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*

March 25, 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
March 25, 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].

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 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.

1.6 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 downscaling 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 Tables

Data types of each column in the DataFrame df

X	int64
Y	int64
month	object
day	object
FFMC	float64
DMC	float64
DC	float64
ISI	float64
temp	float64
RH	int64
wind	float64
rain	float64
target	float64

Table 3.1: Data types of each column in the DataFrame df

3.3 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

3.4 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 3.2: First 10 rows of Data frame

3.5 Figure

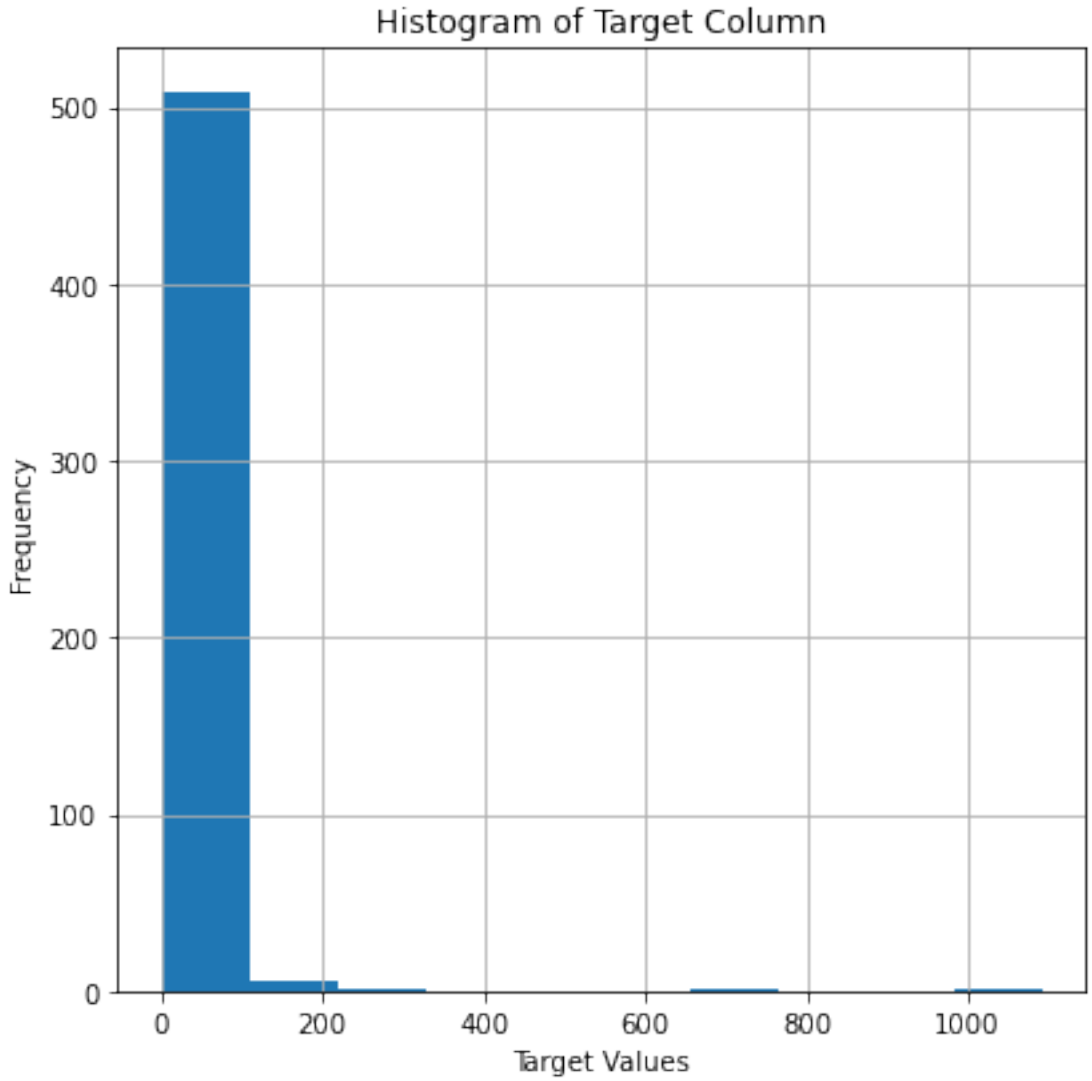

Figure 3.1: Histogram representation of Target column and Frequency.

Histograms for each column in the DataFrame using a specified layout

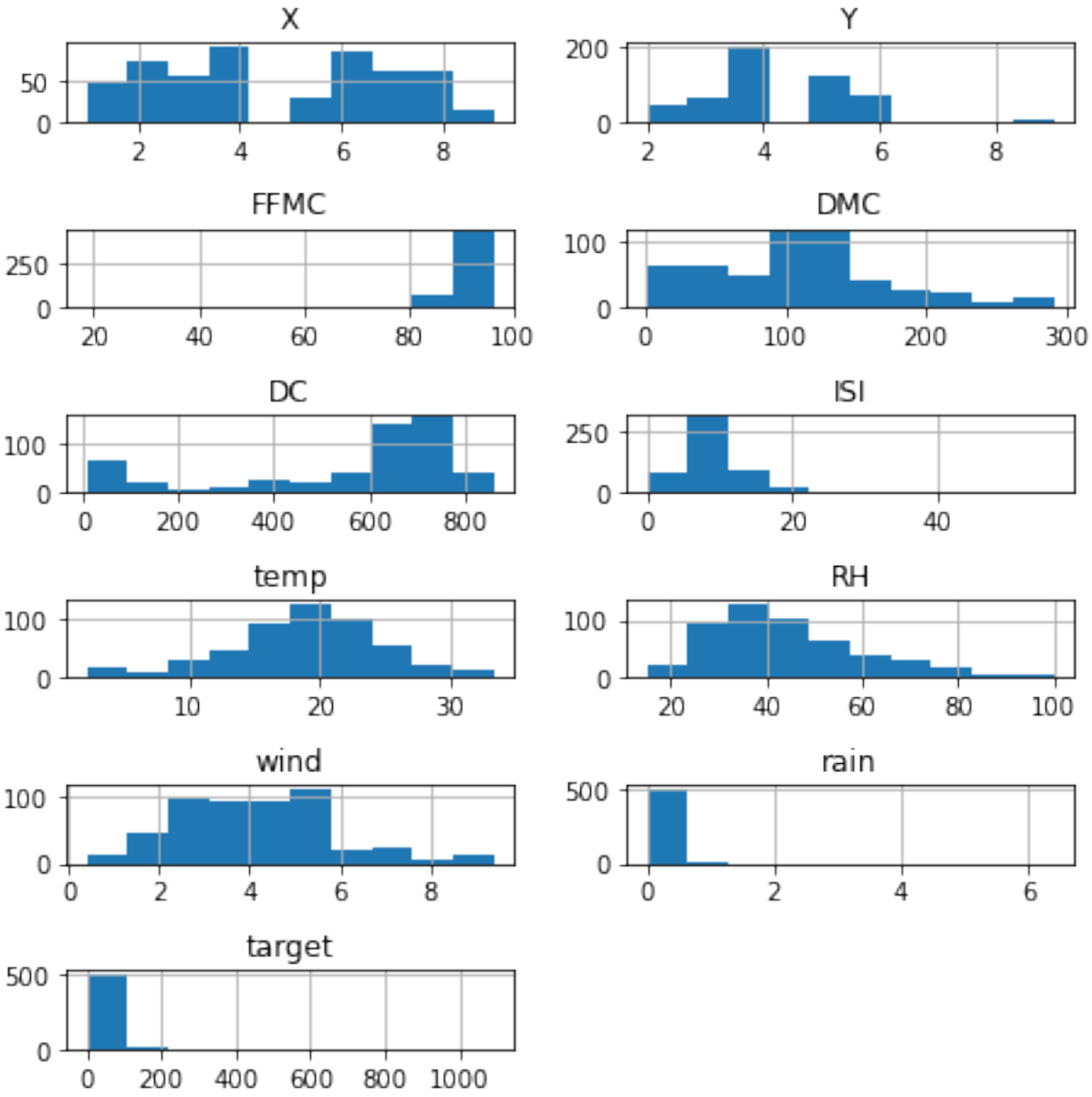

Figure 3.2: Histograms for each column in the DataFrame using a specified layout.

Bar plot showing the average target value for each month

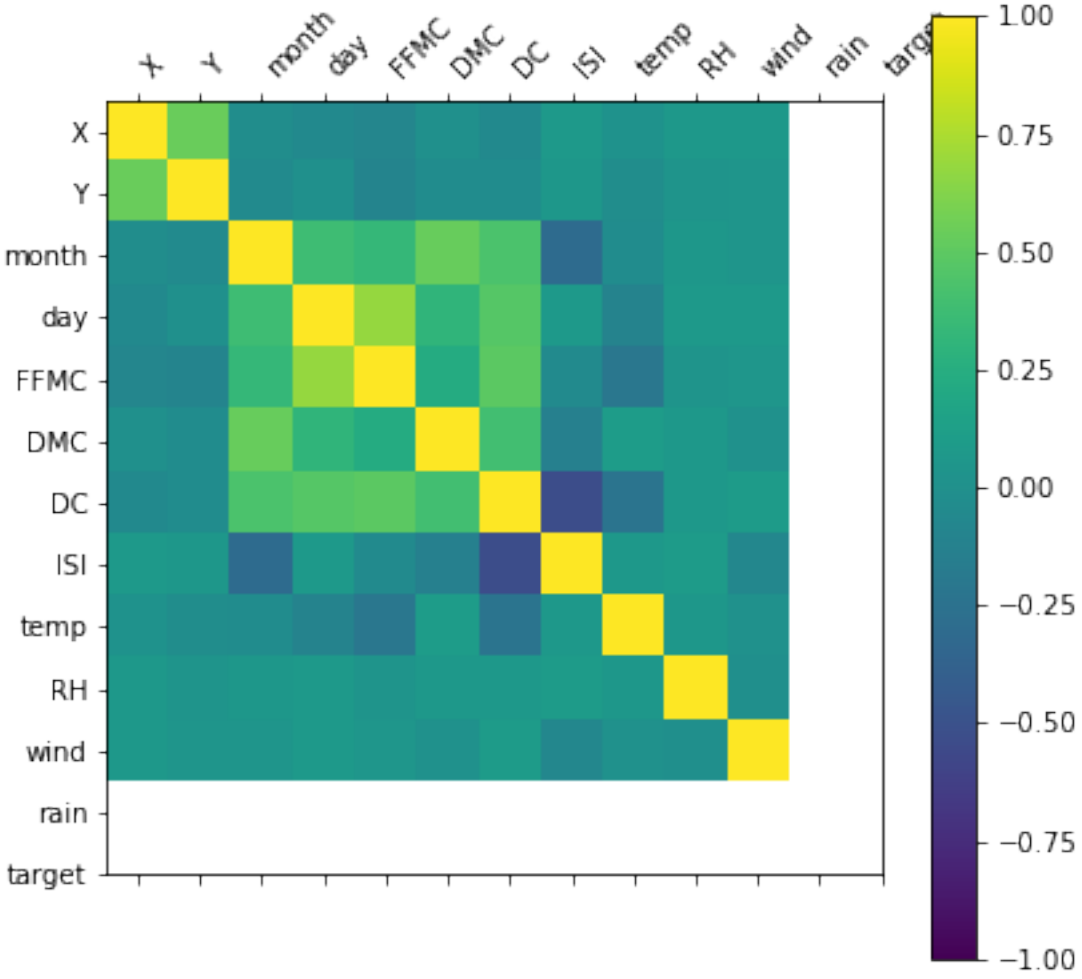

Figure 3.3: Bar plot showing the average target value for each month.

3.6 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.

3.7 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.

- 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.

- 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.

Chapter 4

Results

The results chapter tells a reader about your findings based on the methodology you have used to solve the investigated problem. For example:

- If your project aims to develop a software/web application, the results may be the developed software/system/performance of the system, etc., obtained using a relevant methodological approach in software engineering.
- If your project aims to implement an algorithm for its analysis, the results may be the performance of the algorithm obtained using a relevant experiment design.
- If your project aims to solve some problems/research questions over a collected dataset, the results may be the findings obtained using the applied tools/algorithms/etc.

Arrange your results and findings in a logical sequence.

4.1 A section

...

4.2 Example of a Table in LaTeX

Table 4.1 is an example of a table created using the package LaTeX"booktabs." do check the link: wikibooks.org/wiki/LaTeX/Tables for more details. A table should be clean and readable. Unnecessary horizontal lines and vertical lines in tables make them unreadable and messy. The example in Table 4.1 uses a minimum number of liens (only necessary ones). Make sure that the top rule and bottom rule (top and bottom horizontal lines) of a table are present.

Table 4.1: Example of a table in LaTeX

Bike		
Type	Color	Price (£)
Electric	black	700
Hybrid	blue	500
Road	blue	300
Mountain	red	300
Folding	black	500

4.3 Example of captions style

- The **caption of a Figure (artwork) goes below** the artwork (Figure/Graphics/illustration). See example artwork in Figure **??**.
- The **caption of a Table goes above** the table. See the example in Table 4.1.
- The **caption of an Algorithm goes above** the algorithm. See the example in Algorithm **??**.
- The **caption of a Listing goes below** the Listing (Code snippet). See example listing in Listing **??**.

4.4 Summary

Write a summary of this chapter.

Chapter 5

Discussion and Analysis

Depending on the type of project you are doing, this chapter can be merged with "Results" Chapter as " Results and Discussion" as suggested by your supervisor.

In the case of software development and the standalone applications, describe the significance of the obtained results/performance of the system.

5.1 A section

Discussion and analysis chapter evaluates and analyses the results. It interprets the obtained results.

5.2 Significance of the findings

In this chapter, you should also try to discuss the significance of the results and key findings, in order to enhance the reader's understanding of the investigated problem

5.3 Limitations

Discuss the key limitations and potential implications or improvements of the findings.

5.4 Summary

Write a summary of this chapter.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

Typically a conclusions chapter first summarizes the investigated problem and its aims and objectives. It summaries the critical/significant/major findings/results about the aims and objectives that have been obtained by applying the key methods/implementations/experiment set-ups. A conclusions chapter draws a picture/outline of your project's central and the most signification contributions and achievements.

A good conclusions summary could be approximately 300–500 words long, but this is just a recommendation.

A conclusions chapter followed by an abstract is the last things you write in your project report.

6.2 Future work

This section should refer to Chapter 4 where the author has reflected their criticality about their own solution. The future work is then sensibly proposed in this section.

Guidance on writing future work: While working on a project, you gain experience and learn the potential of your project and its future works. Discuss the future work of the project in technical terms. This has to be based on what has not been yet achieved in comparison to what you had initially planned and what you have learned from the project. Describe to a reader what future work(s) can be started from the things you have completed. This includes identifying what has not been achieved and what could be achieved.

A good future work summary could be approximately 300–500 words long, but this is just a recommendation.

Chapter 7

Reflection

Write a short paragraph on the substantial learning experience. This can include your decision-making approach in problem-solving.

Some hints: You obviously learned how to use different programming languages, write reports in LaTeXand use other technical tools. In this section, we are more interested in what you thought about the experience. Take some time to think and reflect on your individual project as an experience, rather than just a list of technical skills and knowledge. You may describe things you have learned from the research approach and strategy, the process of identifying and solving a problem, the process research inquiry, and the understanding of the impact of the project on your learning experience and future work.

Also think in terms of:

- what knowledge and skills you have developed
- what challenges you faced, but was not able to overcome
- what you could do this project differently if the same or similar problem would come
- rationalize the divisions from your initial planed aims and objectives.

A good reflective summary could be approximately 300–500 words long, but this is just a recommendation.

Note: The next chapter is "**References**," which will be automatically generated if you are using BibTeX referencing method. This template uses BibTeX referencing. Also, note that there is difference between "References" and "Bibliography." The list of "References" strictly only contain the list of articles, paper, and content you have cited (i.e., refereed) in the report. Whereas Bibliography is a list that contains the list of articles, paper, and content you have cited in the report plus the list of articles, paper, and content you have read in order to gain knowledge from. We recommend to use only the list of "References."

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'.

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)

Appendix A

An Appendix Chapter (Optional)

Some lengthy tables, codes, raw data, length proofs, etc. which are **very important but not essential part** of the project report goes into an Appendix. An appendix is something a reader would consult if he/she needs extra information and a more comprehensive understating of the report. Also, note that you should use one appendix for one idea.

An appendix is optional. If you feel you do not need to include an appendix in your report, avoid including it. Sometime including irrelevant and unnecessary materials in the Appendices may unreasonably increase the total number of pages in your report and distract the reader.

Appendix B

An Appendix Chapter (Optional)

...