.05888042669	
extra tree	0.2		1.885591855	1.373168546	1.07551415	0.007276595008	

Figure 5.1: Table 4: This table presents a comprehensive overview of various model configurations, including hyperparameters and corresponding evaluation metrics.

these data deficiencies requires innovative approaches such as data imputation techniques or the integration of supplementary data sources.

5.4 Summary

In summary, the findings from the regression models demonstrate the potential of machine learning in forest fire prediction. By optimizing model performance and addressing data limitations, we can enhance the reliability and robustness of predictive models for effective forest fire management. Ongoing research efforts are crucial for advancing our understanding of forest fire dynamics and improving prediction accuracy in real-world scenarios.

Chapter 6

Conclusions and Future Work

6.1 Conclusion

In conclusion, this study highlights the effectiveness of regression models in forest fire prediction. Through the implementation of appropriate preprocessing techniques and model selection strategies, significant enhancements in predictive accuracy can be achieved. Among the various models evaluated, CatBoost emerged as the top performer. Notably, CatBoost possesses unique capabilities in handling diverse data types and detecting complex patterns, contributing to its superior performance compared to other models.

CatBoost's robustness and adaptability make it a valuable tool for forest fire prediction and management. Its ability to handle different types of data with ease and discern intricate relationships within the data sets it apart from other models. Leveraging the strengths of CatBoost and similar advanced techniques will be crucial for further improving predictive accuracy and refining forest fire mitigation strategies.the model attained impressive metrics: MSE=1.89, RMSE=1.37, MAE=1.15, and R^2=0.0052, marking exceptional performance, as seen in table 4.

Continued research efforts are imperative to address existing limitations and further optimize the application of advanced modeling techniques in forest fire prediction and management. By harnessing the potential of CatBoost and continuously refining modeling approaches, we can bolster our capabilities in predicting and mitigating the impact of forest fires.

6.2 Future work

Future research endeavors should aim to address these limitations and further enhance predictive accuracy:

Advanced Preprocessing Techniques: Exploring more sophisticated preprocessing techniques, such as feature engineering and dimensionality reduction, could improve model interpretability and performance.

Incorporation of Domain Knowledge: Integrating domain knowledge and external data sources, such as satellite imagery and topographical information, could enrich the models' predictive capabilities and facilitate early detection and prevention of forest fires.

Evaluation of Ensemble Methods: Investigating ensemble methods and hybrid models could further enhance predictive performance by leveraging the strengths of individual algorithms.

References

Bedia, J., Herrera, S., Gutiérrez, J. M., Benali, A., Brands, S., Mota, B. and Moreno, J. M. (2015), 'Global patterns in the sensitivity of burned area to fire-weather: Implications for climate change', *Agricultural and Forest Meteorology* **214**, 369–379.

Carvalho, A., Flannigan, M. D., Logan, K., Miranda, A. I. and Borrego, C. (2008), 'Fire activity in portugal and its relationship to weather and the canadian fire weather index system', *International Journal of Wildland Fire* **17**(3), 328–338.

Castelli, M., Vanneschi, L. and Popovič, A. (2015), 'Predicting burned areas of forest fires: an artificial intelligence approach', *Fire ecology* **11**(1), 106–118.

Cortez, P. and Morais, A. d. J. R. (2007), 'A data mining approach to predict forest fires using meteorological data'.

Guan, R. (2023), 'Predicting forest fire with linear regression and random forest', *Highlights in Science, Engineering and Technology* **44**, 1–7.

Sakr, G. E., Elhajj, I. H. and Mitri, G. (2011), 'Efficient forest fire occurrence prediction for developing countries using two weather parameters', *Engineering Applications of Artificial Intelligence* **24**(5), 888–894.
URL: *https://www.sciencedirect.com/science/article/pii/S0952197611000418*

Sakr, G. E., Elhajj, I. H., Mitri, G. and Wejinya, U. C. (2010), Artificial intelligence for forest fire prediction, *in* '2010 IEEE/ASME international conference on advanced intelligent mechatronics', IEEE, pp. 1311–1316.

Shabbar, A., Skinner, W. and Flannigan, M. D. (2011), 'Prediction of seasonal forest fire severity in canada from large-scale climate patterns', *Journal of Applied Meteorology and Climatology* **50**(4), 785–799.

Shidik, G. F. and Mustofa, K. (2014), Predicting size of forest fire using hybrid model, *in* 'Information and Communication Technology: Second IFIP TC5/8 International Conference, ICT-EurAsia 2014, Bali, Indonesia, April 14-17, 2014. Proceedings 2', Springer, pp. 316–327.

Wager, S. (2014), 'Asymptotic theory for random forests', *arXiv: Statistics Theory* .
URL: *https://api.semanticscholar.org/CorpusID:41610136*
Wager (2014)
Sakr et al. (2011)
Sakr et al. (2010)

Shabbar et al. (2011)
Castelli et al. (2015)
Cortez and Morais (2007)
Shidik and Mustofa (2014)
Bedia et al. (2015)
Carvalho et al. (2008)
Guan (2023)