# Forest Fire Prediction Using Heterogeneous Data Sources and Machine Learning Methods

by

Parveen Kaur

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2023

## Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

**Abstract**

Forest fires pose a significant and urgent threat to ecosystems and human lives, necessitating accurate prediction for effective mitigation strategies. Predicting forest fires has been a longstanding challenge due to the complex and dynamic nature of fire behavior. Traditional approaches to forest fire prediction, dating back to the 1950s, relied on simplistic statistical models and manual observations to identify fire-prone areas. However, these classical solutions were limited in their ability to capture the intricate interplay of various environmental factors that influence fire ignition. Since then, the field of forest fire prediction has undergone remarkable advancements, driven by the availability of heterogeneous data sources, advancements in computing power, and the emergence of machine learning techniques. The advent of remote sensing technologies, weather stations, and geospatial data has provided rich and diverse datasets for analyzing fire-related variables such as weather conditions, vegetation indices, topography, and historical fire records. Furthermore, the rapid progress in machine learning algorithms has enabled the development of sophisticated models capable of extracting meaningful patterns and relationships from these large-scale and complex datasets. These advancements have revolutionized forest fire prediction by improving the performance and reliability of predictive models, facilitating proactive decision-making, and enhancing the effectiveness of mitigation strategies.

Our study employs a comprehensive data collection framework to enhance forest fire prediction capabilities. The framework integrates data from remote sensing satellites, ground-based weather stations, and other relevant sources, facilitating the capture of crucial meteorological, biophysical, and topographical attributes. By leveraging these heterogeneous data sources, we create a unified database that spans a substantial 18-year period and offers a high temporal resolution for detailed analysis. However, one of the primary challenges encountered in forest fire prediction is the issue of data imbalance, where the number of non-fire instances significantly surpasses fire instances in the dataset. To address this challenge, advanced spatial subsampling, and downsampling techniques are employed, effectively mitigating the data imbalance issue and ensuring a more balanced representation of fire and non-fire instances for model training. Leveraging machine learning methods such as Random Forest, XGBoost, and MultiLayer Perceptron, our study evaluates the performance of these models in forest fire prediction. The results reveal the impressive performance of XGBoost, achieving an impressive ROC-AUC score of 87.2% and a sensitivity of 75%. This study highlights the importance of incorporating meteorological data and fire history to improve prediction performance and showcases the potential of machine learning techniques in addressing forest fire prediction challenges. The findings contribute to proactive risk assessment, robust mitigation strategies, and preserving ecosystems and human lives.

**Acknowledgements**

I would like to express my deepest gratitude to my thesis advisor, Dr. Sagar Naik, for their invaluable guidance, expertise, and unwavering support throughout this research journey.

I am profoundly grateful to my parents, brother, and loving husband for their unwavering love, encouragement, and support. Their guidance and sacrifices have shaped my path and fueled my determination to succeed.

I extend my thanks to my team members and colleagues for their collaboration, inspiration, and technical expertise, which have greatly contributed to the success of this research project.

I am also thankful to all the participants and organizations who generously shared their data and insights for this study, as their cooperation has been vital to the research.

Lastly, I am sincerely thankful to all individuals and institutions who have played a role, no matter how big or small, in shaping this research endeavor.

## Dedication

This thesis is dedicated to the love and support of my family, my husband, and to the guidance of a higher power.

# Table of Contents

**List of Figures** ix

**List of Tables** xi

**1 Introduction** 1

1.1 Motivation . . . 1

1.2 Objective . . . 3

1.2.1 Research Questions . . . 6

1.2.2 Contributions . . . 6

1.3 Outline of Thesis . . . 8

**2 Background and Literature Review** 9

2.1 Forest fires and Machine Learning . . . 9

2.1.1 Forest Fire Database . . . 10

2.1.2 Machine Learning Algorithms . . . 10

2.1.3 Imbalanced Data . . . 12

2.1.4 Evaluation Metrics for Imbalanced Data . . . 13

2.2 Notions . . . 15

2.2.1 Machine Learning . . . 15

2.2.2 Classification Problem . . . 16

2.2.3 Decision Trees . . . 17

2.2.4 Random Forest . . . 20
2.2.5 XGBoost . . . 22
2.2.6 Multilayer Perceptron . . . 23
2.3 Imbalanced Data . . . 25
2.3.1 Random Undersampling technique . . . 27
2.3.2 Near Miss Undersampling technique . . . 29
2.4 Performance Metric . . . 36
2.4.1 Issue with Prediction Accuracy as Performance Metric . . . 36
2.4.2 Performance Metric for Imbalanced data . . . 37

**3 Data Collection Framework and Application 41**
3.1 Framework for Data Collection . . . 42
3.1.1 Grid Cells . . . 43
3.1.2 Data Conversions . . . 44
3.2 Data Collection Framework for Alberta . . . 48
3.2.1 Provincial Boundary . . . 50
3.2.2 Alberta Grid cells . . . 50
3.2.3 Meteorological Data . . . 52
3.2.4 Copernicus Meteorological Data . . . 54
3.2.5 Biophysical Data . . . 56
3.2.6 Topographical Data . . . 59
3.2.7 Fire Data . . . 62
3.2.8 Data Integration and Combination . . . 66

**4 Methodology of Handling Data Imbalance and Prediction Model 68**
4.1 Data Imbalance Handling . . . 68
4.1.1 Changes in Spatial Resolution . . . 70
4.1.2 Data Spatio-subsampling . . . 70

4.1.3 Data Augmentation . . . 71
4.1.4 Impact of Spatio-subsampling and Data Augmentation . . . 74
4.2 Forest Fire Prediction Modelling . . . 75
4.2.1 Preprocessing . . . 75
4.2.2 Train Test Splitting . . . 77
4.2.3 Pipeline . . . 78
4.2.4 Cross Validation Folds . . . 80
4.2.5 Parameter Grid . . . 80
4.2.6 Fitting model in Grid Search . . . 81
4.3 Comprehensive Approach for Imbalance Handling . . . 82

**5 Experiments and Results** **84**
5.1 Comparison of various models performance . . . 84
5.2 Methodological Considerations . . . 87
5.2.1 Decide downsampling technique . . . 87
5.2.2 Decide downsampling ratio . . . 89
5.2.3 Importance of Downsampling . . . 91
5.3 Ablation Study: Importance of each feature . . . 92
5.4 Computation time analysis . . . 97
5.5 Challenges encountered during development . . . 98
5.5.1 Spatial Resolution Changes . . . 98
5.5.2 Imbalance Ratio of Data . . . 99
5.5.3 Large Volume of Data . . . 100
5.5.4 Availability of data . . . 100

**6 Conclusion & Future Work** **101**
6.1 Future Work . . . 101
6.2 Conclusion . . . 103

**References** **105**

# List of Figures


1.1 Research Objective . . . . . 3
1.2 Data features . . . . . 4

2.1 Decision Tree Components . . . . . 17
2.2 Random Forest Classifier . . . . . 20
2.3 eXtreme Gradient Boosting Classifier . . . . . 22
2.4 Multilayer Perceptron . . . . . 23
2.5 Imbalanced Data for Forest Fire Event . . . . . 25
2.6 Random Undersampling Method . . . . . 27
2.7 Near Miss Undersampling . . . . . 29
2.8 Near Miss version 1 Undersampling . . . . . 31
2.9 Near Miss version 2 Undersampling . . . . . 33
2.10 Near Miss version 3 Undersampling . . . . . 35
2.11 Confusion Matrix . . . . . 38
2.12 Receiver Operating Curve . . . . . 40

3.1 Data Collection Framework . . . . . 42
3.2 Data Collection Framework for Alberta . . . . . 47
3.3 Alberta 10 km by 10 km Grid Layer . . . . . 50
3.4 Alberta 10 km by 10 km Centroid Layer . . . . . 51
3.5 Meteorological data input . . . . . 52

3.6 Township ID joined with grid_id (met_out2) . . . 53
3.7 Goal format of Meteorological data (met_of) . . . 53
3.8 Copernicus Meteorological Data for each grid cell (cmet_out1) . . . 54
3.9 Final Copernicus Meteorological Data (cmet_of) . . . 55
3.10 Normalized Difference Vegetation Index Data for Alberta (ndvi_in) . . . 56
3.11 Mean NDVI Values per Grid Cell (ndvi_out1). . . . 57
3.12 Final NDVI data (ndvi_of) . . . 58
3.13 Digital Elevation Model of Alberta (topo_out1) . . . 59
3.14 Topographical data (topo_out2, topo_out3, topo_out4) . . . 60
3.15 Final Topographical data (topo_of) . . . 61
3.16 Fires in Alberta in last 18 years . . . 62
3.17 Fire data (fire_out1) . . . 63
3.18 Grid_id assigned fire data (fire_out2) . . . 64
3.19 Final fire data (fire_of) . . . 64
3.20 Raw Data . . . 66

4.1 Modelling workflow . . . 69
4.2 Data . . . 73
4.3 Spatio-subsampling and Data Augmentation . . . 74
4.4 Parameter Grid . . . 81

5.1 ROC Curve of Classifiers . . . 86
5.2 Performance vs Downsampling ratio from 0 to 1 . . . 89
5.3 Performance vs Downsampling ratio from 0.01 to 0.1 . . . 90
5.4 Ablation Study for Random Forest. . . . 93
5.5 Ablation Study for XGBoost. . . . 93

# List of Tables


1.1 Contributions and addressed Research Questions . . . . . 7

2.1 Machine Learning Model Literature . . . . . 11
2.2 Literature Overview Regarding Imbalanced Data . . . . . 13
2.3 Performance Metrics Literature Review . . . . . 14

3.1 Location-dependent features goal format . . . . . 45
3.2 Location and time-dependent features goal format . . . . . 45
3.3 Abbreviations . . . . . 48
3.4 Nomenclature of the input and output of data . . . . . 48
3.5 Data Features . . . . . 49
3.6 Data Sources . . . . . 49

4.1 Data . . . . . 74
4.2 Steps to Deal with Imbalance in Forest Fire Data . . . . . 82

5.1 Comparison of performance of different classifiers . . . . . 85
5.2 Different downsampling techniques on Random Forest . . . . . 87
5.3 Effect of undersampling on Random Forest prediction performance . . . . . 91
5.4 Effect of undersampling on XGBoost prediction performance . . . . . 91
5.5 Abbreviations . . . . . 94
5.6 Random Forest Feature Importance . . . . . 94

5.7 XGBoost Feature Importance . . . . . 95
5.8 CPU time for Random Forest and XGBoost . . . . . 97

# Chapter 1

# Introduction

## 1.1 Motivation

Forest fires are a natural part of many ecosystems, but they can also have significant impacts on the environment. There are a few long-term benefits of wildfires, like nutrient cycling and disease control, but these are outweighed by the negative impacts of wildfires that are significant in the short term. The trends of wildfires show a significant rise over the decades. In April 2020, the number of fire alerts across the globe was up by 13% compared to the previous year[24]. A greater number of more intense fires will release millions of extra tonnes of carbon, decimate biodiversity, destroy vital ecosystems, impact economies, and people, threaten property and livelihoods, and cause severe long-term health problems for millions around the world[24]. Copernicus Atmosphere Monitoring Service reported that global wildfires and vegetation fires in 2022 generated 1,455 mega-tonnes of carbon emissions[4]. The higher carbon emissions lead to high temperatures and climate change, which results in more frequent fires in drier conditions, as explained by Climate Feedback Loop Fueling US Fires[4]. Conference Board of Canada [19] estimates that wildfires cost the Canadian economy an average of $1 billion per year. In 2016 in Alberta, an uncontrolled wildfire in the Fort McMurray region led to the evacuation of over 88,000 residents and cost $456 million or 0.1 percent off of real GDP in Alberta[19].

The apocalyptic prospect makes it essential for the researchers, planners, and agencies responsible for fire management and mitigation to have practices for better forest fire warning and prediction systems. There are broadly two actions that are carried out to reduce the risk and mitigate wildfires[8]: (a) prediction of wildfire occurrence, detection of wildfire, predicting the spread of wildfire, identifying the potential danger areas for

wildfire; and (b) decide development, deployment, and financing of wildfire mitigation and elimination resources. Innovative wildfire warning tools and prediction models are required in order to improve wildfire management, moderation, and evacuation practices. Particularly, predicting the fire ignition - *the potential that a wildfire may break out in a certain area before its actual ignition* would offer useful mitigation and disaster planning capabilities. Prediction of forest fires is essential, in order to allocate resources efficiently, respond quickly to prevent fire activity and its spread, and have better wildfire emergency response. The identification of high fire risk regions can help us prevent wildfires by shutting down the electricity line in that region and introducing forest fire vegetation breaks in such regions to limit the spread of fires.

Wildfire ignition prediction can be done by modeling the relationship between the identified influential factors and fire danger. However, as pointed out by [9], forest fire ignition and behavior is a complicated process; it is the result of nonlinear and complex relations between various factors, such as ignition source, fuel content, climate, and topography. The complex nature of the problem makes it difficult for the predictive models to accurately predict the future occurrence of wildfire and help planners with reliable models to use when making crucial mitigation decisions[8]. The selection of appropriate factors involved in the wildfire occurrence and deciding the method to model the forest fire occurrence is important because these can have a big impact on the prediction performance[8].

The advent of various remote sensing technologies and machine learning approaches provides enormous opportunities to build models to predict wildfires. Therefore, we investigate the potential of data combined from various heterogeneous data sources and machine learning for forest fire prediction.

## 1.2 Objective

The objective of this work is to develop a model to predict forest fire ignition. *Prediction of forest fire ignition is formulated here as a classification problem, given the data features model classifies it into forest fire ignition (1) and non-ignition (0).* The objective of this thesis is represented in Fig. 1.1. We assume the availability of historical data from various sources such as weather stations, remote satellites, and other sources. We develop a framework to combine these datasets into a single database. These data variables come from different sources of data such as weather stations, and remote satellites, and different formats of data such as CSV files, GeoTIFF images, and shape files. These datasets from different sources usually have different spatial and temporal resolutions. To combine such datasets to generate a single database various spatial and temporal factors are considered.

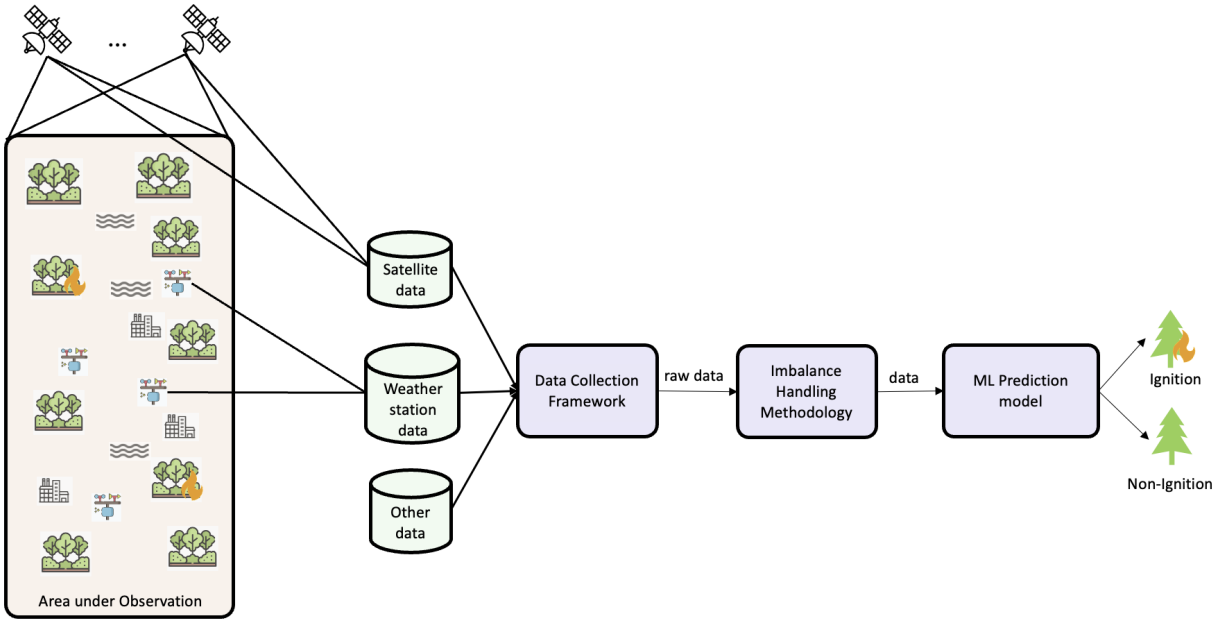



Figure 1.1: Research Objective

The Forest Fire dataset is highly imbalanced which means that the number of fire events is very less than the number of non-fire events. The issue with the highly imbalanced dataset is that model trained on such data tends to favor the majority class and totally ignore the minority class which is fire events in our scenario. So we need a methodology to deal with the imbalance in the Forest Fire Data. Prediction using ma-

chine learning techniques depends on various Forest Fire Danger Conditions (FFDC). The FFDC considered four different types of data: Meteorological variable (MV), Biophysical variable(BV), Topographical variable (TV), and Other variable (OV) as shown in Fig. 1.2. A high-level model of systems for the prediction of forest fire ignitions using FFDCs is denoted by MFFDC as given in Equ. 1.1.

$$MFFDC = f(MV, BV, TV, OV) \tag{1.1}$$

The output of the prediction model is Ignition (1) or Non-Ignition (0) given the input vector which has all the MV, BV, TV, and OV data variables. The Visual representation of the overall work of the study is summarised in 1.1. We investigate the potential in data combined from various resources for forest fire prediction using two tree-based ensemble models and one neural network-based model. For this purpose, we propose, a "Wildfire ignition prediction dataset" that combined various meteorological, topographical, and biophysical features. In addition, we address the crucial problem of imbalance in the data on forest fires.

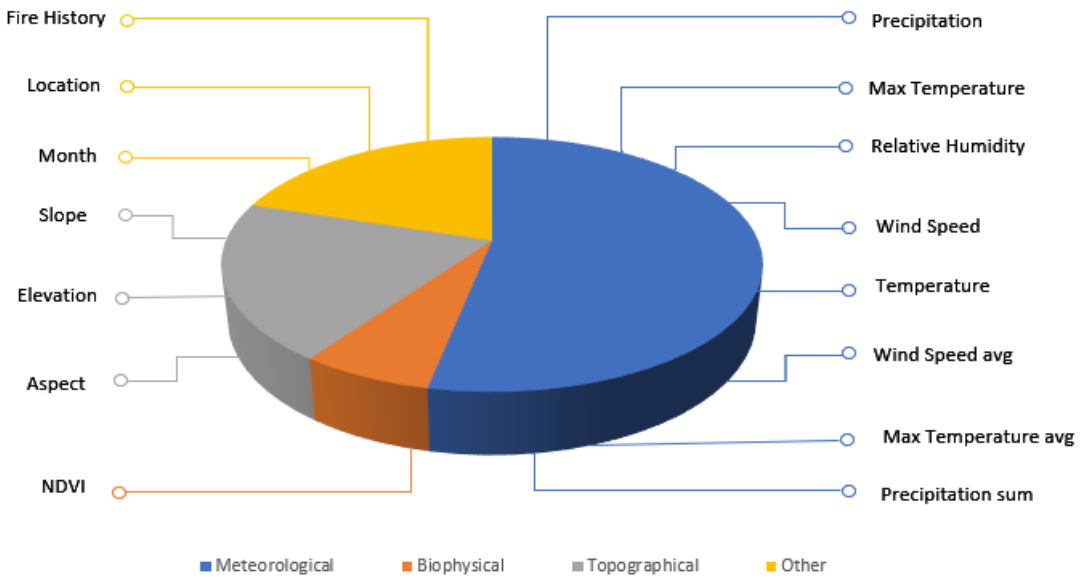



Figure 1.2: Data features

This study presents a novel framework aimed at integrating data from diverse sources into a consolidated database that encompasses various Forest Fire Danger Conditions (FFDC). The objective is to establish a comprehensive and unified repository of features collected from different sources, thereby facilitating in-depth research within the domain of forest fires. Meteorological data, derived from weather stations distributed across the Alberta region, is incorporated. Biophysical data, obtained through remote sensing technology, and topographical data, acquired from remote sensing satellites, are also included. This framework enables the seamless integration of data from multiple sources into a cohesive data file.

The ignition of forest fires is a multifaceted process involving numerous contributing factors. To capture the complexity of the phenomenon, this research integrates factors from different sources that describe various aspects of the prevailing conditions. By leveraging machine learning techniques, it becomes possible to discern intricate relationships between the FFDC and wildfires, thereby enabling the prediction of forest fire occurrences.

In addition to the existing features of Meteorological variables (MV), Topographical Variables (TV), and Biophysical Variables (BV), three additional FFDCs are incorporated into the dataset. The significance of their inclusion is analyzed to ascertain their contribution to the overall predictive capabilities. The location serves as a crucial factor, as certain regions may exhibit a higher risk of forest fires compared to others. Furthermore, considering the influence of seasonal variations, the month of occurrence is incorporated as an FFDC. The combination of month and location aims to enhance the model's performance by capturing inherent features that may not be explicitly represented in the dataset. Moreover, the study investigates the potential impact of incorporating historical fire data for a specific region on the model's ability to predict forest fires. This exploration recognizes the importance of wildfire frequency in a given region as an additional predictive factor. Finally, the research examines how weather trends in preceding days contribute to the occurrence of wildfires, further enhancing the predictive capabilities of the model.

### 1.2.1 Research Questions

RQ1. What are the underlying factors contributing to the ignition of forest fires?

RQ2. What methods can be employed to effectively integrate data from diverse formats and heterogenous data sources into a unified database?

RQ3. How does the combination of meteorological, biophysical, and topographical data enhance the predictive capabilities of the model for forest fire ignition?

RQ4. What are the major challenges associated with wildfire data and what strategies can be employed to address them?

RQ5. How to deal with the high imbalance in the forest fire data?

RQ6. What machine learning methods can effectively predict forest fire ignitions?

RQ7. How do various factors related to wildfires contribute to the prediction of forest fire ignitions?

RQ8. To what extent does the inclusion of fire history information, temporal information (e.g., month) and spatial information (e.g., grid location) improve the performance of forest fire ignition prediction?

### 1.2.2 Contributions

As highlighted earlier, this thesis makes significant contributions in three key areas pertaining to forest fire prediction: dataset aggregation, handling imbalanced data, and creation of a machine learning prediction model.

C1. Creation of a data collection framework for collecting data of different types and resolutions from various sources such as remote satellites and weather stations and aggregating it into a single database. Instantiate this framework by creating a wildfire prediction database with a total of 18 features, which is a combination of Meteorological, Biophysical, Topographical, and Other attributes captured daily over 18 years, with a spatial spanning across 661,848 $km^2$ at the spatial resolution of 10 $km$ by 10 $km$, presented in CSV format with each row representing a unique day and location. Thus framework can help us to create a single database from any number of sources.

C2. Proposed data imbalance handling technique for wildfire ignition prediction. The dataset generated from contribution C1 is highly imbalanced (observed imbalance ratio is 84000 non-fire is to 1 fire event), so we addressed the data imbalance using change of spatial resolution, spatio-subsampling, different downsampling techniques, and downsampling ratios. Handling the imbalanced nature of data leads to improved wildfire ignition predictability of the model.

C3. Creation of machine learning model for wildfire ignition prediction. We use three approaches (i) Random Forest (RF), (ii) eXtreme Gradient Boosting (XGB), and Multi-Layer Perceptron (MLP) neural network to check the reliability of machine learning models for wildfire ignition predictability. An ablation study showing the importance of each feature present in the data is explored.

The experimental results show that the ROC-AUC score of ensemble model XGBoost is the highest scoring at 87.2% among the three models compared. The best undersampling technique chosen for our dataset was the Near Miss 3 undersampling technique which gave out these best results at the downsampling ratio of 0.05. While comparing the contribution of each attribute to the correct prediction of fire and non-fire classes, meteorological data and fire history data show a big impact on the performance of the model.

Tab. 1.1 provides an overview of the research questions addressed by each contribution in the thesis. The table showcases how the three contributions correspond to different research questions discussed in various chapters. Contribution 1 (C1) addresses research questions RQ1, RQ2, providing insights and solutions related to this Data Collection domain. Contribution 2 (C2) focuses on RQ4 and RQ5, contributing valuable strategies to deal with dataset imbalance issues. Lastly, Contribution 3 (C3) encompasses RQ3, RQ6, RQ7, and RQ8, offering comprehensive analysis and outcomes in relation to these research questions about the predictive ability of the model and feature importance. Through these contributions, the thesis covers a wide range of research questions, providing a holistic understanding of the forest fire prediction domain.

| Contribution | Research Question |
|---|---|
| C1 | RQ1, RQ2 |
| C2 | RQ4, RQ5 |
| C3 | RQ3, RQ6, RQ7, RQ8 |

Table 1.1: Contributions and addressed Research Questions

## 1.3 Outline of Thesis

The remaining chapters of this thesis are structured as follows:

Chapter 2 provides a comprehensive overview of the background information related to machine learning and its application in the context of forest fire prediction. It explores relevant literature and introduces key concepts, terminologies, and techniques associated with machine learning models and classifiers employed in this research work.

Chapter 3 presents the framework developed for data collection from diverse sources and the subsequent aggregation of the collected data into a unified database. This chapter details the methodology used to integrate data from various heterogeneous sources, such as remote satellites and weather stations, into a single cohesive dataset for forest fire prediction.

Chapter 4 focuses on the data analysis and preparation phase, highlighting the significant changes made to the dataset. It also elucidates the architectural design of the machine learning classifier employed in this study, providing insights into the model's structure and configuration.

In Chapter 5, the results of the experimental evaluations are presented. This includes an examination of different machine learning models, downsampling techniques, and downsampling ratios on the dataset. Additionally, an ablation study is conducted to assess the influence of the inclusion of various columns of data on the performance of the predictive models.

Lastly, Chapter 6 serves as the conclusion of this research work, summarizing the key findings, contributions, and implications. Furthermore, future directions and potential areas for improvement and expansion of this research are discussed, highlighting avenues for further exploration and development in the field of forest fire prediction.

# Chapter 2

# Background and Literature Review

This chapter explores the fundamental aspects and literature review underlying our research on forest fire prediction using machine learning methodologies. The escalating threat posed by forest fires necessitates the development of accurate and timely prediction models to support proactive fire management and prevention strategies. To accomplish this objective, a comprehensive understanding of the existing body of knowledge in forest fire prediction and machine learning is essential. This chapter presents a systematic review of pertinent literature, examining the latest advancements and methodologies employed in forest fire prediction. Furthermore, it provides a detailed exposition of fundamental concepts and techniques in machine learning that will be pivotal in the development and optimization of our predictive models. By establishing a robust theoretical framework, we aim to facilitate the effective application of machine learning algorithms within the context of forest fire prediction. This research endeavor aims to contribute to the advancement of wildfire management and the preservation of our precious natural ecosystems.

## 2.1 Forest fires and Machine Learning

In recent years, the advent of machine learning techniques has shown great promise in revolutionizing the field of forest fire prediction. This literature review aims to critically examine the existing body of research on the application of machine learning algorithms for forest fire prediction, highlighting the Forest Fire Database, Machine Learning Algorithms, Imbalanced data issues, and Evaluation Metrics employed by different studies, and identifying key challenges and opportunities in this rapidly evolving domain.

### 2.1.1 Forest Fire Database

Extensive research efforts have been dedicated to identifying the primary drivers of forest fires, yielding valuable insights. As reviewed by [6], forest fire ignition emerges from a complex interaction among multiple factors, including meteorological conditions, topography, human activities, vegetation, and fuel types. Notably, meteorological features have been established as pivotal elements in forest fire prediction, particularly when integrated with a diverse array of variables, as demonstrated by [6] and [12].

Within the realm of meteorological features, Temperature, and Relative Humidity have been singled out as particularly influential in neural network-based wildfire prediction models [10]. Furthermore, researchers such as [6] and [20] have highlighted the significant role of topographical attributes, such as slope and aspect, in assessing wildfire risk. The Normalized Difference Vegetation Index (NDVI) has emerged as a critical factor for spatial forest fire prediction, employing artificial neural network-based models [1].

In our research, we extend beyond the dataset employed by [6] by incorporating additional dimensions, including topographical, temporal, and weather trends, to enhance our model's learning capabilities. The importance of these features has motivated us to integrate a diverse range of variables, spanning meteorological, topographical, biophysical, and fire history aspects, specifically tailored to the Alberta region. Moreover, our research incorporates fire history and weather trends to further enrich our dataset. The recognition of the need for high-quality datasets and a comprehensive data collection framework, encompassing various features, underscores our commitment to developing robust wildfire risk prediction models.

### 2.1.2 Machine Learning Algorithms

In recent years, machine learning techniques have gained significant attention in assessing the forest fire domain, as highlighted in the comprehensive review by [9]. Researchers have explored a range of models, including Multilayer Perceptron, Support Vector Machine, Backpropagation Neural Network, Logistic Regression, Random Forest, and XGBoost, to tackle this complex problem. Studies conducted by [12], [6], and [21] have demonstrated the superiority of neural network-based models over support vector machines for wildfire prediction. Notably, the Multilayer Perceptron model outperformed Logistic Regression when employed for forest fire prediction [6], prompting us to incorporate it into our research to assess its predictive reliability for our specific dataset.

| Author | Summary/Learning |
|---|---|
| [6], [21], [12] | The neural network-based model performs better than the support vector machine for wildfire prediction. |
| [6] | The multi-layer perceptron model performed best for forest fire prediction when compared to Logistic Regression. |
| [8] | The prediction capability of RF is better than SVM, so we include Random Forest as a model for the forest fire prediction |
| [7] | While predicting the forest fires in Turkey using the random forest, xgb regressor, Decision Tree, and Linear Regression showed that the Random forest gave the best results. |
| [25] | For forest fire risk prediction employed XGBoost, SVM, and Logistic Regression based solely on the meteorological data and concluded that XGBoost had the best predictive abilities. |

Table 2.1: Machine Learning Model Literature

In one study, [21] employed a neural network with a backpropagation algorithm on meteorological and weather index features to classify fire and non-fire cells. While our dataset is considerably larger compared to that of [21], enabling the development of more generalized models applicable to other regions as well, we recognize the importance of their approach and findings.

Furthermore, [8] demonstrated the superior prediction capabilities of Random Forest over Support Vector Machines, leading us to include Random Forest as one of the models for forest fire prediction. Similarly, [23] compared Random Forest, XGBoost, and Balanced Random Forest for wildfire risk prediction and found promising results. In the context of forest fire detection and prediction, [22] successfully employed Random Forest. However, our dataset is substantially larger, and we address the challenge of imbalanced forest fire data, enhancing the applicability of our results.

Moreover, [7] examined forest fire prediction in Turkey using various models such as Random Forest, XGBoost Regressor, Decision Tree, and Linear Regression, with Random Forest yielding the best performance in terms of metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Accuracy. Drawing inspiration from these findings, we incorporate Multi-Layer Perceptron, Random Forest, and XGBoost into our work, leveraging their demonstrated effectiveness in forest fire prediction tasks.

Additionally, for forest fire risk prediction, [25] explored XGBoost, SVM, and Logistic Regression solely based on meteorological data, concluding that XGBoost exhibited the highest predictive abilities. In our research, we extend beyond this approach by incorpo-

rating a broader range of data factors, augmenting the predictive power of XGBoost. The collective insights and outcomes of these notable studies as shown in Tab. 2.1, provide the foundation and motivation for our selection of Multi-Layer Perceptron, Random Forest, and XGBoost as integral components of our research methodology.

### 2.1.3 Imbalanced Data

During the analysis of research papers focused on machine learning for forest fire prediction, a significant research gap has been identified regarding the recognition of imbalanced data as a fundamental issue within this domain. The prevailing approach to data collection often neglects the inherent data imbalance associated with forest fire occurrences. Several studies [12, 1, 8] have been observed to adopt a methodology in which an equal number of non-fire points are randomly selected alongside fire points, resulting in a dataset composition characterized by a 1:1 ratio. However, it is crucial to address the imbalanced nature of forest fire data in this research context. The composition of the dataset has a substantial impact on the performance of predictive models, making the obtained results unsuitable for practical deployment scenarios where the imbalance ratio between non-fire and fire points is substantially higher. The comprehensive overview of the studies regarding the imbalance issue is depicted in Tab. 2.2. Refer to Section 2.3 for more understanding of Imbalanced data and the Imbalance ratio is defined in Eq. 2.1

In the realm of wildfire prediction, a study conducted by [11] identifies the challenge of data imbalance, particularly concerning wildfire occurrences. They employ a deep learning-based model to forecast the extent of wildfire areas, focusing on the disparity between large-scale and small-scale forest fires. Conversely, in [6], a different forest fire prediction model is constructed, but with a predefined imbalance ratio of 1.4:1. Non-fire points are randomly selected without adequately addressing the underlying data imbalance. Similarly, researchers in [23] and [26] acknowledge the issue of dataset imbalance, yet employ relatively low predefined imbalance ratios of 10:1 and 3:1, respectively.

In contrast, our approach entails comprehensive data collection spanning 18 years, allowing experimentation across various imbalance ratios to select non-fire points that yield optimal results. As outlined in our Methodology section, the observed imbalance ratio in real-world scenarios is notably high. Furthermore, while [5] addresses the imbalance issue using random oversampling techniques for the most represented class, our analysis focuses on undersampling techniques, specifically exploring three versions of the Near Miss undersampling technique.

| Paper | Imbalance addressed | Imbalance ratio | Time Range | Geographical area | Performance Metric |
|---|---|---|---|---|---|
| [26] | No | 3:1 | 1 year | Indonesia | AUC |
| [23] | No | 10:1 | 20 years | Heilongjiang | accuracy, precision, recall, auc, fscore |
| [6] | No | 1.4:1 | 4 years | Jiangxi Province | Accuracy,AUC |
| [12] | No | 1:1 | 8 years | Guangxi Zhuang | accuracy, precision, recall, and f1 value |
| [10] | No | 1:1 | 1 year | North Lebanon | precision, sensitivity, specificity, accuracy, ROC AUC |
| [20] | No | - | 5 years | Kuala Selangor | Accuracy |
| [8] | No | - | 6 years | Dayu County, China | ROC AUC |

Table 2.2: Literature Overview Regarding Imbalanced Data

Compared to these studies, we collect all the data and compare different data sampling techniques, choosing the imbalance ratio that gives us the best predictive ability for our model. The tests are performed on the original unsampled data, which gives stakeholders confidence in the model's prediction ability. This is very important from the stakeholders' point of view, as the models needed by planners and agencies should be able to predict wildfire occurrence on the original data, which usually has high imbalance. In such cases, randomly selecting the non-fire data points may not accurately represent the original data, and therefore, the models may not provide an accurate representation of the problem.

### 2.1.4 Evaluation Metrics for Imbalanced Data

Performance metrics play a crucial role in evaluating the effectiveness of machine learning models on imbalanced data, as demonstrated in studies by [26], [5], and [23]. In the context of forest fire prediction, specific performance metrics hold significant importance, such as sensitivity. This metric measures the accurate classification of fire events by the model,

relative to the total number of fire events. A high sensitivity score is valuable for planners as it helps determine the necessary budget and resource allocation for mitigation purposes. Additionally, identifying areas where fires are unlikely to occur is equally important.

| Performance Metric | Paper | Significance/Limitation |
|---|---|---|
| Accuracy | [26], [23], [6], [12], [10], [20], [1], [11] | Overall correctness measure, but will mislead in case of imbalanced data; may appear high even with low minority class detection. |
| Precision | [23], [12], [10] | Measures accuracy of fire predictions in imbalanced data; indicates the proportion of correctly predicted fire instances. |
| Sensitivity | [23], [12], [10] | Measures correct prediction of fire instances; essential for detecting actual fire occurrences. |
| Specificity | [10] | Measures correct prediction of non-fire instances; crucial for identifying areas without fire. |
| ROC AUC | [23], [8], [10], [26], [5], [11] | Measures the model's ability to distinguish fire and non-fire classes; evaluates the overall performance across different thresholds. |
| F-score | [12], [23] | Harmonic mean of precision and recall; but may incline towards majority class |

Table 2.3: Performance Metrics Literature Review

Increasing the sensitivity of the model may lead to an increase in false positive rates, thereby potentially decreasing the model's specificity. Our objective is to achieve a high sensitivity while maintaining a reasonable level of specificity. Studies conducted by [26] and [5], which address the issue of imbalanced data, utilize the Area Under the Curve (AUC) as a performance metric. In our research, in addition to AUC, we employ other metrics such as sensitivity and specificity to provide a clearer representation of the model's performance for both fire and non-fire classes.

In contrast, [23] adopts a combination of various performance metrics including accuracy, precision, recall, AUC, and F-score. However, we refrain from using accuracy as it does not accurately reflect the predictive ability of the model, particularly for the minority class, which corresponds to fire cells in our case. By employing a comprehensive set of performance metrics tailored to the imbalanced nature of the data, our evaluation framework ensures a more accurate assessment of the model's predictive capabilities. Tab. 2.3 shows the importance of each performance metrics in case of imbalanced data and the research works that used them.

## 2.2 Notions

This section provides an essential foundation for understanding the machine learning techniques and concepts employed in this research. This section introduces and describes key components, including Random Forest (RF) and XGBoost (XGB), Multilayer Perceptron which are popular and effective algorithms for predictive modeling. Moreover, it explores the challenge of imbalanced data and the implications it poses for model performance. To address this issue, undersampling techniques such as Random Undersampling and Near Miss Undersampling are introduced as effective approaches for balancing the dataset. Additionally, performance metrics are discussed, highlighting their significance in evaluating the performance and effectiveness of the developed models. Through a comprehensive exploration of these notions, this section establishes a solid understanding of the fundamental components and methodologies utilized in the subsequent analysis and experimentation.

### 2.2.1 Machine Learning

Machine learning refers to the ability of a machine to learn from data and make decisions or predictions without explicit programming. It involves the development of statistical methods and algorithms that enable computers to improve their performance on specific tasks by identifying patterns in data.

There are three main categories of machine learning algorithms: supervised, unsupervised, and reinforcement learning. These categories differ based on the availability of the output variable during the training process. The distinctions can be understood as follows:

- In supervised learning, the training dataset contains the output feature that the model needs to learn. The model learns to map input variables to the provided output.
- Unsupervised learning involves datasets where no output feature is available. Instead, the model learns patterns and structures inherent in the data.
- Reinforcement learning is based on actions and rewards. The model learns by performing actions and receiving feedback or rewards, aiming to maximize cumulative rewards.

This research focuses on the supervised learning technique, which can be further categorized into classification and regression. The main difference lies in the type of output variable the model predicts.

- Classification models learn to predict categorical variables based on input features. For example, an image classification model could classify whether an image depicts a dog or a cat.
- Regression models, on the other hand, learn to predict continuous variables based on input features. For instance, a regression model might predict the stock price for the next day.

### 2.2.2 Classification Problem

The research problem we are addressing involves the prediction of forest fires, which is classified as a classification problem. Our objective is to predict whether a given location on a specific day will experience a fire based on input features. This classification task involves categorizing input features into two classes: "fire" and "non-fire" cells. The classification task can be further categorized as follows:

- Binary Classification: This task aims to assign data points to one of two potential classes.
- Multi-class Classification: In this task, data points are assigned to one of more than two potential classes, with multiple classes serving as the output feature for model training.
- Multi-label Classification: The objective of multi-label classification is to assign data points to one or more classes. The output feature contains multiple classes, and the model assigns two or more class labels to input data points.
- Imbalanced Classification: This type of classification problem deals with imbalanced data, where the samples of one class are significantly fewer (referred to as the minority class) than the samples of another class (referred to as the majority class). Most imbalanced classification tasks are binary classifications.

Specifically, our research focuses on the problem of imbalanced binary classification, which is further explained in section 2.3 of this chapter. In previous studies, several classification techniques have shown promising results for forest fire prediction, and we concentrate on three techniques that have achieved good performance in this context.

### 2.2.3 Decision Trees

A "decision tree" is a tree-like structure that represents a series of decisions based on certain features. The data is recursively split at decision nodes, creating multiple branches. The tree consists of two types of nodes: decision nodes and leaf nodes. Decision nodes are where the data is divided into sub-nodes, while leaf nodes represent the final outcomes or decisions.

The structure of a decision tree shown in Fig. 2.1 includes various components, such as the root node, decision nodes, leaf nodes, subtrees, and branches. The root node is the starting point of the decision tree and represents the entire dataset. It is recursively divided into nodes that represent different partitions of the dataset. Decision nodes are the points where decisions are made and further divided into sub-nodes. Leaf nodes cannot be further divided and provide the final results of the decision tree.

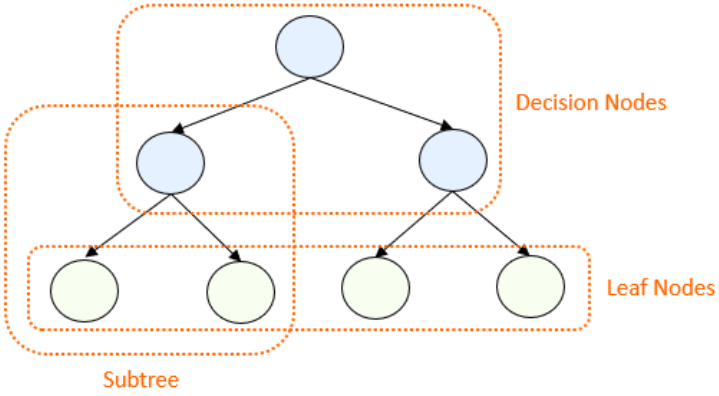



Figure 2.1: Decision Tree Components

The process of splitting involves dividing a node into sub-nodes based on specific conditions of selected features, effectively partitioning the data. Each resulting partition forms a branch or subtree of the decision tree. Pruning is a technique used to determine the optimal length of the decision tree, addressing the problem of overfitting. It involves removing or collapsing branches or nodes that do not significantly contribute to improving the predictive ability of the model.

**Construction of Decision Tree**

When constructing a decision tree, the selection of the splitting feature at each node is a crucial step. An attribute selection method is utilized to determine the most appropriate

feature. This process is carried out recursively until certain stopping criteria are met to form the complete tree. The choice of attribute selection method depends on whether the target variable is categorical or continuous. In the case of categorical classification, the Gini Index or Information methods are commonly used to identify the optimal feature for splitting at each node.

At each node, the feature that yields the highest value of the selected metric is chosen to perform the split. The objective is to minimize randomness within the resulting subsets and create more homogeneous subsets. In our implementation, the Gini Index is employed to split each node. This index quantifies the likelihood that a randomly selected sample would be misclassified if its label were assigned randomly based on the class distribution within the subset. The Gini Index ranges from 0 to 1, where a value of 0 represents a perfectly pure dataset (all data points belong to the same category) and a value of 1 denotes a perfectly impure dataset (data points are evenly distributed across all categories).

*Gini index = 1 - (sum of squares of probabilities of each class in the subset)*

To determine the best splitting feature at each node, the Gini Index is calculated for all available features. The feature that results in the greatest reduction in the Gini Index is selected for the node split. The calculation of the Gini Index involves summing the squares of the probabilities of each class in the subset. These probabilities are calculated as the ratios of the number of samples in each class to the total number of samples in the subset.

### Problem with Decision Trees

When decision trees are allowed to grow without any limitations, they often encounter a common issue known as overfitting. This occurs when the decision tree creates a leaf node for each individual row of the training data, resulting in the model achieving 100% accuracy on the training set. However, such a model tends to learn not only the relevant patterns but also the noise and irrelevant information present in the data. As a consequence, the model performs poorly when applied to unseen test data, indicating a lack of generalization ability.

To address the problem of overfitting in decision trees, various techniques have been developed. One approach we are particularly interested in is ensemble learning methods. These methods involve combining multiple decision trees to create a more robust and accurate model. The subsequent section delves into the discussion of the Random Forest Classifier, which is motivated by the need to mitigate overfitting issues associated with individual decision trees.

### Ensemble Learning Methods

A single model may not yield satisfactory results due to its limited predictive power. To overcome this limitation, Ensemble Learning methods have emerged as effective techniques in machine learning. These methods involve the combination of multiple weak models to create a single strong model with enhanced predictive capabilities and improved generalization performance on unseen data. The fundamental concept underlying Ensemble Learning is the amalgamation of diverse models to leverage their collective intelligence.

There are several approaches to combining models within the Ensemble Learning framework. One such approach is **Bagging**, which entails training multiple instances of the same model in parallel on different partitions of the data. In the context of classification, the final prediction is determined through majority voting based on the individual model outputs. This method is particularly effective when combining weak models characterized by low bias but high variance. By aggregating their predictions, a single model with reduced bias and variance can be obtained.

Another prominent Ensemble Learning method is **Boosting**. In contrast to Bagging, Boosting involves sequential training of multiple instances of the same model. Each subsequent model focuses on learning from the mistakes of its predecessors, placing a higher emphasis on correcting the errors made by the previous models. The iterative nature of Boosting aims to create a collective model that progressively reduces bias. By assigning greater weightage to the incorrect predictions of the preceding models, the subsequent models strive to rectify and improve upon those errors.

Through these Ensemble Learning techniques, we can harness the diversity and complementary strengths of multiple models to enhance the overall predictive performance and robustness of the resulting ensemble model.

### 2.2.4 Random Forest

The Random Forest algorithm serves as an extension of ensemble methods for decision trees, aiming to enhance their predictive capabilities. It consists of a collection of uncorrelated decision trees created on subsets of the data. In the context of classification tasks, the final prediction is determined through majority voting on the class labels generated by individual trees. The underlying principle behind Random Forest is the notion of "the wisdom of crowds," where the collective decision-making of multiple trees outperforms the performance of individual trees. Even if certain trees within the forest are incorrect, the consensus of the ensemble leads to accurate overall predictions.

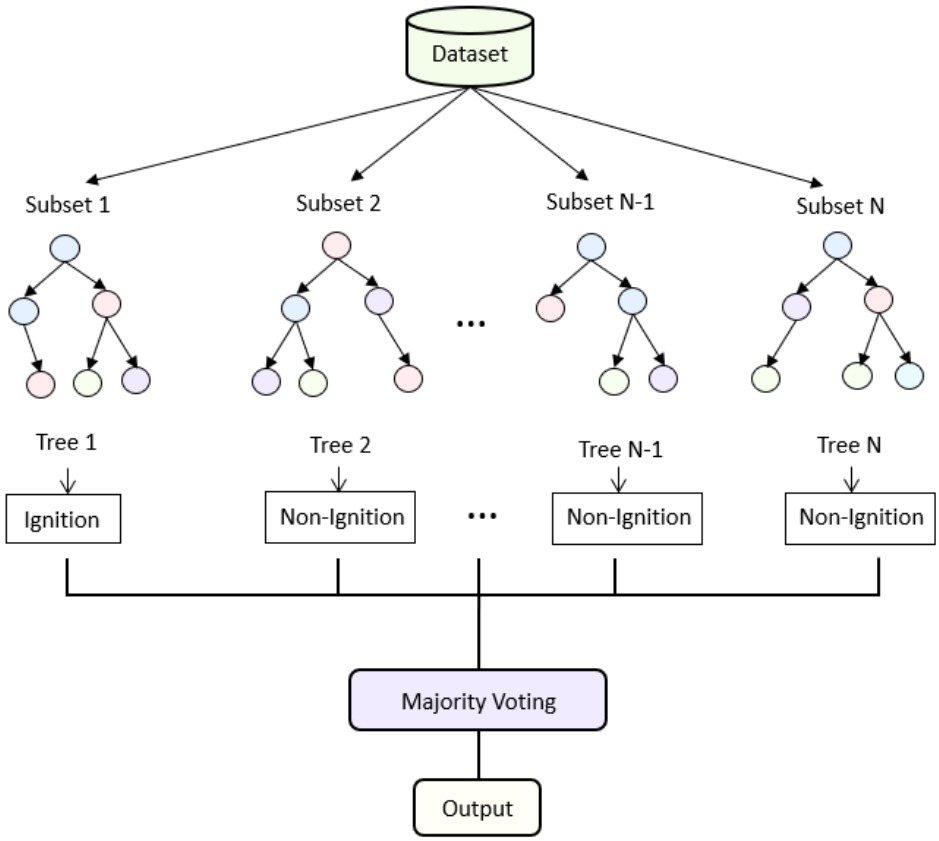



Figure 2.2: Random Forest Classifier

The structure of the Random forest is depicted in Fig. 2.2. The superior performance of Random Forest compared to individual decision trees can be attributed to two key factors:

- Each individual tree contains valuable information: If each individual tree within the Random Forest is a strong classifier, the ensemble model will exhibit lower error rates and better overall performance.
- Reduced correlation among the trees: The success of Random Forest hinges on minimizing the correlation between the trees in the ensemble. As correlation decreases, the error of the Random Forest decreases, leading to higher performance. This is achieved through two key techniques: bagging and feature randomness

Bagging is employed to create multiple subsets of the data, ensuring diversity within the ensemble. By randomly sampling the data with replacement, subsets of the same size are generated. For example, if the training data is 10, 20, 30, 40, 50, one of the subsets could be 10, 10, 30, 50, 50. Training individual trees on different subsets decreases the correlation among the trees, enhancing the overall performance of the Random Forest.

Feature randomness is another crucial aspect of Random Forest. It involves selecting a subset of features for each decision tree. If the original data contains F features, a feature subset of size f (where $f < F$) is chosen for each tree. Importantly, this feature subset remains constant throughout the construction of the Random Forest. By training each tree on a different subset of features, the algorithm ensures greater diversity and reduced correlation among the trees.

The size of the feature sample, denoted by f, plays a significant role in the model's performance. Decreasing f reduces both the correlation and strength of the trees while increasing f increases both the correlation and strength. Hence, finding an optimal value of f is crucial, striking a balance between uncorrelated trees and strong classifiers.

Once the Random Forest is built, the model is trained on the data, and predictions are made using majority voting based on the class labels generated by the ensemble. Notably, there are several hyperparameters that can be fine-tuned to improve the model's performance, including the node size, the number of trees in the forest, and the number of features sampled.

### 2.2.5 XGBoost

XGBoost, also known as eXtreme Gradient Boosting, is a powerful machine learning algorithm widely used for classification and regression tasks. The structure of the XGBoost is shown in Fig. 2.3. It builds a predictive model by combining multiple weak learners, such as decision trees, in a sequential manner. The key idea behind XGBoost is to iteratively train these weak learners and improve their performance over time.

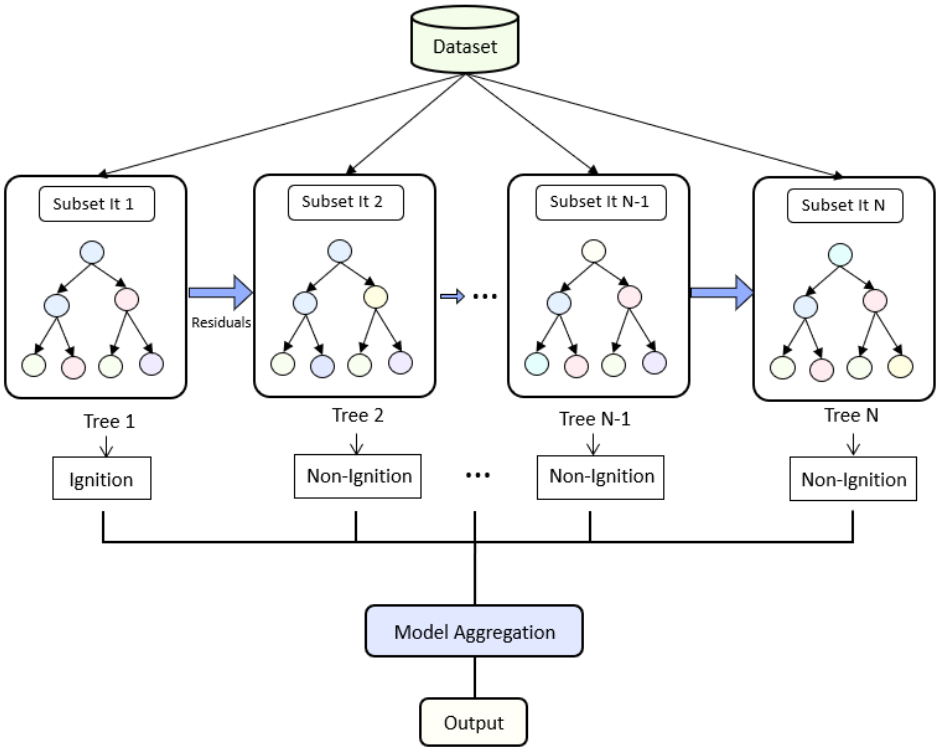



Figure 2.3: eXtreme Gradient Boosting Classifier

The construction of an XGBoost model involves several steps. It starts with an initial prediction, typically the mean value of the target variable. Then, the algorithm calculates the residuals, which are the differences between the actual and predicted values. It trains a weak learner, such as a decision tree, to predict these residuals. The process is repeated iteratively, with each new weak learner aiming to minimize the error between the predicted and actual values.

To enhance its performance, XGBoost incorporates regularization techniques. These

techniques prevent overfitting by reducing the impact of each individual learner and introducing diversity through the subsampling of features. XGBoost also utilizes gradient-based optimization, adjusting the parameters of each weak learner to minimize the loss function. Additionally, it employs an efficient data structure called the weighted quantile sketch to speed up the computation of splitting points during tree construction.

One of the notable features of XGBoost is its ability to handle missing values in the input data. It automatically learns the best direction to assign missing values during the model construction process, making it robust in the presence of incomplete data.

Overall, XGBoost is a powerful algorithm that combines the strengths of multiple weak learners to create a strong predictive model. Its regularization techniques, gradient-based optimization, efficient data structures, and handling of missing values contribute to its accuracy and versatility.

### 2.2.6 Multilayer Perceptron

The Multi-Layer Perceptron (MLP) is a popular and widely used neural network architecture for supervised learning tasks. It consists of multiple layers of interconnected artificial neurons, mimicking the structure and functionality of the human brain.

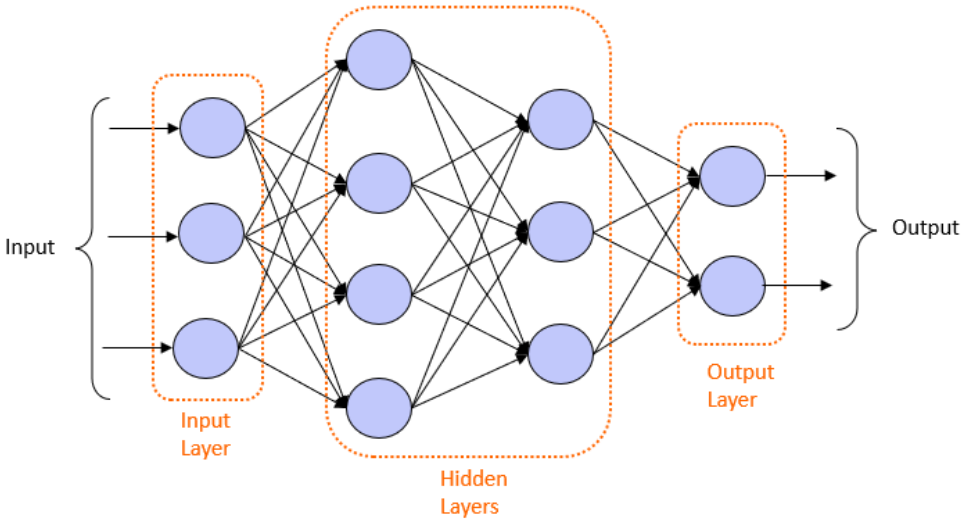



Figure 2.4: Multilayer Perceptron

The MLP is composed of an input layer, one or more hidden layers, and an output

layer as depicted in Fig. 2.4. Each neuron in a layer is connected to all the neurons in the adjacent layers. The input layer receives the features of the input data, and the output layer produces the predicted output or classification.

The key idea behind the MLP is to learn complex patterns and relationships in the data by adjusting the weights and biases of the neurons through a process called backpropagation. During training, the network is presented with labeled examples, and it makes predictions based on the current weights and biases. The prediction error is then calculated, and the weights and biases are adjusted in a way that minimizes the error. This iterative process continues until the network reaches a satisfactory level of accuracy.

One of the strengths of MLP is its ability to learn non-linear relationships between the input and output. The hidden layers, with their activation functions, introduce non-linear transformations that enable the network to capture intricate patterns in the data. This flexibility makes MLP suitable for a wide range of complex tasks, including image recognition, natural language processing, and time series analysis.

To train an MLP effectively, it requires a large amount of labeled training data. Additionally, careful considerations must be given to the architecture design, such as the number of hidden layers, the number of neurons in each layer, and the choice of activation functions. These design choices can greatly impact the learning capacity and performance of the MLP.

In conclusion, the Multi-Layer Perceptron is a versatile neural network architecture that excels at learning complex patterns and relationships in data. Its layered structure, backpropagation algorithm, and non-linear activation functions enable it to handle various challenging tasks. With appropriate design and sufficient training data, MLP can achieve high accuracy and generalization capabilities.

## 2.3 Imbalanced Data

Imbalanced data refers to a dataset where one class label is significantly more prevalent than the other, resulting in an imbalanced classification problem. The majority class represents the class with a higher number of instances, while the minority class corresponds to the class with a lower number of instances. The ratio between the number of instances in the minority class and the majority class is defined as the imbalance ratio.

$$\text{Imbalance ratio} = \frac{\text{number of non-fire points}}{\text{number of fire points}} \tag{2.1}$$

Equ. 2.1 quantifies the imbalance ratio, which is calculated as the number of non-fire data points divided by the number of fire data points. This ratio provides a measure of the class imbalance in our forest fire ignition prediction problem, where fire occurrences are rare events compared to non-fire instances.

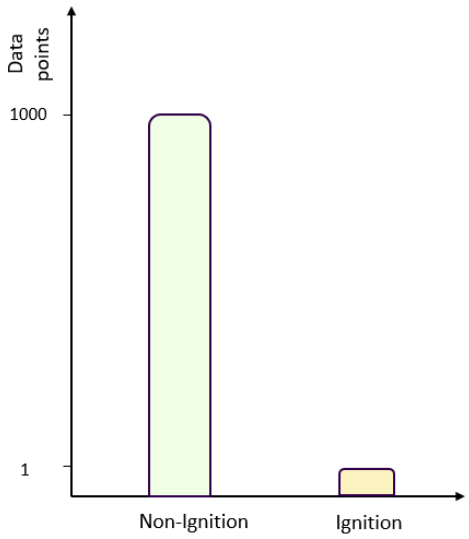



Figure 2.5: Imbalanced Data for Forest Fire Event

Fig. 2.5 provides a visual representation of highly imbalanced data, showcasing the unequal distribution of class labels. It serves as an illustration to highlight the challenge posed by imbalanced data in traditional classification models. By visually depicting the disparity between the majority and minority classes, the figure emphasizes the imbalance issue and its impact on classification performance, particularly for the minority class.

When training traditional classification models on imbalanced data, it is assumed that the class distribution is unbiased. However, these models tend to favor the majority class, leading to suboptimal classification performance for the minority class. This presents a challenge in forest fire ignition prediction, as accurately predicting the minority class (fire ignitions) is of utmost importance.

In our research, we encounter the imbalance issue as the number of fire ignitions is much lower than the number of non-ignition instances. Consequently, the minority class consists of fire ignitions, while the majority class comprises non-ignition cases. Our primary objective is to develop machine learning models capable of accurately predicting the minority class despite the class imbalance.

To mitigate this issue, various resampling techniques are employed to adjust the class distribution in the training data, aiming for a more balanced representation. Two commonly used resampling methods are oversampling and undersampling.

**Oversampling:** As the name suggests, oversampling involves increasing the number of samples in the minority class while keeping the majority class samples the same. This approach aims to create a new dataset with a higher number of minority samples, thus addressing the class imbalance.

**Undersampling:** On the other hand, undersampling reduces the number of samples in the majority class, giving the minority class a higher representation. There are several techniques for undersampling, such as random undersampling and near miss undersampling, which are further explained in Sections 2.3.1 and 2.3.2 of this thesis.

$$\text{Downsampling ratio} = \frac{\text{num of fire points}}{\text{num of non-fire points}} \tag{2.2}$$

The downsampling ratio, denoted in Eq. 2.2, is a predetermined value that is specified prior to the application of the undersampling technique. It serves as a crucial parameter for the undersampling method, providing guidance on the proportion of non-fire data points to be retained. The downsampling ratio is calculated as the ratio of the number of fire data points to the number of non-fire data points, representing the inverse of the imbalance ratio. By controlling the downsampling ratio, we can effectively adjust the class distribution and create a more balanced representation of the data for training our predictive models. It should be noted that it is the inverse of the Imbalance ratio which is given in Eq. 2.1.

### 2.3.1 Random Undersampling technique

Random Undersampling, as its name implies, reduces the samples of the majority class by randomly removing instances. The samples in the minority class remain unchanged. This technique allows us to achieve a target imbalance ratio by determining the desired proportion of the minority class and subsequently dropping a corresponding number of samples from the majority class.

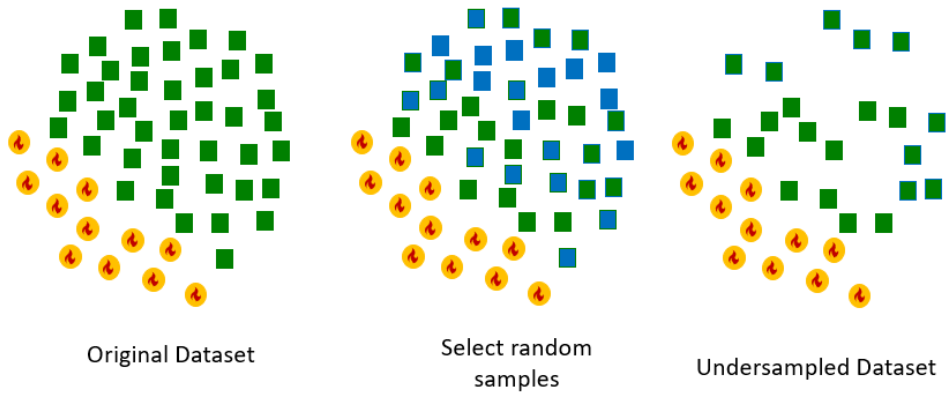



Figure 2.6: Random Undersampling Method

To illustrate the process of Random Undersampling and its impact on the dataset, Fig. 2.6 provides a visual representation. The figure consists of three subfigures, each depicting a different stage of the undersampling technique.

The first subfigure represents the original dataset, where both fire and non-fire samples are present. The majority class (non-fire) is visually more prominent due to its higher number of instances, while the minority class (fire) is represented by a smaller number of points.

In the second subfigure, random samples from the majority class are selected to be retained, while the minority class remains unchanged. This represents the application of Random Undersampling, where a portion of the majority class is preserved to create a more balanced dataset. The imbalance ratio is adjusted by selectively keeping samples from the majority class.

Finally, the third subfigure displays the resulting dataset after the removal of non-fire samples. Here, the majority class has been reduced to achieve the desired imbalance ratio, resulting in a more balanced representation of fire and non-fire instances.

By visually depicting the impact of Random Undersampling, this figure provides a clearer understanding of how the technique modifies the class distribution in the dataset. It demonstrates the reduction in the number of non-fire samples while retaining the original distribution of the fire instances.

Random Undersampling offers several advantages, including computational efficiency due to its straightforward implementation. By randomly selecting samples for removal, it avoids the need for complex computations or expensive algorithms. However, it is important to note that this method also has limitations. The main drawback is the potential loss of valuable information about the problem domain. Since the samples are removed randomly, there is a possibility of discarding instances that may contain crucial insights or important patterns relevant to forest fire prediction. Therefore, careful consideration should be given to the potential trade-off between computational efficiency and the risk of losing informative data when employing Random Undersampling.

### 2.3.2 Near Miss Undersampling technique

Near Miss undersampling is another valuable technique for addressing the class imbalance in datasets. Unlike Random Undersampling, which randomly removes samples from the majority class, Near Miss undersampling focuses on selecting a subset of majority class samples based on their proximity to the minority class instances in the feature space. The underlying idea is to retain the most informative majority class samples that are close to the minority class, as they are likely to contain valuable insights for classification.

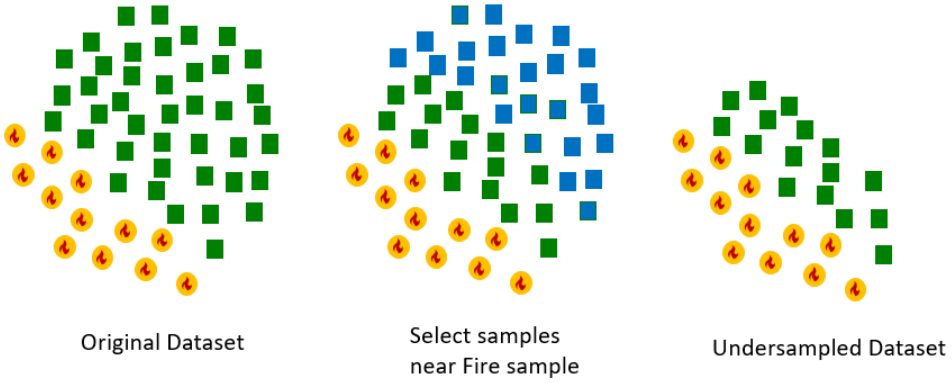



Figure 2.7: Near Miss Undersampling

Fig. 2.7 showcases the application of Near Miss undersampling in the context of forest fire prediction. The figure consists of three subfigures, each representing a different step in the Near Miss undersampling process.

In the first subfigure, we have the original dataset, which contains both fire and non-fire samples. This dataset exhibits a severe class imbalance, with the majority class (non-fire) overwhelming the minority class (fire).

The second subfigure demonstrates the outcome of the Near Miss undersampling technique. Here, the focus is on selecting samples from the majority class that are in close proximity to the fire samples. These selected samples, highlighted in the figure, aim to capture the crucial information and characteristics related to fire occurrences. By retaining only the samples near the fire instances, we aim to create a subset that is more representative of the minority class, enhancing the model's ability to learn from the informative majority class samples.

Finally, the third subfigure displays the result of further refinement in the Near Miss undersampling process. In this step, only the nearest non-fire samples to the fire instances

are kept. By narrowing down the selection to the nearest non-fire samples, we ensure that the retained majority class samples are the most relevant and influential for predicting fire events.

This figure visually demonstrates how Near Miss undersampling selectively retains majority class samples based on their proximity to the minority class, enabling the creation of a more balanced and informative dataset for forest fire prediction.

The Near Miss algorithm encompasses three different versions, each employing a distinct method for selecting the subset of majority class samples. In the subsequent sections, we will explore the details of the three versions of the Near Miss algorithm and discuss their individual merits and considerations.

### Near Miss 1

The Near Miss 1 technique focuses on selecting samples from the majority class based on their average distance from the $N$ minority samples in the feature space. This approach aims to retain the majority class instances that are closest to the minority class instances, thereby training the model to differentiate between these samples. The parameter $N$ can be chosen based on the specific requirements of the use case, allowing flexibility in determining the number of minority class samples to consider in the distance calculation.

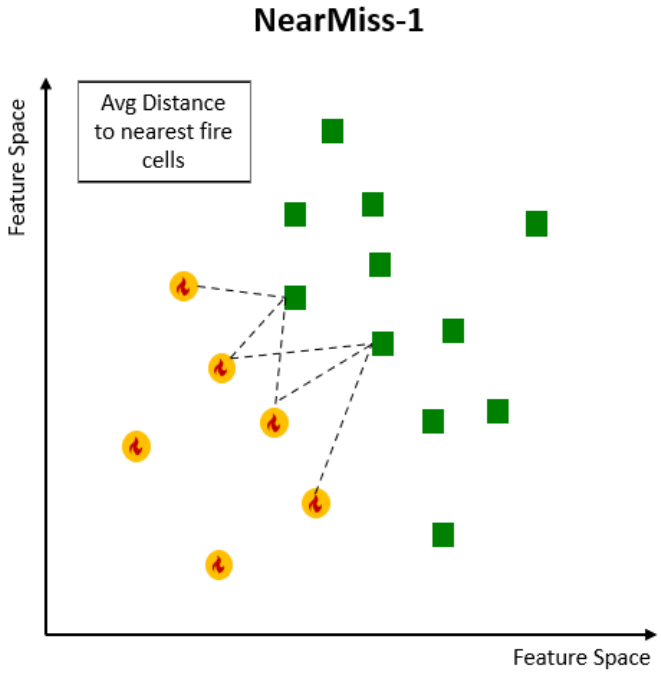



Figure 2.8: Near Miss version 1 Undersampling

In Fig. 2.8, the figure showcases the application of Near Miss 1 for forest fire prediction. The minority class, represented by fire instances, is distinguished from the majority class, represented by non-fire instances. The diagram demonstrates the process of selecting the non-fire samples with the smallest average distance from $N$ fire samples. In the figure, it shows that the value of $N$ taken here is three, and the distance is being calculated from the three nearest fire samples, and those non-fire samples are selected whose average distance to the three nearest fire samples is minimum.

The Near Miss 1 technique offers several benefits. Focusing on the non-fire samples that are closest to the fire instances, allows for a more targeted and informative representation of the majority class. This approach helps address the issue of class imbalance and provides a more balanced dataset for training the fire prediction model. Additionally, by considering the average distance, the technique takes into account the overall proximity between the two classes, enhancing the model's ability to capture the subtle patterns and characteristics associated with fire events. However, it may discard potentially informative non-fire samples farther from the fire instances and assumes that proximity reflects relevance. We will experiment and see how this performs for our case of forest fire prediction.

### Near Miss 2

The Near Miss 2 technique focuses on selecting samples from the majority class based on their average distance from the $N$ farthest minority samples in the feature space. This approach aims to retain the majority class instances that are closest to the farthest minority class instances, capturing the samples that are more representative of the minority class. The parameter $N$ determines the number of farthest minority class samples to consider in the distance calculation.

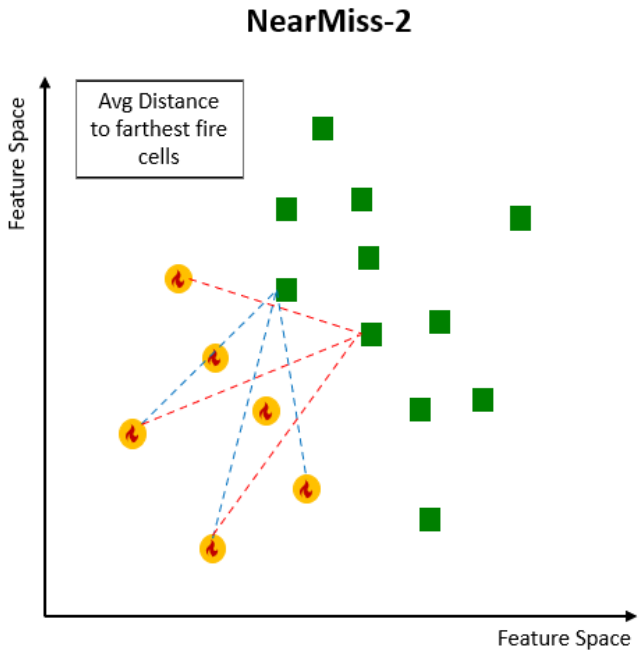



Figure 2.9: Near Miss version 2 Undersampling

In Fig. 2.9, the figure showcases the application of Near Miss 2 for forest fire prediction. The minority class, represented by fire instances, is distinguished from the majority class, represented by non-fire instances. The diagram demonstrates the process of selecting the non-fire samples with the smallest average distance from the $N$ farthest fire samples. In this illustration, the parameter $N$ is set to three, indicating that the distance is calculated from the three farthest fire samples, and the non-fire samples with the minimum average distance are selected.

The Near Miss 2 technique offers several benefits. Focusing on the non-fire samples that are closest to the farthest fire instances, helps address the class imbalance issue and

ensures a more balanced representation of the majority class. This approach considers the proximity between the two classes, allowing the model to capture the relevant patterns and characteristics associated with fire events. However, it may discard non-fire samples that are farther from the farthest fire instances, potentially losing some valuable information. The experiments will show it this undersampling technique behave on our forest fire imbalanced data.

### Near Miss 3

The Near Miss 3 undersampling technique follows a two-step process that involves two parameters, $M$ and $N$. In the first step, for each minority sample, we select the $M$ nearest neighbors from the majority class in the feature space. In the second step, we further filter the selected majority samples by considering their average distance from the $N$ nearest negative samples.

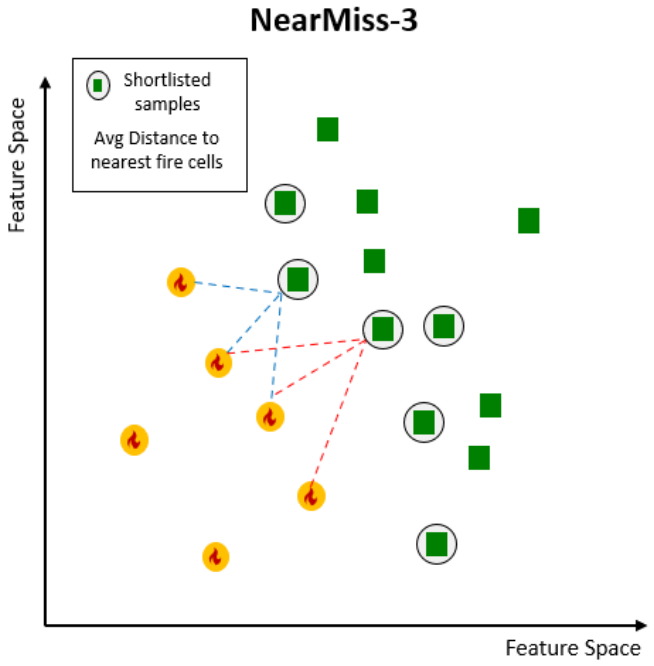



Figure 2.10: Near Miss version 3 Undersampling

As depicted in Fig. 2.10, the application of Near Miss 3 for forest fire prediction involves the following steps. In the first step, the $M$ nearest non-fire samples are chosen (shown as circles non-fire samples) for each fire sample. Then, in the second step, only those non-fire samples with the largest average distance to the $N$ nearest fire samples are retained.

The Near Miss 3 technique offers a unique approach to undersampling by considering the proximity between the minority (fire) and majority (non-fire) class samples in a two-step process. Selecting the non-fire samples that are nearest to the fire samples and then retaining only those with the largest average distance to the nearest fire samples, aims to improve the balance and informativeness of the dataset for forest fire prediction.

In our research, we will explore the application of the Near Miss 3 undersampling technique to our forest fire prediction problem. By examining its performance and considering its advantages and limitations, we aim to determine its suitability and effectiveness for our specific scenario.

## 2.4 Performance Metric

Performance Metrics play a crucial role in evaluating the effectiveness and reliability of machine learning models. They provide objective measures to assess the model's performance and guide decision-making processes. In the context of imbalanced data, where one class is significantly underrepresented compared to the other, selecting appropriate performance metrics becomes even more critical. In this section, we will describe and analyze various performance metrics in the context of forest fire prediction with imbalanced data. We will highlight the importance of choosing suitable metrics that account for the imbalanced nature of the data and effectively evaluate the model's ability to predict fire occurrences. By understanding and utilizing the appropriate performance metrics, we can assess the model's performance accurately, focusing on its ability to identify the minority class, which is crucial for effective fire management and prevention.

### 2.4.1 Issue with Prediction Accuracy as Performance Metric

During the evaluation of machine learning models, accuracy is commonly used as a performance metric. However, this approach may not always be suitable or appropriate, particularly when dealing with imbalanced datasets. The choice of performance metric should be carefully considered, taking into account the specific use case and characteristics of the dataset. In this section, we will examine the limitations of using accuracy as a performance metric for machine learning models trained on imbalanced datasets.

Let us consider a scenario where we have a dataset with binary class labels: True and False. The imbalance ratio between these classes is 100:1, meaning that for every True class sample, there are 100 False class samples. Now, suppose we develop a machine learning model that simply outputs "False" for all predictions. When we evaluate this model on a test dataset, we would achieve an accuracy of 99%. However, this high accuracy is misleading because the model has not learned anything meaningful from the minority class (True). In this case, accuracy fails to reflect the model's performance in correctly identifying the True class samples.

In our specific use case, where the objective is to predict the occurrence of forest fires, the desired output is "True" for fire samples and "False" for non-fire samples. It is crucial for our model to learn the distinguishing characteristics of fire instances and accurately predict their presence. However, if we solely rely on accuracy as the performance metric, the model may simply predict the majority class (Non-Fire) for all samples, resulting in high accuracy without effectively capturing the minority class (Fire). This highlights the need for alternative performance metrics that can better assess the model's ability to learn and predict the minority class, enabling more meaningful training and evaluation processes.

### 2.4.2 Performance Metric for Imbalanced data

When evaluating the performance of machine learning models on imbalanced data, it is important to use appropriate performance metrics that consider the imbalanced nature of the dataset. In our study, we have selected the following performance metrics to assess the effectiveness of our models:

**Confusion Matrix:** The confusion matrix is a commonly used performance metric for classification problems. It provides a tabular representation of the predicted and actual values. Fig. 2.11 visually illustrates a confusion matrix. For our case, the confusion matrix consists of four terms:

- True Ignition/True Positive (TP): The number of actual Ignition samples that are correctly classified as Ignition by the model.
- False Ignition/False Positive (FP): The number of actual Non-Ignition samples that are incorrectly classified as Ignition by the model.
- True Non-Ignition/True Negative (TN): The number of actual Non-Ignition samples that are correctly classified as Non-Ignition by the model.
- False Non-Ignition/False Negative (FN): The number of actual Ignition samples that are incorrectly classified as Non-Ignition by the model.

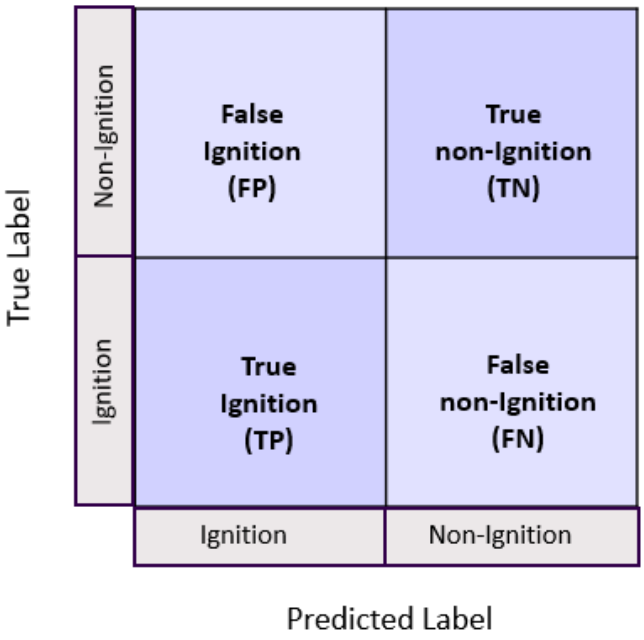



Figure 2.11: Confusion Matrix

The confusion matrix helps us derive several other performance metrics, including recall, sensitivity, specificity, and accuracy. It provides insights into the model's performance by indicating where the model might be making mistakes. In the context of imbalanced datasets, the confusion matrix is particularly valuable in understanding how well the model is able to learn and classify the minority class.

By analyzing the values in the confusion matrix, we can calculate performance metrics that are specifically designed for imbalanced datasets. These metrics allow us to assess the model's ability to correctly identify the minority class and account for the class imbalance.

**Sensitivity (Recall)**: Sensitivity, also known as the true ignitions rate or recall of the ignitions class, is a significant performance metric derived from the confusion matrix. It quantifies the ability of the model to correctly identify ignition samples out of the total actual ignition samples.

The formula for sensitivity is given by Equation 2.3:

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{2.3}$$

In the context of forest fire prediction, sensitivity plays a crucial role in evaluating how effectively the model can identify and predict forest fires among all the actual fire occurrences. A high sensitivity score indicates that the model is successful in detecting the majority of forest fires. This performance metric is particularly important as it focuses on the minority class, which is the occurrence of fire, allowing us to assess the model's ability to accurately predict fire events.

By measuring sensitivity, we can gain insights into the model's capability to capture the characteristics and patterns associated with forest fires. It helps us understand how well the model is performing in terms of identifying the rare and critical fire events, which is the main objective of our research.

**Specificity**: Specificity, also known as the true non-ignitions rate or recall of the non-ignition class, measures the ability of a classification model to correctly identify non-ignition samples out of the total actual non-ignition samples.

The formula for specificity is given by Equation 2.4:

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{2.4}$$

In the context of forest fire prediction, it is crucial to evaluate specificity as it quantifies the model's accuracy in classifying non-fire events correctly among all the non-fire instances. A high specificity score indicates that the model is effective in identifying and predicting the absence of fire accurately. This information is valuable for planners and decision-makers as it helps in resource allocation and planning for mitigation purposes. Knowing where a fire is not likely to happen enables better utilization of budget and resources.

It is important to note that there can be a trade-off between sensitivity and specificity. When the true positive rate (sensitivity) of the model increases, it is possible for the false positive rate to also increase, potentially leading to a decrease in specificity. However, our goal is to achieve a balance between sensitivity and specificity, aiming for a model with good sensitivity while maintaining a decent level of specificity.

**ROC_AUC:** The Receiver Operating Characteristic Area Under the Curve (ROC_AUC) is a performance metric commonly used in binary classification tasks. It measures the overall discriminative ability of the model by plotting the true ignition rate (sensitivity) against the false ignition rate (1-specificity) at various classification thresholds.

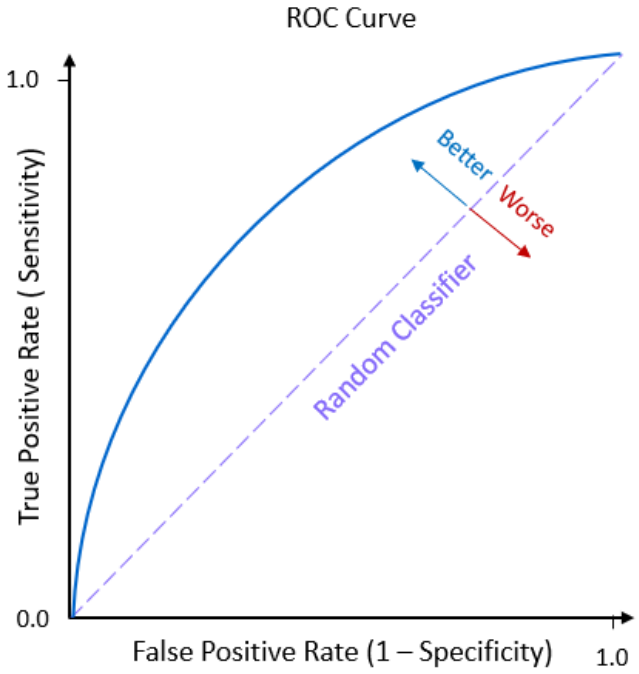



Figure 2.12: Receiver Operating Curve

As shown in Fig. 2.12, the 0.5 ROC_AUC score denotes the performance of the Random Classifier. A higher ROC_AUC score indicates a better ability of the model to distinguish between the two classes and a lower score denotes a worse ability to distinguish between two classes.

ROC_AUC is particularly valuable for forest fire prediction as it allows us to optimize the model based on both sensitivity (ability to correctly identify fire instances) and specificity (ability to correctly identify non-fire instances). By considering the entire range of classification thresholds, ROC_AUC provides a comprehensive evaluation of the model's ability to handle the imbalanced nature of the data and make accurate predictions for both fire and non-fire instances.

# Chapter 3

# Data Collection Framework and Application

This section provides a comprehensive and in-depth examination of our data acquisition framework, which plays a crucial role in integrating diverse types of features encompassing meteorological variables, biophysical variables, and topographical variables. The successful handling of such heterogeneous data necessitates the adoption of a systematic approach, wherein data originating from various sources undergo a standardized conversion process, enabling subsequent consolidation into a unified dataset through the utilization of our well-established framework.

Our Data Collection Framework is meticulously applied within the context of the province of Alberta, located in Canada. This particular region proves to be highly conducive to our research objectives, primarily due to its extensive historical record of forest fires spanning multiple decades, as well as the availability of comprehensive topographical data that encompasses the region's diverse landscape. Covering a total area of 661,848 $km^2$, Alberta boasts a substantial proportion of land dedicated to forested areas, accounting for approximately 61% of its overall land area. These factors underline the significance and relevance of analyzing the patterns and dynamics of forest fires within this region.

The comprehensive analysis of forest fire trends in Alberta is founded upon the invaluable dataset provided by Alberta.ca[18]. This dataset offers a wealth of information regarding the occurrence and characteristics of wildfires in the region, allowing for a detailed examination of the patterns and dynamics associated with forest fire occurrences. The consistent and recurring nature of these wildfires year after year underscores the urgent need for a thorough understanding of the factors contributing to their ignition and

spread. Specifically, this section of the thesis addresses the research questions RQ1 and RQ2, which pertain to the identification of underlying factors contributing to forest fire ignition and the development of a comprehensive data collection framework, respectively.

## 3.1 Framework for Data Collection

Our primary objective, following the processing of various features, is to establish a standardized tabular format that ensures consistent spatial and temporal resolutions. This standardization allows for the seamless integration of data tables using a unique identifier of Grid_id and date shared among them, as depicted in Fig. 3.1. The details of the process is described in the coming subsections.

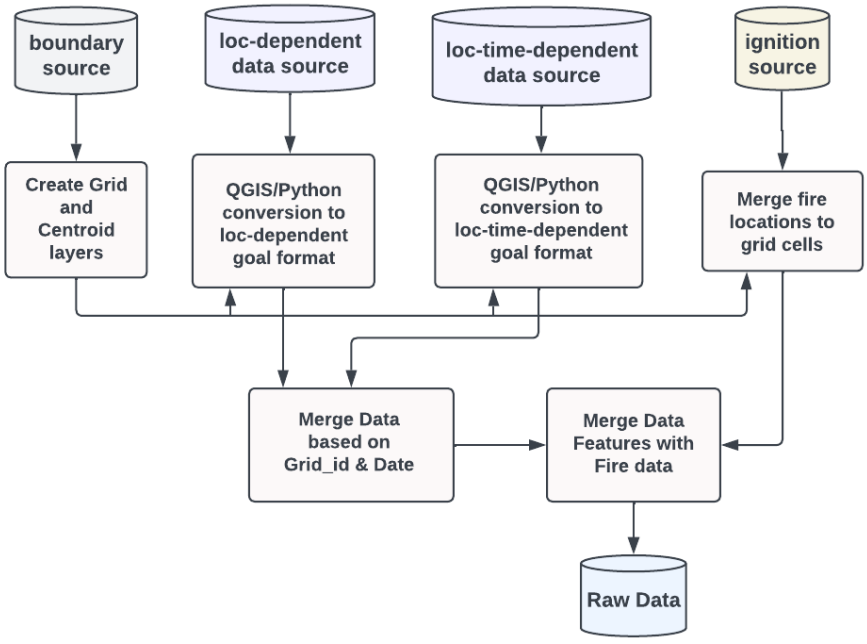



Figure 3.1: Data Collection Framework

### 3.1.1 Grid Cells

The datasets considered for the analysis of forest fires fall under the category of geo-databases, which are databases that incorporate location references on Earth. These datasets consist of various features that exhibit varying values depending on their geographical coordinates, namely latitude and longitude. Each feature is collected by different authorities, leading to variations in formats and resolutions. To standardize the datasets and ensure uniformity, a grid-based approach was employed, dividing the entire region into grid cells using a spatial resolution of 10 km by 10 km.

By utilizing this grid system, each grid cell is assigned a single value for each data feature. In cases where a feature exhibits variability within a grid cell, the values are aggregated and assigned to that specific grid cell. To facilitate the integration of diverse datasets with a shared location, a unique Grid_id is assigned to each grid cell. This Grid_id serves as a representation of the cell's location and will be utilized in joining various datasets that correspond to the same location.

This approach enables the consolidation of disparate data sources into a standardized format, providing a foundation for comprehensive analysis and modeling of forest fires within the region of interest.

#### Spatial Resolution

The selection of a spatial resolution is an essential consideration for our dataset. After careful evaluation, a spatial resolution of 10 km by 10 km was chosen. This resolution provides a sufficiently large area that can effectively capture fire ignition points within the forested regions.

When determining the spatial resolution of the combined data, it is crucial to consider the resolution of each individual dataset that contributes to the combination. The approach is to identify the dataset with the lowest resolution and use that as the basis for the combined data.

In our case, the meteorological data exhibits the lowest resolution, approximately 10 km by 10 km. By adopting this resolution for the entire Alberta region, we are able to acquire adequate data for understanding feature trends and wildfire patterns.

It is worth noting that employing a higher resolution of 1 km by 1 km would introduce challenges, such as a higher class imbalance within the combined dataset. Conversely, utilizing a lower resolution of 100 km by 100 km would encompass larger regions, potentially

resulting in multiple fire ignition points falling within the same grid cell and exacerbating the class imbalance issue. Therefore, the chosen 10 km by 10 km spatial resolution strikes a balance between capturing meaningful information and managing class imbalances effectively.

#### Temporal Resolution

In order to align with the specific research objective of analyzing wildfire ignitions, a temporal resolution of one day was selected. This resolution is deemed adequate for capturing the onset of fires accurately.

Taking into account the examination of fire data spanning 18 years, sourced from [18], it was observed that the majority of fires occurred during the period between April and October. Consequently, the temporal dimension of the data encompasses each day from April 1st to October 31st for the years 2000 to 2018. This temporal range effectively covers the critical months associated with wildfire incidents, facilitating a comprehensive analysis of fire patterns and trends over the specified time span.

### 3.1.2 Data Conversions

The dataset comprises diverse data formats, including raster files and CSV files, each with distinct spatial and temporal resolutions. To address these variations, we employ an analytical approach to define a desired format for two distinct categories of variables within the geodatabase.

The data encompass two distinct categories of features:

- **Location-dependent features**: The first category of features in our dataset comprises location-dependent variables. These features are solely influenced by geographical location and remain constant over time, specifically within a short-term timeframe of a few decades. Examples of such features include Slope, Aspect, and Elevation, which are topographical characteristics. The desired tabular format for these location-dependent features is illustrated in Tab. 3.1. To achieve a standardized format for these features, we assign a Unique Id to represent each location, denoted by the Grid_id.
- **Location and time-dependent features**: The second category of features in our dataset encompasses variables that are influenced by both time and location. These

| Grid_id | Feature name |
|---|---|
| | |

Table 3.1: Location-dependent features goal format

features exhibit variations in value based on specific dates and geographical locations. An example of such a feature is temperature, which can differ across different locations and also vary for a particular location on different days. The meteorological and biophysical features included in our data fall under this category. To achieve a standardized format for these time and location-dependent features, we establish a Unique Id by combining the Grid_id and date, representing each specific combination of location and time. The desired tabular format for these features is demonstrated in Tab. 3.2.

| Grid_id | date | Feature name |
|---|---|---|
| | | |

Table 3.2: Location and time-dependent features goal format

The data transformation process to achieve the desired tabular format involves leveraging the capabilities of QGIS and Python. The specific steps undertaken for each dataset may vary depending on its original format, which can include raster, shape, or CSV files. These steps are designed to address the unique characteristics of each dataset and ensure consistency in the final format.

The goal is to convert each dataset into the appropriate format based on its location dependency or location and time dependency. Fig. 3.1 shows the details of the framework of data collection. For location-dependent features, such as topographical variables (e.g., Slope, Aspect, Elevation), the data is organized into a tabular format where each location is assigned a unique identifier, represented by the Grid_id. This allows for a standardized representation of the location-dependent features across the dataset.

For location and time-dependent features, such as meteorological and biophysical variables, the data is structured in a tabular format that incorporates both the Grid_id and the corresponding date. This combination of location and time creates a unique identifier for each specific combination, enabling the integration of time-dependent information with spatial context.

The combined datasets, now in a standardized tabular format, can be merged based on the unique identifiers. This data integration process ensures that the location and time-

dependent features are properly aligned, providing a comprehensive dataset for further analysis and modeling.

The data collection framework presented in Fig. 3.1 provides a visual representation of the overall process, highlighting the sequence of steps involved in converting and combining the diverse data sources. It serves as a blueprint for reproducing the data collection process for other regions of interest, allowing researchers to apply the framework to different geographical areas and extend the analysis beyond the scope of the Alberta region.

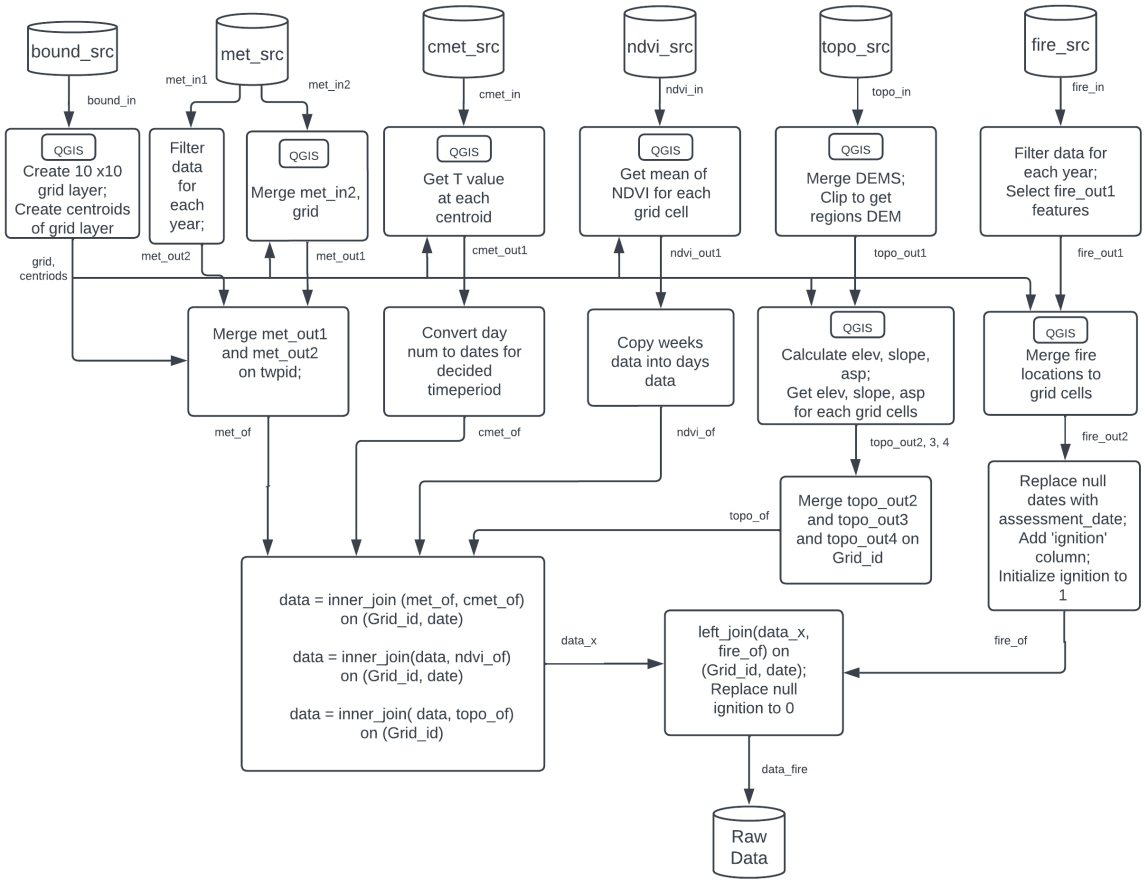



Figure 3.2: Data Collection Framework for Alberta

| Key | Value |
|---|---|
| bound | Boundary Data |
| met | Meteorological Data |
| ndvi | NDVI Data |
| topo | Topography Data |
| fire | Fire Data |
| cmet | Copernicus Meteorological Data |
| Grid_id | Grid id |
| date | Date of ignition |
| Prcp | Precipitation |
| MaxT | Maximum Temperature |
| RelHum | Relative Humidity |
| Ws | Average 10 meter wind speed |
| T | Temperature at 2 meter |
| NDVI | Normalized Difference Vegetation Index |
| slope | Slope |
| elev | Elevation |
| asp | Aspect |
| ignition | Ignition |
| twpid | Township id |
| DEM | Digital Elevation Model |

Table 3.3: Abbreviations

| Name | Meaning |
|---|---|
| x_src | Source of data of type x, x ∈ bound, met, cmet, ndvi, topo, fire |
| x_in | Inputs from data source x |
| x_inj | Multiple inputs from data source x, j ∈ 1,2 |
| x_outj | Intermediate outputs after processing of x type of data, j ∈ 1,2, 3,4 |
| x_of | Final output after processing of x type of data |

Table 3.4: Nomenclature of the input and output of data

## 3.2 Data Collection Framework for Alberta

The creation of a comprehensive dataset for the province of Alberta is a fundamental aspect of our research, as it forms the foundation for analyzing and modeling forest fire dynamics in the region. Leveraging the data collection framework, we have systematically integrated various types of features to construct a unified dataset that encompasses meteorological variables, biophysical variables, and topographical variables. This dataset serves as a vital resource for examining the relationships between these diverse factors and their impact on wildfire occurrences. By applying the framework to the specific context of Alberta, we have successfully combined data from different sources, standardized their formats, and ensured consistent spatial and temporal resolutions. The Fig. 3.2 shows the data collection

| Name | Type | Features |
|---|---|---|
| bound_in | Shape file | Alberta Boundary shape file |
| grid | Shape file | 10 × 10 grid layer |
| centroid | Shape file | Centroids layer of 10 × 10 grid layer |
| met_in | csv | twpid, date, Prcp, MaxT, RelHum, Ws, Prcp_f, MaxT_f, RelHum_f, Ws_f |
| met_in2 | csv | twpid, latitude, longitude |
| met_out1 | csv | twpid, date, Prcp, MaxT, RelHum, Ws, latitude, longitude |
| met_of | csv | Grid_id, date, Prcp, MaxT, RelHum, Ws |
| cmet_in | raster | 365 bands col1, . . . , col365, temp of each day of year |
| cmet_out1 | csv | Grid_id, col1, . . . , col365, temp of each day of year |
| cmet_of | csv | Grid_id, date, T |
| ndvi_in | raster | 26 bands 1 band for each week from April to October |
| ndvi_out1 | csv | Grid_id, col1_mean,... , col26_mean NDVI mean for each week |
| ndvi_of | csv | Grid_id, date, NDVI |
| topo_in | raster | 3 Digital Elevation models |
| topo_out1 | raster | 1 DEM for complete region of Alberta |
| topo_out2 | csv | Grid_id, slope |
| topo_out3 | csv | Grid_id, asp |
| topo_out4 | csv | Grid_id, elev |
| topo_of | csv | Grid_id, slope, asp, elev |
| fire_in | csv | fire_id, year, date, assessment_date, latitude, longitude |
| fire_out1 | csv | Date, assessment_date, latitude, longitude |
| fire_out2 | csv | Grid_id, date, assessment_date, latitude, longitude |
| fire_of | csv | Grid_id, date, ignition |
| data_x | csv | Grid_id, date, Prcp, MaxT, RelHum, Ws, T, NDVI, slope, elev, asp |
| data_fire | csv | Grid_id, date, Prcp, MaxT, RelHum, Ws, T, NDVI, slope, elev, asp, ignition |

Table 3.5: Data Features

| Name | Type | Features |
|---|---|---|
| bound_src | shape file | https://open.alberta.ca/opendata/gda-4d939041-851b-4848-bd30-44dbf129e16c |
| met_src | csv | https://acis.alberta.ca/acis/township-data-viewer.jsp |
| cmet_src | shape file | https://cds.climate.copernicus.eu/cdsapp!/dataset/reanalysis-era5-complete?tab=overview |
| ndvi_src | raster | https://open.canada.ca/data/en/dataset/44ced2fa-afcc-47bd-b46e-8596a25e446e |
| topo_src | raster | https://earthexplorer.usgs.gov/ |
| fire_src | csv | https://www.alberta.ca/wildfire-maps-and-data.aspx |

Table 3.6: Data Sources

framework applied for Alberta. Tables 3.3, 3.4, 3.5, and 3.6 show the description of various aspects of the framework. This meticulously crafted dataset not only captures the unique characteristics of the Alberta region but also provides a solid basis for conducting in-depth analyses and developing accurate predictive models to enhance our understanding of forest fire behavior in this area.

### 3.2.1 Provincial Boundary

The determination of Alberta's provincial boundary relied on the utilization of the Alberta Census Boundaries - Current (2021) dataset, obtained from Open Government Data - Alberta [17]. Among the shape files available within the Alberta Census Boundaries dataset, specifically, Alberta Census Division 2021 was selected for our analysis. The data required no additional processing and was utilized in its original form. According to Statistics Canada, the linear resolution of this dataset is reported to be 1 meter. Given that our grid cells have a resolution of 10 km by 10 km, no further validation or adjustment was deemed necessary, and the data was directly incorporated into our research framework.

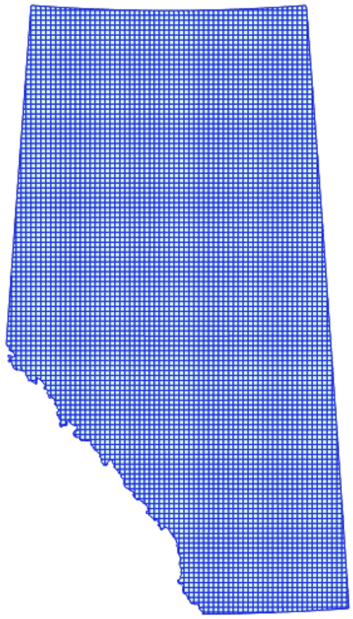


Figure 3.3: Alberta 10 km by 10 km Grid Layer

### 3.2.2 Alberta Grid cells

In order to achieve data standardization, a grid-based approach was employed to harmonize the various datasets. The entire region of Alberta was subdivided into grid cells using a spatial resolution of 10 km by 10 km. Each grid cell represents a distinct geographic unit within the region. For each data feature, a single value was assigned to every grid cell. In

cases where the feature exhibited variability within a grid cell, an aggregation of values was assigned. This grid system, depicted in Fig. 3.3, was established using QGIS. Each grid cell was assigned a unique identifier, known as the Grid_id, which serves as a reference to its specific location. This Grid_id plays a crucial role in joining disparate datasets that share the same geographic location. Additionally, centroids were generated for each grid cell using QGIS, as illustrated in Fig. 3.4.

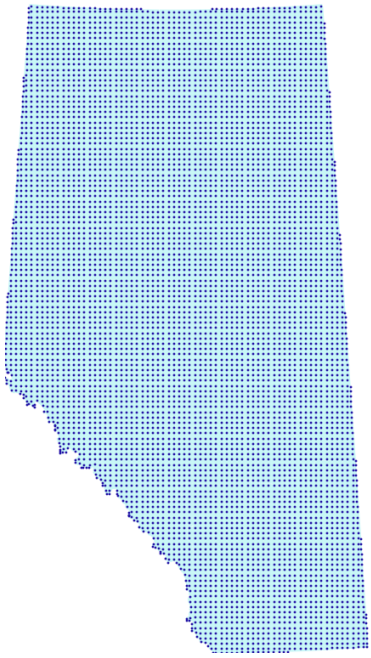


Figure 3.4: Alberta 10 km by 10 km Centroid Layer

### 3.2.3 Meteorological Data

The meteorological data utilized in this study were sourced from Alberta Agriculture, Forestry and Rural Economic Development [2], specifically from the Alberta Climate Information Service (ACIS) dataset, as indicated in Fig. 3.2 under the label met_src. The ACIS dataset comprises weather data collected from various meteorological stations operated by government agencies. While our analysis covers the entire region of Alberta, it is important to note that weather stations are situated at limited locations across the province. To obtain data for the entire region, we relied on ACIS's interpolated historical climate data for Alberta townships. This dataset considers the subdivision of Alberta into townships, each measuring approximately 9.65 km by 9.65 km. Daily gridded data is provided for each township, aligning with our desired spatial resolution of 10 km by 10 km.

In our study, our focus was on analyzing four key variables within the meteorological dataset: Maximum Temperature, Precipitation, Wind Speed, and Relative Humidity. To obtain the necessary data, we utilized the meteorological source (met_src), which consisted of two distinct inputs: met_in1 and met_in2.

| | twpid | date | Prcp | MaxT | RelHum | Ws |
|---|---|---|---|---|---|---|
| 0 | T001R01W4 | 01JAN2000 | 0.0 | 2.86 | 92.28 | 3.36 |
| 1 | T001R01W4 | 02JAN2000 | 0.0 | -5.07 | 86.67 | 14.84 |
| 2 | T001R01W4 | 03JAN2000 | 0.0 | -11.84 | 75.61 | 6.52 |
| 3 | T001R01W4 | 04JAN2000 | 0.0 | 4.75 | 78.27 | 3.28 |
| 4 | T001R01W4 | 05JAN2000 | 0.0 | 3.36 | 84.78 | 17.28 |

(a) Weather data features (met_in1)

| twpid | Lat | Long |
|---|---|---|
| T001R01W4 | 49.0432 | -110.0724 |
| T001R02W4 | 49.0429 | -110.2067 |
| T001R03W4 | 49.0428 | -110.3403 |
| T001R04W4 | 49.0428 | -110.4738 |
| T001R05W4 | 49.0428 | -110.6072 |

(b) Township metadata (met_in2)

Figure 3.5: Meteorological data input

met_in1, as shown in Fig. 3.5a, is a comprehensive CSV file containing climate data for all years. It includes columns such as Township ID (twpid), Date (date), Precipitation (Prcp), Maximum Temperature (MaxT), Relative Humidity (RelHum), and Wind speed (Ws). On the other hand, met_in2, represented in Fig. 3.5b, is another CSV file that includes township IDs (twpid) along with their corresponding latitude (Lat) and longitude (Long).

To streamline the data processing, we first filtered the information within met_in1 for each year, resulting in a separate CSV file named met_out1. This step allowed us to isolate

| twpid | Grid_id |
|---|---|
| T022R14W4 | 682187 |
| T022R14W4 | 682188 |
| T022R14W4 | 682189 |
| T022R14W4 | 682190 |
| T022R14W4 | 682191 |

Figure 3.6: Township ID joined with grid_id (met_out2)

the climate data for individual years and facilitate further analysis.

Next, leveraging the geographical information provided in met_in2, we conducted a spatial join using the software QGIS. This spatial join involved combining the grid layer with met_in2, enabling the creation of a CSV file that establishes an association between each Grid_id and its corresponding Township ID. This association between grid cells and township IDs is depicted in Fig. 3.6, providing a visual representation of the joined data.

| | Grid_id | date | Prcp | MaxT | RelHum | Ws |
|---|---|---|---|---|---|---|
| 0 | 682187 | 2018-04-01 | 0.00 | -9.68 | 81.47 | 7.06 |
| 1 | 682187 | 2018-04-02 | 0.00 | -5.19 | 76.34 | 9.21 |
| 2 | 682187 | 2018-04-03 | 0.00 | -5.57 | 82.82 | 9.38 |
| 3 | 682187 | 2018-04-04 | 0.10 | -4.75 | 81.22 | 12.63 |
| 4 | 682187 | 2018-04-05 | 0.00 | -7.69 | 73.13 | 16.13 |

Figure 3.7: Goal format of Meteorological data (met_of)

To get the goal format of these location and time-dependent features, we performed a join operation between met_out1 and met_out2 based on the shared township ID. This join operation resulted in the generation of a tabular format known as met_of, as shown in Fig. 3.7. The met_of format includes columns such as Grid_id, date, precipitation, Maximum Temperature, Relative Humidity, and wind speed. The overall process, as described, is illustrated in Fig. 3.2, providing a comprehensive overview of the meteorological data collection and transformation steps.

### 3.2.4 Copernicus Meteorological Data

The daily temperature records at 12 noon were obtained from ERA5-Land[13], a comprehensive meteorological dataset. To specifically capture the meteorological conditions within the region of Alberta, the data collection was confined to a defined geographical extent, with North = 60, South = 48, East = -109, and West = -121 serving as the boundaries. The temperature data was initially available in the form of a raster file (cmet_in), where each of the 365 bands(columns) represented the temperature data for a specific day.

**Algorithm 1** QGIS Steps Copernicus Data

1: Download data from CDS website.
2: Create a new project
3: Add boundary layer.
4: Add data as a raster layer.
5: Reproject data using wrap (reproject).
6: Clip the layer by the extent of the boundary.
7: Import centroids shape file.
8: Perform Point Sampling analysis.
9: Select the days and grid id column.
10: Export the data in csv format (cmet_out1).

To extract the temperature values at specific locations, we performed a point sampling analysis using the Geographic Information System (GIS) software QGIS. This analysis involved overlaying the cmet_in raster file with a shapefile containing the centroid locations of interest. The resulting output was a CSV file (cmet_out1), which captured the association between the Grid_id (representing the location) and the corresponding temperature data for each day of the year. The detailed steps of this process are outlined in Alg. 1.

| id | temp_1 | temp_2 | temp_3 | temp_4 | temp_5 | temp_6 | temp_7 | temp_8 | ... | temp_356 | temp_357 | temp_358 | temp_359 | temp_360 | temp_361 | temp_362 | temp_363 | temp_364 | temp_365 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 682187 | 236.42056 | 261.82727 | 260.60316 | 258.45537 | 256.90065 | 269.82703 | 264.12549 | 263.51064 | ... | 267.22847 | 263.48256 | 261.60460 | 265.76256 | 265.40895 | 265.30519 | 262.90131 | 268.11090 | 271.23093 | 254.74339 |
| 682188 | 236.42056 | 261.82727 | 260.60316 | 258.45537 | 256.90065 | 269.82703 | 264.12549 | 263.51064 | ... | 267.22847 | 263.48256 | 261.60460 | 265.76256 | 265.40895 | 265.30519 | 262.90131 | 268.11090 | 271.23093 | 254.74339 |
| 682189 | 236.42056 | 261.82727 | 260.60316 | 258.45537 | 256.90065 | 269.82703 | 264.12549 | 263.51064 | ... | 267.22847 | 263.48256 | 261.60460 | 265.76256 | 265.40895 | 265.30519 | 262.90131 | 268.11090 | 271.23093 | 254.74339 |
| 682190 | 236.42056 | 261.82727 | 260.60316 | 258.45537 | 256.90065 | 269.82703 | 264.12549 | 263.51064 | ... | 267.22847 | 263.48256 | 261.60460 | 265.76256 | 265.40895 | 265.30519 | 262.90131 | 268.11090 | 271.23093 | 254.74339 |
| 682191 | 236.42056 | 261.82727 | 260.60316 | 258.45537 | 256.90065 | 269.82703 | 264.12549 | 263.51064 | ... | 267.22847 | 263.48256 | 261.60460 | 265.76256 | 265.40895 | 265.30519 | 262.90131 | 268.11090 | 271.23093 | 254.74339 |

Figure 3.8: Copernicus Meteorological Data for each grid cell (cmet_out1)

To further refine the temperature data, we utilized Python scripting to transform cmet_out1 into the desired tabular format, known as cmet_of (Fig. 3.9). This involved

restructuring the data by converting the days into columns and the dates into rows. The resulting format enabled a more convenient and structured representation of the temperature data for analysis.

| Grid_id | date | T |
|---|---|---|
| 682187 | 2006-04-01 | 273.50450 |
| 682187 | 2006-04-02 | 274.15900 |
| 682187 | 2006-04-03 | 272.44263 |
| 682187 | 2006-04-04 | 275.07300 |
| 682187 | 2006-04-05 | 278.41235 |

Figure 3.9: Final Copernicus Meteorological Data (cmet_of)

As part of our data filtering process, we retained temperature data from 1st April to 31st October, omitting the remaining time periods. This selection was based on the focus of our study, which centered around the period when forest fire incidents are more likely to occur in Alberta. By narrowing down the temporal scope, we obtained a subset of data that was specifically relevant to our research objectives.

Through the rigorous implementation of these steps, we successfully acquired, processed, and standardized the meteorological data, enabling a comprehensive analysis of temperature patterns and their relationship to forest fire occurrences.

### 3.2.5 Biophysical Data

The research incorporates the utilization of biophysical data, specifically focusing on the Normalized Difference Vegetation Index (NDVI). The NDVI provides valuable insights into vegetation dynamics, and for this study, historical AVHRR satellite images with a spatial resolution of 1 km [3] were employed to derive an accurate representation of the NDVI. These satellite images cover a substantial time span from 1987 to 2021 and are available as raster files, with each year represented by a separate file.

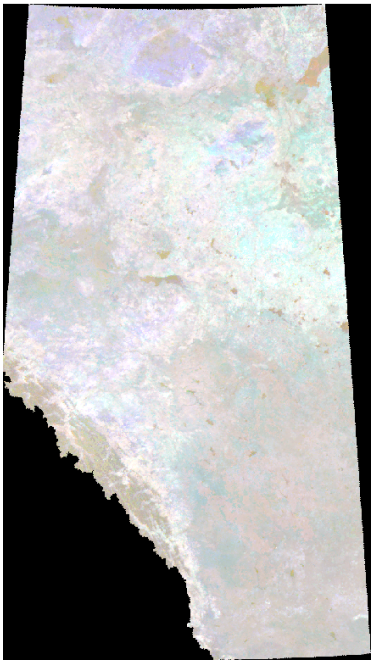


Figure 3.10: Normalized Difference Vegetation Index Data for Alberta (ndvi_in)

The NDVI dataset, known as ndvi_in, encompasses crucial variables such as latitude, longitude, and 26 bands of NDVI values. Each band corresponds to a specific Julian week, ranging from the 15th to the 41st week, spanning the period from April 6th to October 11th. This raster format of the NDVI dataset is visually depicted in Fig. 3.10. To optimize storage efficiency and facilitate streamlined processing, the NDVI values have been rescaled from the original range of [-1; 1] to the rescaled range of [0; 20,000].

To extract meaningful insights from the NDVI data, an essential step involved the application of the mean Multiband Zonal statistics function within the QGIS software. This statistical analysis enabled the calculation of the average NDVI value for each individual

**Algorithm 2** QGIS Workflow for NDVI Data Extraction

1: Create a new project in QGIS.
2: Import the boundary layer and the NDVI data.
3: Reproject the data using wrap reproject to ensure spatial consistency.
4: Clip the data to the Alberta region of interest.
5: Generate a grid layer to define the spatial units for analysis.
6: Apply the Multiband Zonal Statistics function, calculating the mean values for the clipped NDVI data within each grid cell.
7: Export the resulting data as a CSV file.

grid cell. The comprehensive steps involved in this data processing stage are outlined in Alg. 2. As a result of this analysis, a resulting output in the form of a CSV file, named ndvi_out1, was obtained, as illustrated in Fig. 3.11. This output comprises the Grid_id column, as well as 26 additional columns, representing the average NDVI value for each Julian week. The ndvi_out1 dataset serves as an intermediary step in the overall data processing pipeline.

| | id | _b1_mean | _b2_mean | _b3_mean | _b4_mean | _b5_mean | _b6_mean | _b7_mean | _b8_mean | _b9_mean | ... | _b19_mean | _b20_mean | _b21_mean | _b22_mean | _b23_mean | _b24_mean | _b25_mean | _b26_mean |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 682187 | 11338.226885 | 11433.763637 | 11528.291297 | 11623.813657 | 11630.140842 | 11636.528029 | 11948.129096 | 12109.396449 | 12038.637010 | ... | 13072.797312 | 13662.626699 | 14138.655915 | 13900.374040 | 13666.189947 | 13365.178599 | 13064.651113 | 12638.225918 |
| 1 | 682188 | 11267.300547 | 11469.572035 | 11532.208735 | 11596.201655 | 11529.342341 | 11611.857783 | 11943.772456 | 12067.458632 | 12006.080917 | ... | 12979.716177 | 13543.796143 | 13840.351633 | 13757.095265 | 13690.729956 | 13379.314159 | 13068.497065 | 12683.455215 |
| 2 | 682189 | 11441.623585 | 11464.755600 | 11535.862243 | 11609.195515 | 11478.742275 | 11705.227244 | 11935.326645 | 12192.497239 | 12032.754438 | ... | 12865.601066 | 13466.389909 | 13640.922016 | 13698.465477 | 13779.730530 | 13441.397869 | 13103.383941 | 12767.542246 |
| 3 | 682190 | 11454.903742 | 11469.201186 | 11560.102084 | 11653.480408 | 11536.824000 | 11668.120006 | 11926.029720 | 12351.050118 | 12222.832550 | ... | 13218.721915 | 13733.960958 | 14034.406797 | 14004.968012 | 13975.529227 | 13645.169852 | 13315.809227 | 12756.508329 |
| 4 | 682191 | 11415.413720 | 11454.706599 | 11516.435606 | 11579.164614 | 11520.550425 | 11513.910587 | 11952.459182 | 12354.984715 | 12260.135297 | ... | 13206.579760 | 13784.747402 | 14611.080761 | 14202.996261 | 13946.941936 | 13630.862239 | 13315.625266 | 12680.333756 |

Figure 3.11: Mean NDVI Values per Grid Cell (ndvi_out1).

Subsequently, in the Python, further refinement of the ndvi_out1 dataset was conducted to obtain the NDVI data at a daily level, spanning from April 1st to October 31st. This involved a straightforward approach of copying the NDVI values from the corresponding Julian week and assigning them to the respective day. Notably, specific adjustments were made to ensure the availability of complete daily NDVI data. For instance, data from April 6th was replicated for the days from April 1st to April 5th, while data from October 11th was extended to cover the remaining days in October. The detailed steps of this data transformation process are outlined in Alg. 3. These adjustments were crucial to maintaining the temporal continuity and completeness of the NDVI dataset throughout the desired time range.

The final output, represented as ndvi_of and illustrated in Fig. 3.12, showcases the goal format of the NDVI data. This standardized format facilitates further analysis and examination of the relationship between vegetation dynamics and forest fire occurrences

**Algorithm 3** Python Algorithm for NDVI Data Processing

1: Import the NDVI data from ndvi_out1.
2: Copy the data of the first band into the data for the period from 1st April to 5th April.
3: Copy the weekly data into the daily data for the period from 6th April to 11th October.
4: Copy the data of the last band into the data for the period from 12th October to 31st October.
5: Combine the above data to create a complete dataset for one year.
6: Save the dataset as a CSV file.

| | Grid_id | NDVI | date |
|---|---|---|---|
| 0 | 682187 | 10031.230796 | 2018-04-01 |
| 1 | 682188 | 10107.878808 | 2018-04-01 |
| 2 | 682189 | 10128.845456 | 2018-04-01 |
| 3 | 682190 | 10100.401407 | 2018-04-01 |
| 4 | 682191 | 10077.864084 | 2018-04-01 |

Figure 3.12: Final NDVI data (ndvi_of)

within the study region. The availability of this processed and standardized NDVI dataset enhances our ability to explore and interpret the dynamic nature of vegetation patterns and their potential influence on forest fire dynamics.

### 3.2.6 Topographical Data

The topographical data utilized in this research is derived from Digital Elevation Models (DEMs) obtained from Earth Explorer [15], specifically represented as topo_src in Fig. 3.2. For our study, we searched the Shuttle Radar Topography Mission (SRTM) 1 Arc-Second Global dataset [16] and Global Multi-resolution Terrain Elevation Data 2010 (GMTED2010) with a time range from 2000 to 2020 for Alberta region. These DEMs provide valuable information about the elevation and surface characteristics of the study area.

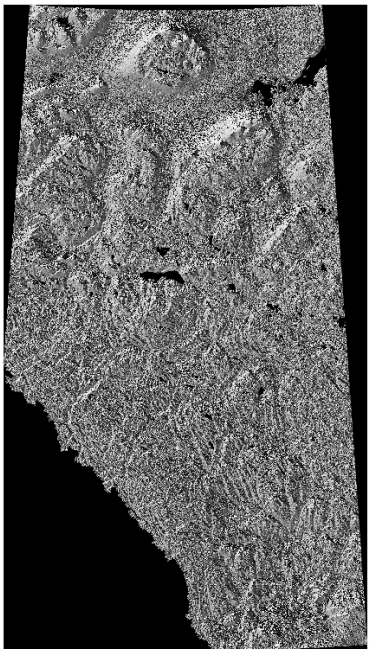


Figure 3.13: Digital Elevation Model of Alberta (topo_out1)

To acquire comprehensive topographical data specifically for Alberta, we downloaded products: GMTED2010N30W120, GMTED2010N50W120, GMTED2010N50W150. These DEMs, referred to as topo_in, were chosen based on their coverage and suitability for our analysis. Within the QGIS software, we combined these DEMs using the steps outlined in Alg. 4. This process involved reprojecting the data, merging the DEMs, and clipping the resulting dataset to the extent of Alberta. The resulting Digital Elevation Model for Alberta provides a detailed representation of the topography within the study area, as visually depicted in Fig. 3.13.

**Algorithm 4** QGIS Steps for combining Alberta DEMs

1: Download the DEMs. (three TIFF files for this case)
2: Merge the DEMs.
3: Reproject the merged file into the common projection.
4: Clip the reprojected DEM to obtain the final Alberta DEM.

| Grid_id | Slope |
|---|---|
| 682187 | 0.794462 |
| 682188 | 1.501793 |
| 682189 | 3.134629 |
| 682190 | 2.237989 |
| 682191 | 2.688662 |

(a) Slope data

| Grid_id | Elev |
|---|---|
| 682187 | 692.111111 |
| 682188 | 685.166667 |
| 682189 | 663.633333 |
| 682190 | 646.972222 |
| 682191 | 642.111111 |

(b) Elevation data

| Grid_id | Asp |
|---|---|
| 682187 | 223.312323 |
| 682188 | 212.090259 |
| 682189 | 221.814934 |
| 682190 | 238.394201 |
| 682191 | 168.393558 |

(c) Aspect data

Figure 3.14: Topographical data (topo_out2, topo_out3, topo_out4)

With the acquired Alberta DEM, we proceeded to extract the necessary topographical features for our analysis. Leveraging the extensive capabilities of QGIS, we employed various features and tools to calculate essential parameters such as slope and aspect from the DEMs. By following the steps outlined in Alg. 5, we derived separate datasets for slope, aspect, and elevation. The slope dataset quantifies the steepness of the land surface, while the aspect dataset indicates the orientation or direction of the slope. The elevation dataset provides information about the vertical height of the terrain. These datasets were then merged based on the common Grid_id, resulting in a consolidated CSV file that encompasses the topographical features of slope, elevation, and aspect. This consolidated dataset, referred to as topo_of, represents the desired goal format for topographical data, as depicted in Fig. 3.15.

**Algorithm 5** QGIS Steps for getting Slope, Elevation and Aspect for each Grid cell.

1: Create new project
2: Import Grid Layer, boundary layer
3: Calculate Slope and Aspect.
4: Save the csv files

| Grid_id | Slope | Elev | Asp |
|---|---|---|---|
| 682187 | 0.794462 | 692.111111 | 223.312323 |
| 682188 | 1.501793 | 685.166667 | 212.090259 |
| 682189 | 3.134629 | 663.633333 | 221.814934 |
| 682190 | 2.237989 | 646.972222 | 238.394201 |
| 682191 | 2.688662 | 642.111111 | 168.393558 |

Figure 3.15: Final Topographical data (topo_of)

The availability of this standardized topographical dataset enables comprehensive exploration and analysis of the terrain characteristics within the study region of Alberta. It provides valuable insights into the landscape features, which play a crucial role in understanding the dynamics of forest fires. By incorporating topographical data into our analysis, we can better assess the influence of terrain on fire behavior, identify vulnerable areas, and develop effective strategies for fire management and prevention.

### 3.2.7 Fire Data

Data pertaining to the locations of forest fires was acquired from the Alberta wildfire records, maintained by the Alberta Forest Service since 1931 [18]. Over the years, the method of record keeping has evolved, with the current system being the Fire Information Resource Evaluation System (FIRES), which serves as a centralized database. The data is available in the CSV (Comma Separated Values) format, encompassing various features related to forest fires, such as fire location latitude, fire location longitude, fire start date, assessment datetime, and fire year.

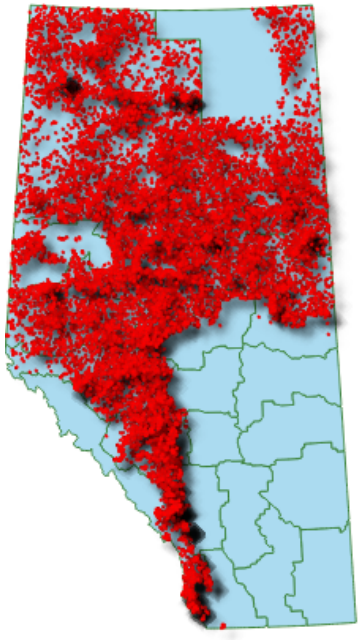


Figure 3.16: Fires in Alberta in last 18 years

The fire start date represents the date when a fire actually initiated, and its determination depends on various factors related to the cause of the fire. Meanwhile, the assessment datetime field in the dataset refers to the date and time when the wildfire was evaluated by the employees of the Wildfire Management Division, providing a level of confidence in the data's accuracy. In cases where the fire start date was null, indicating that the actual start date was unknown, the assessment datetime was utilized as a substitute. The dataset can be visualized in QGIS, as depicted in Fig. 3.16, where it is evident that there are no fires outside the designated boundary, indicating the absence of errors in the dataset.

| date | assessment_datetime | latitude | longitude |
|---|---|---|---|
| 2018-04-14 15:00:00 | 2018-04-14 17:10 | 49.573420 | -114.363340 |
| 2018-04-14 15:00:00 | 2018-04-14 17:25 | 49.573420 | -114.363340 |
| 2018-04-24 13:00:00 | 2018-04-24 14:10 | 51.183467 | -114.525467 |
| 2018-04-22 12:00:00 | 2018-04-26 15:25 | 49.690832 | -113.964052 |
| 2018-04-28 12:15:00 | 2018-04-28 13:54 | 51.615800 | -115.389017 |

Figure 3.17: Fire data (fire_out1)

**Fire Data Analysis:** A visual analysis of the data reveals that there are relatively few wildfires in the Southwest Region of Alberta (Fig. 3.16). The majority of Alberta's land area can be categorized into two regions: the Green Area, which constitutes the forested portion and is characterized by a relatively low population density, and the White Area, which is more densely populated and encompasses central and southern Alberta. As observed in Fig. 3.16, the majority of wildfires occur in the Green Area, while the White Area records significantly fewer incidents. This disparity can be attributed to the fact that approximately 75% of the White Area is privately owned, resulting in limited public fire records for this region. Statistical analysis of the data reveals that, on average, there are approximately 1256 wildfires per year.

**Algorithm 6** QGIS Steps for assiging Grid_ids to fires

1: Create new project
2: Import Grid Layer
3: Add fire data csv as delimited text layer
4: Perform Spatial Join with base layer as grid layer and other layer as fire data layer (one to one intersection).
5: Export joined csv file.

**Fire Data Preparation** : Fire Data Preparation is a crucial step in the analysis of the wildfire dataset. To begin, the fire data for each year was filtered to include only the necessary features, such as the date, assessment date, latitude, and longitude. This filtering process resulted in a refined CSV file that captures the relevant information for further analysis and visualization (Fig. 3.17). Next, in order to facilitate the integration of fire data with other datasets, a series of processing steps were performed in QGIS. The algorithm outlined in Alg. 6 was followed to assign Grid_ids to the fire records based on their corresponding geographic locations. This process involved spatially joining the fire data with the grid layer, resulting in a new CSV file that includes the assigned Grid_id for each fire record (Fig. 3.18).

| id | date | assessment_datetime | latitude | longitude |
|---|---|---|---|---|
| 4706 | 2018-04-14 | 2018/04/14 | 49.57342 | -114.36334 |
| 4706 | 2018-04-14 | 2018/04/14 | 49.57342 | -114.36334 |
| 4564 | 2018-04-24 | 2018/04/24 | 51.183467 | -114.525467 |
| 5076 | 2018-04-22 | 2018/04/26 | 49.690832 | -113.964052 |
| 3815 | 2018-04-28 | 2018/04/28 | 51.6158 | -115.389017 |

Figure 3.18: Grid_id assigned fire data (fire_out2)

Throughout the fire data preparation phase, special attention was given to handling null values in the dates. In cases where the fire start date was missing, the assessment date was used as a substitute. This ensures that each fire event has a valid date associated with it, maintaining the integrity of the dataset.

| | Grid_id | date | ignition |
|---|---|---|---|
| 0 | 4706 | 2018-04-14 | 1 |
| 1 | 4706 | 2018-04-14 | 1 |
| 2 | 4564 | 2018-04-24 | 1 |
| 3 | 5076 | 2018-04-22 | 1 |
| 4 | 3815 | 2018-04-28 | 1 |

Figure 3.19: Final fire data (fire_of)

Furthermore, to provide additional context and facilitate further analysis, a new column named "ignition" was introduced in the fire dataset. This column was initialized with a value of 1 for all fire records, indicating the occurrence of a fire event. By including this

column, the fire dataset becomes more informative, allowing for the distinction between fire events and non-fire events when merged with other datasets. The resulting processed fire dataset, as depicted in Fig. 3.19, represents a comprehensive and standardized representation of the fire data. It incorporates the necessary information for spatial and temporal analysis, enabling a deeper understanding of the patterns and dynamics of wildfires in Alberta.

In summary, the fire data preparation process involved filtering the dataset, assigning Grid_ids based on geographic locations, handling null values in the dates, and introducing an "ignition" column. These steps ensure the reliability and compatibility of the fire data, setting the stage for meaningful analysis and interpretation in the context of the broader research objectives.

### 3.2.8 Data Integration and Combination

In the process of combining and creating our required database for Alberta, all the processed data features are merged based on their unique identifiers. The location-dependent features are associated with the unique identifier Grid_id, while the location and time-dependent features are identified by a combination of Grid_id and date. This ensures that each record in the combined dataset can be uniquely identified and linked to its corresponding location and time information.

**Algorithm 7** Data Integration and Combination

1: Import all the datasets for one year.
2: Inner join Meteorological and Remote data based on (Grid_id and date).
3: Inner Join the above data Temp data from csv based on (Grid_id and date).
4: Inner join the above Topographical Data with the above data based on (Grid_id).
5: Merge (Joined data, Fires data) left join based on (Grid_id and date)
6: Set the Null values in the ignition column in the joined data to 0 (no fire ignition).
7: Save the data_fire for that year.

To achieve the integration of the different data features, a series of join operations are performed. The algorithm outlined in Alg. 7 provides a comprehensive guide for joining the datasets. Initially, an inner join is performed between the meteorological data (met_of) and the corrected meteorological data (cmet_of) based on the common Grid_id and date. The resulting dataset is then further joined with the NDVI data (ndvi_of), again using the Grid_id and date as the joining criteria. This step ensures that the meteorological and NDVI data are aligned and associated with the appropriate locations and time periods.

| | Grid_id | date | Prcp | MaxT | RelHum | Ws | NDVI | T | Slope | Elev | Asp | fire | ignition |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 55.0 | 2018-04-01 | 0.000000 | -3.660000 | 47.090000 | 5.480000 | 3804.284842 | 251.046360 | 68.835158 | 339.524835 | 273.715633 | 0.0 | 0 |
| 1 | 55.0 | 2018-04-02 | 0.000000 | -4.490000 | 55.070000 | 7.160000 | 3804.284842 | 253.516570 | 68.835158 | 339.524835 | 273.715633 | 0.0 | 0 |
| 2 | 55.0 | 2018-04-03 | 0.000000 | -6.210000 | 66.880000 | 11.210000 | 3804.284842 | 258.258130 | 68.835158 | 339.524835 | 273.715633 | 0.0 | 0 |
| 3 | 55.0 | 2018-04-04 | 1.370000 | -6.470000 | 68.060000 | 8.290000 | 3804.284842 | 260.493790 | 68.835158 | 339.524835 | 273.715633 | 0.0 | 0 |
| 4 | 55.0 | 2018-04-05 | 0.000000 | -8.000000 | 59.490000 | 9.990000 | 3804.284842 | 255.341710 | 68.835158 | 339.524835 | 273.715633 | 0.0 | 0 |

Figure 3.20: Raw Data

Subsequently, the combined dataset is joined with the topographical data (topo_of) based on the shared Grid_id. This integration allows for the incorporation of topographical

features into the dataset, providing additional insights into the relationship between terrain characteristics and the occurrence of wildfires.

Finally, the resulting dataset, denoted as data_x, is left joined with the fire data (fire_of), linking the fire occurrences to their respective locations and time periods. To handle cases where there are no fire events, the null values in the ignition column are replaced with 0, indicating a non-fire event. This comprehensive integration of fire data completes the creation of the raw data, which represents a combined dataset for one year in Alberta, encompassing all three types of variables (Fig. 3.20).

To obtain a comprehensive database for Alberta, these steps are performed iteratively for each year of the available data, spanning a total of 18 years. The iterative process ensures the inclusion of historical information and enables the analysis of long-term trends and patterns in the dataset.

By combining and organizing the diverse datasets, our database provides a holistic and comprehensive representation of the various factors influencing wildfires in Alberta. This integrated dataset serves as a valuable resource for further analysis and exploration, enabling a deeper understanding of the dynamics, relationships, and trends related to wildfires in the region.

# Chapter 4

# Methodology of Handling Data Imbalance and Prediction Model

This section outlines the methodology utilized to develop a forest fire prediction model based on imbalanced data, encompassing contributions C2 and C3. This chapter presents the methodology employed for handling data imbalance in the context of forest fire prediction and outlines the workflow for creating a robust prediction model as represented in Fig. 4.1. Dealing with imbalanced data poses a significant challenge in developing accurate and reliable prediction models, as forest fire incidents are relatively rare compared to non-fire instances in the dataset. To address this issue, various techniques have been proposed in the literature, and this chapter extensively explores and evaluates these approaches. Moreover, this chapter provides a detailed description of the step-by-step workflow involved in constructing a forest fire prediction model, encompassing data preprocessing, algorithm selection, model training, and evaluation. By implementing an effective methodology for handling data imbalance and developing a robust prediction model, we aim to enhance the performance and effectiveness of forest fire prediction, ultimately contributing to the advancement of fire management strategies and the protection of our natural ecosystems.

## 4.1 Data Imbalance Handling

The dataset obtained from the data collection process encompasses a comprehensive 18-year timeframe, incorporating data from multiple reliable sources. Upon conducting a thorough analysis, a pronounced class imbalance became evident, as there exists a significantly larger

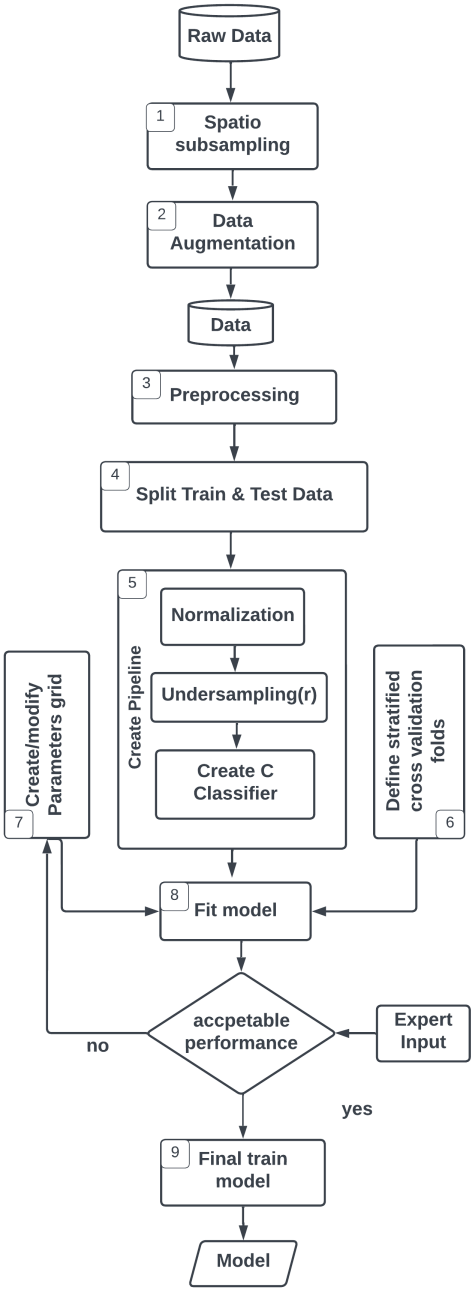



Figure 4.1: Modelling workflow

number of non-fire cells compared to fire cells within the dataset. Such a severe data imbalance can adversely impact the performance of machine learning models when directly employed for training. Consequently, addressing this issue necessitates a comprehensive approach with three steps: Changes in spatial resolution, data spatio-subsampling, and implementation of undersampling techniques to rebalance the dataset. However, even after undertaking undersampling, it was observed that the resulting machine learning model exhibited unsatisfactory performance. To overcome this limitation, it became apparent that additional features must be incorporated into the dataset to enhance the model's ability to effectively classify fire and non-fire cells. By incorporating these supplementary features, we aim to improve the predictive capabilities of the model, leading to more accurate forest fire predictions and bolstering the efficacy of fire management strategies.

### 4.1.1 Changes in Spatial Resolution

Initially, a spatial resolution of 1 km by 1 km was chosen for the data collection process. Consequently, all the collected data was converted to this resolution and combined, resulting in a substantial dataset of approximately 288 GB. However, the imbalance ratio of the data was found to be extremely high at 84,000:1, posing a significant challenge for subsequent analysis.

$$Imbalance\ ratio = 84,000 : 1$$

Increasing the resolution would result in a decrease in the imbalance between fire and non-fire samples. Since fire occurrences are rare events, the number of fires within a cell remains the same regardless of the cell size, whether it is 1 km by 1 km or increased to 10 km by 10 km. However, decreasing the resolution from 1km by 1km to 10 km by 10km would reduce the number of non-fire cells, thereby reducing the class imbalance. It is crucial to strike a balance when increasing the resolution to avoid situations where multiple fire events fall within the same cell, as this would decrease the overall number of fire cells and complicate the imbalance ratio. More details on the challenges encountered with spatial resolution are given in section 4.1.1. After doing changes in the spatial resolution the imbalance ratio of the data was reduced to **995:1**.

### 4.1.2 Data Spatio-subsampling

This study introduces a novel technique known as spatio-subsampling, which offers a solution to the data imbalance issue encountered in forest fire prediction. The spatio-

subsampling technique referred as box 1 in Fig. 4.1 aims to mitigate data imbalance by selectively removing data chunks from regions where no fire events have been observed over the past decades. Through a comprehensive analysis of the Fire Data, specific regions characterized by prevalent non-fire instances were identified, namely the Southeastern parts and sub-region of northern parts of Alberta. Leveraging the Grid_ids associated with these regions, we executed the spatio-subsampling approach by eliminating samples from our dataset that corresponded to these areas. By strategically removing non-fire cells, a substantial reduction in the imbalance ratio from **995:1 to 806:1** was achieved. Consequently, the non-fire cells were reduced by one-fourth, resulting in a significant improvement in the overall balance of the dataset.

The proposed spatio-subsampling technique offers a valuable approach that can be adopted by researchers across various regions to effectively address data imbalance challenges in their respective studies. By employing this technique, researchers can enhance the quality and reliability of their forest fire prediction models, ultimately contributing to more accurate fire management strategies.

### 4.1.3 Data Augmentation

The term "data augmentation" mentioned in this study refers to the process of enhancing the dataset by adding new columns that contain previously unavailable information. By incorporating additional relevant features, we aim to improve the overall performance of the prediction model. The objective is to expand the dataset with valuable data attributes that were not initially included, enabling the model to gain a more comprehensive understanding of the underlying patterns and relationships. The process is denoted by box 2 in Fig. 4.1. Through data augmentation, we can equip the model with a richer set of information, allowing it to make more accurate and reliable predictions for forest fire occurrences. This approach contributes to the refinement of fire management strategies and aids in mitigating potential risks associated with forest fires.

1. **Addition of month column**: Certain seasons exhibit higher fire probabilities, and specific regions experience elevated temperatures during particular times. To incorporate this temporal aspect, we added a "month of date" column to the dataset. This addition aims to enhance the model's classification of fire and non-fire cells by considering the impact of seasons and time on climatic and fire conditions. We encoded the month column using a label encoder, treating it as a categorical variable. This enables the model to better capture the influence of time-related factors, contributing to improved fire prediction capabilities.

2. **Addition of Grid location**: The variation in the risk of fires across different regions is widely acknowledged. Recognizing this, we sought to incorporate the location of each cell as a parameter in our data analysis. To achieve this, we introduced a categorical column called Grid_id into the dataset. The purpose of adding this column was to account for additional, yet unidentified, factors that are specific to certain regions and were not originally represented in our dataset. By including the Grid_id, we aimed to capture the influence of these region-specific factors, thereby enhancing the comprehensiveness of our data and enabling a more accurate analysis of fire occurrences.

3. **Addition of Fire History**: The presence of a significant number of historical fires in a specific region indicates a heightened risk of fire. Therefore, it is crucial to include the Fire History variable in the dataset. The decision to incorporate Fire History into the analysis was influenced by the study conducted by [14], where they utilized the average fire count as a baseline risk factor to predict forest fires, considering various values of other variables. Fire History is quantified as the average daily count of fires within a designated grid cell. This count is calculated by dividing the total number of fires that occurred in the grid cell by the total number of days considered in the dataset's timeframe. The numerical representation given in Eq. 4.1 provides valuable insights into the fire history, reflecting the average occurrence of fires within the specified grid cell.

$$\text{Fire History} = \frac{\text{Total num of fires in a grid cell}}{\text{Total num of days in time period}} \tag{4.1}$$

4. **Addition of trend of climatic data**: The occurrence of fire can also be influenced by the weather conditions in preceding days. The trend of climatic data is captured by aggregating the weather information from previous days. The climatic variables, such as Temperature, Precipitation, and Humidity, serve as indicators of these weather conditions. To represent the trend of climatic data in our dataset, we aggregate these variables over a specific number of previous days, denoted as $n$, and the current day is denoted by $d$. In our case, the climatic trend of the previous three days is aggregated for experiment purposes.

Hence, in Eq. 4.2, 4.3, 4.4, we utilize the Average Maximum Temperature (AMT), Total Precipitation (TP), and Average Wind Speed (AWS) from the previous three days as the data column values for the current day.

$$AMT(d) = \frac{\sum_{i=1}^{n} MaxT_{d-i}}{n} \tag{4.2}$$

$$TP(d) = \sum_{i=1}^{n} Prcp_{d-i} \tag{4.3}$$

$$AWS(d) = \frac{\sum_{i=1}^{n} Ws_{d-i}}{n} \tag{4.4}$$

| | Grid_id | date | Prcp | MaxT | RelHum | Ws | NDVI | T | Slope | Elev | Asp | ignition | month | loc | AWS | AMT | TP | fireHis |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 57 | 2000-04-01 | 2.91 | 8.94 | 59.87 | 18.92 | 11867.347 | 280.369 | 47.206 | 890.499 | 215.037 | 0.0 | 4 | 0 | 0.000 | 0.000 | 0.00 | 0.0 |
| 1 | 57 | 2000-04-02 | 0.00 | 9.10 | 58.95 | 6.25 | 11867.347 | 270.108 | 47.206 | 890.499 | 215.037 | 0.0 | 4 | 0 | 0.000 | 0.000 | 0.00 | 0.0 |
| 2 | 57 | 2000-04-03 | 0.00 | 12.85 | 66.62 | 9.34 | 11867.347 | 275.497 | 47.206 | 890.499 | 215.037 | 0.0 | 4 | 0 | 11.503 | 10.297 | 2.91 | 0.0 |
| 3 | 57 | 2000-04-04 | 4.37 | 7.19 | 70.29 | 12.02 | 11867.347 | 275.438 | 47.206 | 890.499 | 215.037 | 0.0 | 4 | 0 | 9.203 | 9.713 | 4.37 | 0.0 |
| 4 | 57 | 2000-04-05 | 5.82 | 3.99 | 78.57 | 7.25 | 11867.347 | 269.662 | 47.206 | 890.499 | 215.037 | 0.0 | 4 | 0 | 9.537 | 8.010 | 10.19 | 0.0 |

Figure 4.2: Data

### 4.1.4 Impact of Spatio-subsampling and Data Augmentation

The impact of the Spatio-subsampling and Data Augmentation techniques on the dataset can be observed in the comprehensive overview provided in Fig. 4.3. In the context of the dataset, let the number of rows and columns in the Raw Data be denoted as $M$ and $N$, respectively. After applying the Spatio-subsampling technique, the number of rows in the resulting Data denoted as $M'$, decreased compared to the original dataset. Additionally, through the process of Data Augmentation, the number of columns in the augmented dataset, denoted as $N'$, expanded in comparison to the initial dataset. This dual transformation seeks to optimize the dataset's predictive capabilities for the subsequent modeling stages. The effects of these transformations and their implications will be further explored and analyzed in the forthcoming Experimentation and Result section, providing valuable insights into the performance of the model.

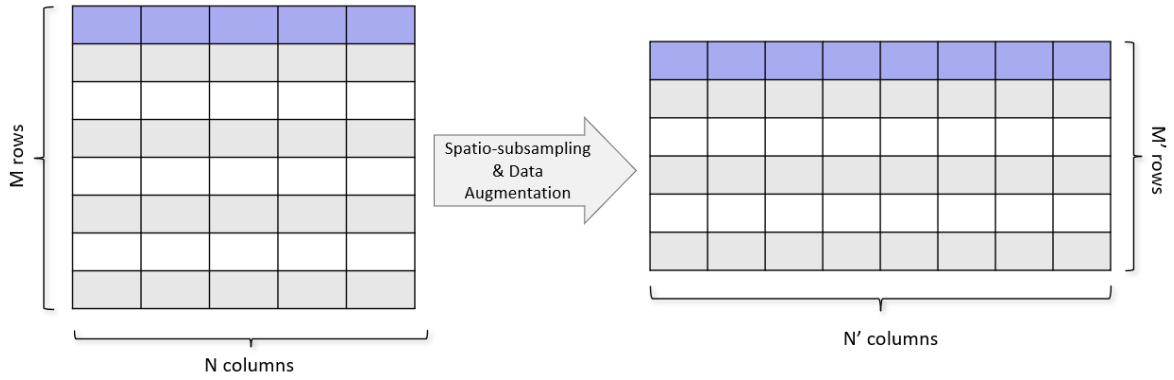



Figure 4.3: Spatio-subsampling and Data Augmentation

Tab. 4.1 provides a visual representation of the data that is the result of the combined processes of Data Spatio subsampling and Augmentation. This curated dataset snippet serves as the input for the subsequent stages of the modeling workflow, specifically the Preprocessing stage. Fig. 4.2 shows the snippet of the result dataset.

| Grid_id | date | Prcp | MaxT | RelHum | Ws | T | NDVI | slope |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | | |

| elev | asp | ignition | month | loc | AMT | TP | AWS | fireHis |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | | |

Table 4.1: Data

## 4.2 Forest Fire Prediction Modelling

The prediction modeling section encompasses a series of essential steps and techniques aimed at developing accurate and reliable prediction models for forest fire incidents. This section covers various subsections, including data preprocessing, train-test splitting, pipeline construction, normalization, undersampling, classifier selection, and grid search optimization as given in Alg. 8. The data preprocessing stage involves cleaning and transforming the dataset to ensure its quality and suitability for analysis. The train-test split is performed to evaluate the model's performance on unseen data, while the pipeline construction facilitates the seamless integration of preprocessing and modeling steps. Normalization techniques are applied to standardize the data and improve model convergence. Undersampling is employed to address data imbalance issues and enhance the model's ability to learn from rare fire events. Classifier selection involves choosing an appropriate algorithm that best suits the prediction task, considering factors such as interpretability and performance. Grid search optimization is conducted to fine-tune hyperparameters and optimize the model's performance. By following this comprehensive approach, we aim to develop robust prediction models that contribute to effective forest fire management and mitigation strategies.

### 4.2.1 Preprocessing

A comprehensive preprocessing procedure was carried out to meticulously prepare the dataset for modeling purposes, as depicted in box 3 of Fig. 4.1. During this preprocessing stage, special attention was given to addressing the presence of NULL values in certain rows pertaining to the topographical features. To ensure the completeness and consistency of the data, these NULL values were replaced with zero, a commonly employed approach that signifies the absence of a value for those specific features. This step is crucial as it ensures that the dataset remains intact and free from missing values, enabling accurate and reliable analyses in subsequent modeling stages.

To ensure the quality and integrity of the dataset, a thorough examination was conducted using boxplot visualizations. This analysis aimed to identify and address potential outliers, which are data points that deviate significantly from the expected distribution. Remarkably, no outliers were detected within the dataset, indicating that the data adheres closely to the expected range and distribution.

Furthermore, the categorical variable in the dataset underwent a transformation using the Label Encoder technique. This transformation is necessary to convert categorical

**Algorithm 8** Prediction Modeling (bold steps refer to boxes in Fig. 4.1)

**Require: Input** Dataset $\mathcal{D}$ containing predictor variables and target variable
**Ensure: Output** Trained prediction model

**Preprocessing:**
- Perform data cleaning to handle missing values
- Encode categorical variables using label encoding

**Train-Test Split:**
- Split $\mathcal{D}$ into training set $\mathcal{D}_{\text{train}}$ and test set $\mathcal{D}_{\text{test}}$

**Pipeline Construction:**
- Define a pipeline that encompasses the following steps:
    - **Normalize** the numerical features in $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{val}}$
    - Perform **undersampling** on $\mathcal{D}_{\text{train}}$ to address data imbalance
    - Define **Classifier**

**Grid Search:**
- Define a grid of hyperparameters to search for the chosen classifier.
- Perform grid search with cross-validation to find the best hyperparameters for the classifier

**Model Training:**
- Fit the pipeline to $\mathcal{D}_{\text{train}}$ using the selected classifier and optimized hyperparameters

**Model Evaluation:**
- Evaluate the trained model on $\mathcal{D}_{\text{test}}$ using appropriate metrics

**return** Trained prediction model

data into a numerical format that can be effectively utilized in various machine learning algorithms. The Label Encoder assigns a unique numerical label to each category, thereby facilitating the analysis and processing of categorical features.

By undertaking these meticulous preprocessing steps, the dataset is carefully refined, ensuring its suitability for subsequent modeling analyses. These steps enhance the overall quality and reliability of the dataset, setting a solid foundation for accurate and robust modeling outcomes.

### 4.2.2 Train Test Splitting

The dataset was subject to a partitioning process, resulting in the creation of two distinct subsets: the training data and the testing data, as visually depicted in box 4 of Fig. 4.1. This partitioning scheme plays a vital role in ensuring the accurate and reliable evaluation of the trained model's performance.

During the training phase, the training subset exclusively served as the target for all subsequent data processing and model training operations. This approach ensures that the model learns and adapts to the patterns and characteristics present in the training data, enabling it to make informed predictions and classifications. Importantly, the testing subset was kept separate and untouched during the training phase, preserving its integrity and allowing for an unbiased assessment of the trained model's performance.

To achieve a well-balanced representation of class labels in both the training and testing subsets, a train-test split ratio of 0.8 was adopted. This ratio designates that 80% of the dataset was allocated to the training data, while the remaining 20% was allocated to the testing data. This partitioning strategy strikes a balance between providing sufficient training data for the model to learn from and ensuring an adequate amount of independent data for unbiased evaluation.

To facilitate the equitable allocation of samples from each class during the partitioning process, the sklearn train-test stratified split method was employed. This method ensures that the distribution of class labels remains preserved in both the training and testing subsets. By maintaining the proportionate representation of different classes in each subset, the partitioning process minimizes the risk of introducing bias and enhances the overall reliability of the subsequent model evaluation.

By adhering to these partitioning procedures, the dataset is effectively divided into training and testing subsets, enabling a comprehensive assessment of the trained model's performance.

### 4.2.3 Pipeline

Pipelines are essential in machine learning model training, particularly when dealing with imbalanced data. They provide a structured and systematic approach, minimizing manual interventions and ensuring consistent handling of imbalances throughout the preprocessing and modeling stages. In our study, we acknowledge the significance of pipelines in addressing the imbalanced nature of the data, as depicted in box 5 of Fig. 4.1. Within the pipeline framework, we adopt a distance-based undersampling technique to mitigate the class imbalance. This technique selectively reduces the number of majority class samples exclusively in the training data, preserving the integrity of the validation data. By doing so, pipelines prevent data leakage and enable an unbiased evaluation of the model's performance on unseen data. Undersampling, also known as downsampling, allows us to create a more balanced representation of the dataset, diminishing the dominance of the majority class while retaining the inherent patterns and relationships in the imbalanced data.

One of the advantages of utilizing pipelines in the context of imbalanced data is their ability to facilitate effective hyperparameter tuning. By encapsulating the undersampling step within the pipeline, we can systematically fine-tune the parameters and configurations specific to the undersampling technique. This streamlined approach enables efficient exploration of various imbalance handling techniques and hyperparameter settings, aiding in the identification of the optimal configuration for our model. With the use of pipelines, we can enhance the reliability and performance of our prediction models on imbalanced datasets, ensuring that the imbalances are properly addressed while maintaining the integrity of the validation data.

1. **Normalization**: The data underwent normalization using the standard scalar from the sklearn library. The inclusion of normalization within the pipeline is essential as it ensures that the normalization process is applied solely to the training data while keeping the validation data separate. By performing normalization in this manner, we maintain the integrity of the validation data and prevent any information leakage between the training and validation subsets.

$$z = \frac{x - \mu}{\sigma} \tag{4.5}$$

   In Eq. 4.5, x is our data variable to be normalized, $\mu$ denotes the mean of the training data and $\sigma$ is the standard deviation of training data, and $z$ is the normalized value of $x$.

2. **Undersampling**: Our approach involves applying undersampling techniques, namely Random undersampling and three versions of Near miss undersampling, to tackle the class imbalance prevalent in the training data. The effectiveness of these techniques lies in their ability to reduce the dominance of the majority class, allowing for a more balanced representation of the data. To control the extent of downsampling, we employed a hyperparameter known as the downsampling ratio, denoted as $r$. This parameter played a crucial role in shaping the composition of the training data and ultimately influenced the performance of our model.

   Choosing an appropriate value for the downsampling ratio, $r$ is a critical step in our modeling process. The selection process involved experimentation and evaluation to assess the impact of different values of $r$ on the model's performance. By systematically varying the downsampling ratio and analyzing the corresponding outcomes, we aimed to identify the optimal configuration that strikes a balance between reducing class imbalance and preserving the essential patterns and information within the data.

   Given the significance of the downsampling ratio in shaping our training data, careful consideration and thorough experimentation were conducted to determine the most effective value for $r$. Through this iterative process, we aimed to achieve an optimal downsampling strategy that effectively addresses the class imbalance while maximizing the predictive power and generalizability of our model.

3. **Classifier**: Within our pipeline, the classification machine learning model plays a central role in our study. The choice of classifier, denoted as $C$, significantly impacts the model's predictive capabilities and overall performance. To facilitate a comprehensive comparative analysis, we have carefully selected three distinct classifiers: Random Forest classifier, XGBoost Classifier, and Multi-layer Perceptron classifier.

   By utilizing these three classifiers, we aim to comprehensively assess their performance and determine the most suitable model for our forest fire prediction task. Through rigorous experimentation and evaluation, we will examine various performance metrics, ROC_AUC, sensitivity, and specificity, to gain insights into the strengths and limitations of each classifier. This comparative analysis will enable us to make informed decisions regarding the selection of the most effective classifier that yields optimal results in predicting forest fire occurrences.

### 4.2.4 Cross Validation Folds

The performance evaluation of our model on the validation data involves utilizing the Cross Validation technique, as depicted in box 6 of Fig. 4.1. In this study, we employ a 10-fold cross-validation approach to assess the effectiveness of our model.

To ensure a fair and reliable evaluation, the data is divided into ten equally sized subsets, or "folds." Each fold is stratified to maintain a balanced representation of fire and non-fire samples, addressing the data's inherent class imbalance. This stratification ensures that each fold contains a similar ratio of fire to non-fire instances, enabling a robust assessment of the model's performance.

During the cross-validation process, the model is trained on nine folds while the remaining fold is used for validation. This procedure is repeated ten times, with each fold serving as the validation set once. By aggregating the results from the ten iterations, we obtain a comprehensive evaluation of the model's performance across different subsets of the data. By employing the 10-fold stratified cross-validation technique, we aim to obtain accurate and reliable performance estimates for our forest fire prediction model. This approach allows us to assess the model's ability to generalize to unseen data and provides valuable insights into its overall effectiveness.

### 4.2.5 Parameter Grid

The Grid Search method is a powerful technique employed in our study to identify the optimal parameters for our model. By defining a parameter grid, we can systematically explore various combinations of hyperparameters and fine-tune our model for optimal performance. In our case, the undersampling ratio stands as a critical hyperparameter that addresses the class imbalance challenge.

Within the parameter grid, we include a range of undersampling ratios to be evaluated. This allows us to assess the impact of different downsampling ratios on the performance of our model. Through an extensive search over the parameter grid, we can identify the undersampling ratio that yields the best results, effectively mitigating the effects of class imbalance on our forest fire prediction model.

To evaluate the performance of our model under different undersampling ratios, we conducted experiments specifically with the Random Forest Classifier. The selected undersampling ratio was iteratively tuned and evaluated to determine its impact on the model's performance. Once the optimal downsampling ratio was determined for the Ran-

| Hyperparameter | Values |
|---|---|
| n_estimators | [200, 400, 600, ..., 1800, 2000] |
| max_features | ['sqrt', 'auto'] |
| max_depth | [50, 60, 70, .., 200] |
| min_samples_split | [ 5, 10] |
| min_samples_leaf | [1, 2, 4] |
| bootstrap | [True, False] |
| sampling_strategy | [0.0, 0.05, 0.1, 0.15, ..., 0.95, 1.0] |

Figure 4.4: Parameter Grid

dom Forest Classifier, we proceeded to compare the performances of the XGBoost and Multilayer Perceptron Classifiers using the same undersampling ratio.

To visualize the search space and the values explored for the Random Forest Classifier with Near Miss version 3 undersampling, we present Fig. 4.4. This figure illustrates the range of undersampling ratios considered during the grid search, highlighting the values examined to identify the best-performing configuration for our forest fire prediction model.

By employing the Grid Search method and evaluating different undersampling ratios, we aim to fine-tune our model and select the most effective configuration for accurate forest fire prediction. This systematic approach allows us to make informed decisions regarding the choice of undersampling ratio, optimizing the performance of our model in handling imbalanced data.

### 4.2.6 Fitting model in Grid Search

The training process involves training the classifier on the grid of parameters. Each possible value within the parameter grid is used to train the classifier. This training procedure is conducted on the training set, and the performance of the classifier is evaluated on the validation set. The output of this process is the identification of the best performance achieved along with its corresponding parameters. This allows us to determine the optimal combination of parameter values that yield the highest performance for the model.

The model's performance is analyzed across different parameter values. If satisfactory, we proceed; otherwise, we adjust the parameters and evaluate performance. A major focus

is on the downsampling ratio. Upon achieving good performance at a specific ratio ($r$), nearby values are explored for further analysis. Once optimal performance is attained, the model is trained on the complete dataset using best-performing parameters, as represented by box 9 in Fig. 4.1.

The outcome of this process is the trained model, which incorporates the best possible parameters identified through training on the training data. Subsequently, the model is subjected to testing on the test data, and its performance is carefully evaluated and analyzed.

This model can be tested on the test data and performance can be analyzed. Various results can be analyzed and conclusions can be drawn from this performance. The results and observations from the modeling are discussed in Chapter 5.

## 4.3 Comprehensive Approach for Imbalance Handling

The table presented in Tab. 4.2 showcases the results of applying various techniques to address data imbalance in forest fire data. The table provides information on the effectiveness of each step in reducing the data imbalance and achieving a more balanced dataset. Let's examine the table and draw insights from the data:

Step 1. Change Spatial Resolution: Before implementing this technique, the data imbalance ratio was 84,000:1, indicating a significant imbalance between the majority and minority classes. However, after changing the spatial resolution from 1km by 1km to 10 km by 10 km, the data imbalance ratio improved to 995:1, representing a substantial decline of approximately 98%. This suggests that modifying the spatial resolution significantly contributed to reducing the data imbalance.

| Handling Data Imbalance | Before | After | Effectiveness |
|---|---|---|---|
| Step 1: Change Spatial Resolution | 84,000:1 | 995:1 | 98% decline |
| Step 2: Data Spatio-subsampling | 995:1 | 806:1 | 19% decline |
| Step 3: Near Miss 3 Undersampling | 806:1 | 20:1 | 97% decline |

Table 4.2: Steps to Deal with Imbalance in Forest Fire Data

Step 2. Data Spatio-subsampling: The second step involved data spatio-subsampling, which further reduced the data imbalance. The initial imbalance ratio of 995:1 decreased to 806:1, resulting in a decline of approximately 19%. This indicates that the chosen subset of spatially balanced data contributed to a more equitable representation of the majority and minority classes.

Step 3. Near Miss 3 Undersampling: The final step utilized the Near Miss 3 undersampling technique, resulting in a remarkable decline in the data imbalance. The imbalance ratio decreased from 806:1 to 20:1, representing a significant decline of approximately 97%. This indicates that the Near Miss 3 undersampling technique effectively addressed the data imbalance issue.

Overall, the table highlights the effectiveness of the applied techniques in reducing data imbalance in forest fire data. The progressive steps of changing spatial resolution, data spatio-subsampling, and Near Miss 3 undersampling r esulted in substantial declines in the data imbalance ratio. By achieving a more balanced dataset, these techniques contribute to improving the reliability and performance of predictive models for forest fire prediction.

It is important to note that the presented effectiveness percentages reflect the relative decline in data imbalance for each specific technique after the previous technique is already applied. The results demonstrate the importance of carefully handling data imbalance to enhance the performance and generalizability of machine learning models used in forest fire prediction.

# Chapter 5

# Experiments and Results

The Experiments and Results chapter presents a comprehensive analysis of the performance and outcomes of the developed forest fire prediction models. This chapter encompasses various aspects, including a comparative evaluation of different models, an assessment of performance using different downsampling techniques and ratios, an ablation study to identify the significance of individual components and a discussion on the challenges encountered during the development process. Through a systematic and rigorous experimentation approach, this chapter sheds light on the strengths, weaknesses, and overall effectiveness of the models, providing valuable insights into their capabilities and limitations. The findings and outcomes presented in this chapter contribute to a deeper understanding of forest fire prediction methodologies and lay the groundwork for future advancements in this field.

## 5.1 Comparison of various models performance

Three distinct machine learning models were trained and evaluated for their performance, namely Random Forest, XGBoost, and Multilayer Perceptron models. This section covers the research question RQ6 of contribution C3. The ROC Curves of the classifiers are shown in three classifiers represented in fig. 5.1a, 5.1b, and 5.1c. The results of the comparative analysis, as presented in Tab. 5.1, highlight the varying performance of these models. Notably, XGBoost demonstrated the highest performance with a ROC-AUC score of 0.872. This indicates that XGBoost outperformed both the Random Forest and Multilayer Perceptron models in effectively distinguishing between ignition and non-ignition samples. The favorable scores in terms of sensitivity and specificity further emphasize the performance of the XGBoost model in representing both ignition and non-ignition data. It

achieved the highest number of accurate predictions for fire ignition and the lowest number of missed ignition predictions.

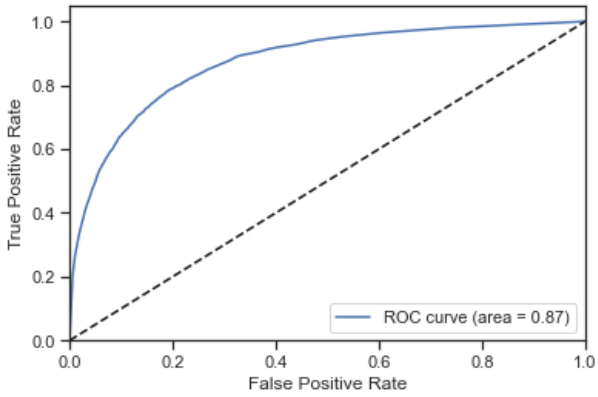
On the other hand, the performance of Random Forest exhibited a slightly lower ROC-AUC score compared to XGBoost. While the sensitivity of Random Forest, representing the rate of correctly predicted fire ignitions, was lower than that of XGBoost, the model showed the highest specificity among the three models. This indicates its ability to predict a significant proportion (89%) of the non-fire cells correctly. Given the focus on improving the prediction of fire cells, XGBoost emerges as the preferred model over Random Forest.

| Model Name | ROC AUC | Sensitivity | Specificity | Correct Ignitions | Missed Ignitions |
|---|---|---|---|---|---|
| Random Forest | 0.869 | 0.66 | 0.89 | 2077 | 1043 |
| **XGBoost** | **0.872** | **0.75** | **0.83** | **2325** | **795** |
| Multilayer Perceptron | 0.769 | 0.04 | 1 | 129 | 2991 |

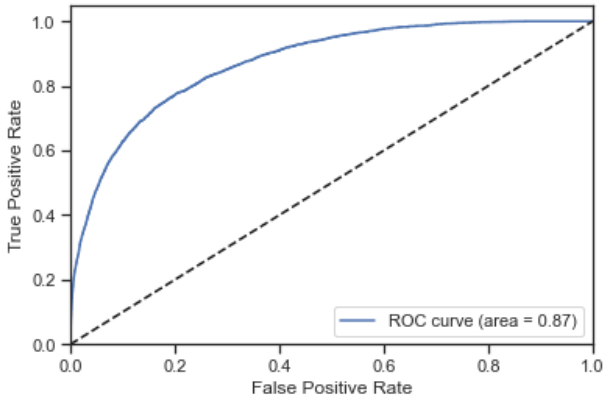
Table 5.1: Comparison of performance of different classifiers

In contrast, the Multilayer Perceptron model displayed limitations in effectively classifying fire and non-fire samples, as evidenced by its lowest ROC-AUC score among the three models. The classifier only managed to predict a small percentage (4%) of the total ignition cells, which is crucial for the specific use case. However, the model exhibited nearly 100% specificity, reflecting its ability to accurately represent non-ignition cells while disregarding fire ignition cells. The performance in correctly predicting non-ignitions and missed ignitions further supports this observation. The suboptimal performance of the Multilayer Perceptron model may be attributed to the high data imbalance and the limited capability of a three-layered network to effectively represent imbalanced data.

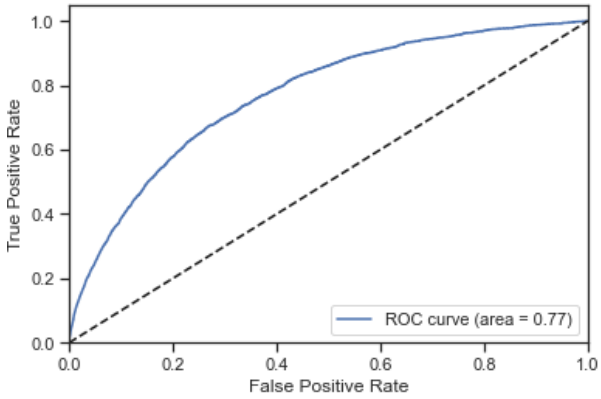
These findings highlight the varying performance and capabilities of the different models in the context of forest fire prediction. The superior performance of XGBoost, followed by Random Forest, suggests the importance of utilizing advanced ensemble learning techniques in handling imbalanced datasets and achieving accurate predictions. The limitations observed in the Multilayer Perceptron model underscore the need for further exploration and experimentation to improve its effectiveness in imbalanced data scenarios.




(a) Random Forest ROC Curve




(b) XGBoost ROC Curve




(c) Multilayer Perceptron ROC Curve

Figure 5.1: ROC Curve of Classifiers

## 5.2 Methodological Considerations

During the training of a machine learning model, numerous methodological considerations come into play. Particularly in the case of imbalanced data, thoughtful decisions must be made regarding the selection of an appropriate undersampling technique and the optimal undersampling ratio for our specific dataset. These choices carry substantial weight in influencing the overall performance and effectiveness of our model. Furthermore, performing feature selection and evaluating feature importance allows us to discern which attributes contribute most significantly to the model's classification performance. Within the realm of our analysis, it is imperative to prioritize the prediction of forest fires, specifically the accurate classification of ignition samples, over the prediction of non-ignition cells. To ensure well-informed decisions, we undertook a comprehensive comparative analysis encompassing a diverse range of parameter options, carefully assessing their impact and ultimately selecting the most suitable choices for our final model. This section covers the research questions RQ4, RQ5 of contribution C2.

### 5.2.1 Decide downsampling technique

In order to address the issue of imbalanced data, we employed downsampling techniques to reduce the number of non-ignition samples in our dataset. The selection of an appropriate downsampling technique is of utmost importance, as different methods can significantly impact the performance of our classifier. To conduct this analysis, we utilized the Random Forest model, applying four downsampling techniques as Random Undersampling, Near Miss 1, Near Miss 2, and Near Miss 3 undersampling techniques to our training data and evaluating the model's prediction performance on the test data.

| Down sampling | ROC AUC | Sensitivity | Specificity | Correct Ignitions | Missed Ignitions |
|---|---|---|---|---|---|
| Random | 0.90 | 0.33 | 1.00 | 1019 | 9095 |
| Near Miss 1 | 0.549 | 0.89 | 0.05 | 2792 | 328 |
| Near Miss 2 | 0.371 | 0.31 | 0.47 | 961 | 2159 |
| **Near Miss 3** | **0.869** | **0.66** | **0.89** | **2077** | **1043** |

Table 5.2: Different downsampling techniques on Random Forest

- The analysis of downsampling techniques using Tab. 5.2 revealed distinct performance variations among the methods. Notably, the Near Miss version 3 technique emerged as the most effective approach, achieving a high ROC-AUC score of approximately 87%. This signifies the model's ability to accurately represent both ignition and non-ignition samples, demonstrating its robustness in capturing the underlying patterns and characteristics of forest fire occurrences.

- In contrast, the Near Miss version 2 technique exhibited the poorest performance among the examined downsampling methods, with a significantly low ROC-AUC score of 37%. This result indicates that the model's ability to distinguish between ignition and non-ignition samples was severely compromised. The model's classification decisions were less reliable and deviated from the desired performance, rendering it less suitable for forest fire prediction.

- The Near Miss version 1 technique demonstrated high sensitivity score. While it exhibited a high sensitivity score of 89%, indicating its proficiency in accurately classifying ignition samples, its specificity score was extremely low at 5%. This suggests that the model predominantly classified a substantial portion of cells as ignition samples, regardless of their true class, leading to a high number of false positives. Consequently, this technique may not provide a well-balanced representation of both ignition and non-ignition samples.

- The Random Undersampling technique, despite displaying a high specificity score of 1.00, showed limited effectiveness in predicting fire occurrences. Its performance was primarily focused on accurately classifying non-fire cells while exhibiting reduced sensitivity and overall performance in capturing ignition samples. As a result, this technique may not be suitable for achieving a balanced and comprehensive representation of both classes in the dataset.

The observed disparities in performance among the downsampling techniques highlight the critical importance of selecting an appropriate method for handling imbalanced data in forest fire prediction. The Near Miss version 3 technique, with its superior performance, offers promising potential for enhancing the performance and reliability of forest fire prediction models. The results emphasize the significance of carefully considering the downsampling approach and its impact on the model's ability to accurately classify and predict fire occurrences.

### 5.2.2 Decide downsampling ratio

The composition of the training dataset is significantly influenced by the chosen downsampling ratio (2.2). As the composition of the dataset directly impacts the model's classification capabilities, the selection of an appropriate downsampling ratio becomes crucial when dealing with imbalanced data. In order to determine the optimal downsampling ratio, we conducted a comparative analysis of the random forest model's performance on Near Miss 3 downsampled data using different downsampling ratios, as depicted in Fig. 5.2 and 5.3.

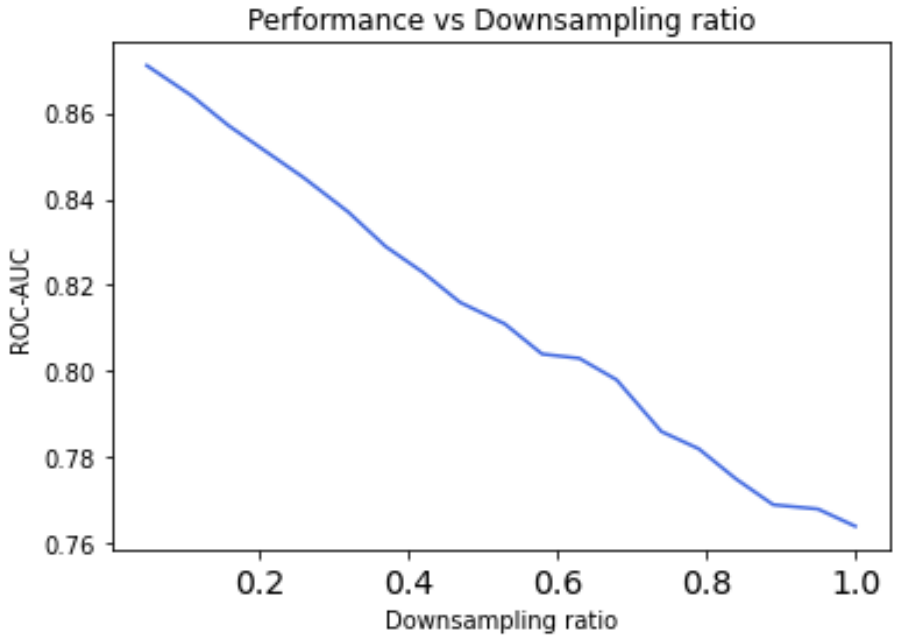



Figure 5.2: Performance vs Downsampling ratio from 0 to 1

Initially, we selected 20 downsampling ratios at equal intervals between 0 and 1. We evaluated the corresponding ROC-AUC scores and aimed to identify the downsampling ratio that yielded the highest performance. Interestingly, we observed a decreasing trend in the random forest model's classification score as the downsampling ratio increased to 1. When the downsampling ratio was set to achieve an equal number of fire and non-fire samples (i.e., $r = 1$), the ROC-AUC score was approximately 0.76. Higher ROC-AUC scores were observed on the lower values of the downsampling ratio of 0.01 to 0.1 as shown in Fig. 5.2. Therefore, we experimented with a downsampling ratio of 0.01 to 0.1 to find the best value of the downsampling ratio as shown in Fig. 5.3.

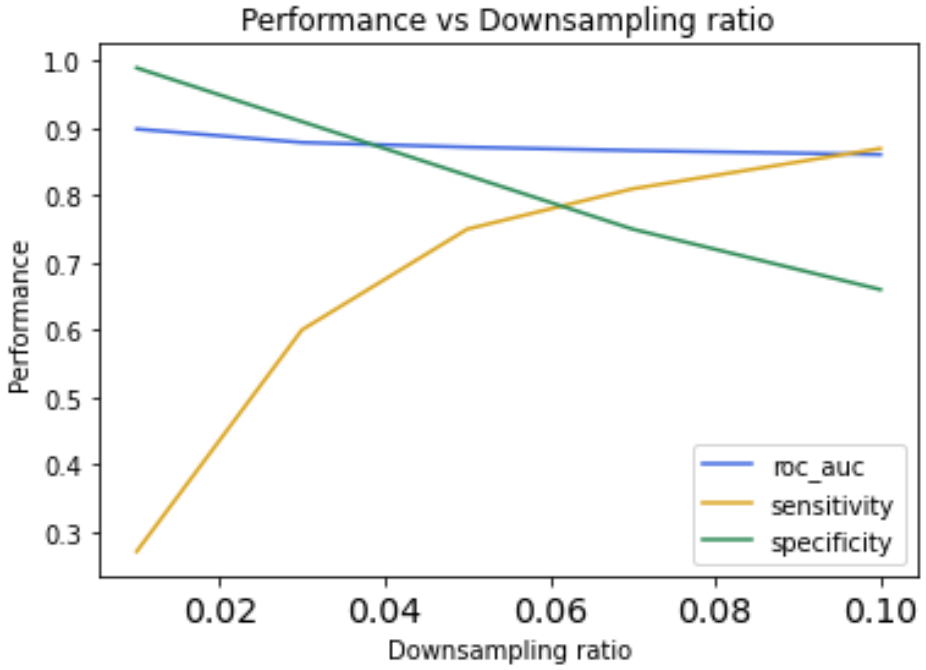



Figure 5.3: Performance vs Downsampling ratio from 0.01 to 0.1

Fig. 5.3 presents the detailed classification score report for downsampling ratios ranging from 0.01 to 0.1. It is important to note that different downsampling ratios yield varying results in terms of sensitivity, specificity, and ROC-AUC scores. While a downsampling ratio of 0.01 achieves the highest ROC-AUC score of 0.89, a closer analysis reveals that it fails to effectively represent the ignition cells, with only 27% of the fire samples being classified correctly. Although the model performs well in classifying non-fire samples (99% specificity), the low sensitivity suggests a limitation in identifying fire samples. As the downsampling ratio increases from 0.01 to 0.1, the ROC-AUC score declines slightly. Considering the trade-off between sensitivity and specificity, the downsampling ratios of 0.05 and 0.07 provide comparable ROC-AUC scores with reasonably acceptable values of sensitivity and specificity. Ultimately, the selection of the best downsampling ratio depends on expert opinion and the specific use case. In our case, we have chosen a downsampling ratio of 0.05, which allows the model to classify 75% of fire samples and 83% of non-fire samples effectively.

### 5.2.3 Importance of Downsampling

The initial analysis of the model's performance on the original data reveals that it is unable to effectively represent the fire samples, with only 14% and 6% of fire samples being correctly predicted for Random Forest and XGBoost, respectively (see Tab. 5.3 and 5.4). This poor performance can be attributed to the high-class imbalance in the original data. However, by applying undersampling techniques to decrease the imbalance in the training data, we observe significant improvement in the model's predictive capability, as demonstrated in the aforementioned tables. Undersampling the training data proves to be an effective approach to enhance the model's ability to accurately classify both the majority and minority classes.

| Down sampling | Imbalance ratio | ROC AUC | Sensitivity | Specificity | Correct Ignitions | Missed Ignitions |
|---|---|---|---|---|---|---|
| Before | **806:1** | 0.77 | **0.14** | 1.00 | 440 | 2680 |
| After | **20:1** | 0.869 | **0.66** | 0.89 | 2077 | 1043 |

Table 5.3: Effect of undersampling on Random Forest prediction performance

| Down sampling | Imbalance ratio | ROC AUC | Sensitivity | Specificity | Correct Ignitions | Missed Ignitions |
|---|---|---|---|---|---|---|
| Before | **806:1** | 0.91 | **0.06** | 1.00 | 184 | 2936 |
| After | **20:1** | 0.872 | **0.75** | 0.83 | 2325 | 795 |

Table 5.4: Effect of undersampling on XGBoost prediction performance

## 5.3 Ablation Study: Importance of each feature

An ablation study assesses the influence of removing individual data features on the performance of Random Forest and XGBoost models. Fig. 5.4 and Fig. 5.5 present the results, where each label represents the model's performance after excluding a specific feature (e.g., month) from the dataset and the data label represents the performance on the complete dataset. Refer to Tab. 5.5 for the abbreviations of the labels. In terms of overall performance on the complete dataset, XGBoost outperforms Random Forest in all performance metrics. The ROC-AUC scores of both models exhibit similar behavior. However, the contribution of each model in predicting fire and non-fire cells varies when different variables are removed from the data. This section covers the research questions RQ7 and RQ8 under contribution C3 and RQ3 of contribution C1.

For XGBoost, the sensitivity is generally better than the Random Forest model for all variables except for meteorological data ($Met$). On the other hand, Random Forest shows higher specificity than XGBoost for all variables except for $Met$. The removal of meteorological data results in a decrease in ROC-AUC for both models. However, for Random Forest, meteorological data contributes more to the prediction of non-fire cells compared to XGBoost.

The best performance for the Random Forest model is achieved when the $avgFeat$ variables are removed. The variables that show a significant dip in ROC-AUC are $fireHis$ and $loc$. Without $fireHis$, the model's ability to predict fire cells decreases from 66% to 56%, highlighting the contribution of $fireHis$ data to fire ignition prediction. Without $loc$, the model's ability to predict fire cells slightly increases from 66% to 68%, but there is a substantial decrease of 25% in the ability to predict non-fire cells. This demonstrates that the grid cell location helps the model better classify non-fire cells.

For XGBoost, the best performance is achieved when using the complete data. The variables that lead to a major dip in ROC-AUC are $fireHis$ and $loc$. Without $fireHis$, the model's ability to predict fire cells decreases from 75% to 67%, indicating the contribution of $fireHis$ data to fire ignition prediction. The specificity of the model also decreases by 2%. Without $loc$, the model's ability to predict fire cells increases from 75% to 81%, but there is a 21% dip in the ability to predict non-fire cells. This underscores the importance of the grid cell location in the model's classification of non-fire cells. Without meteorological data ($Met$), the overall ROC-AUC is lower, but the sensitivity of the model is higher and the specificity is lower. This indicates that meteorological data contributes more to the prediction of non-fire cells.

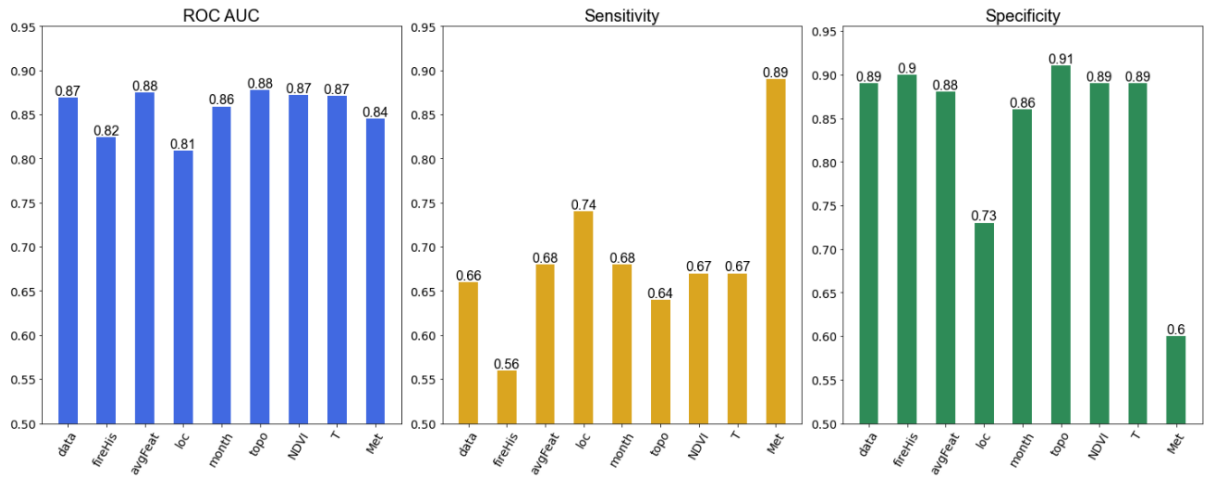



Figure 5.4: Ablation Study for Random Forest.

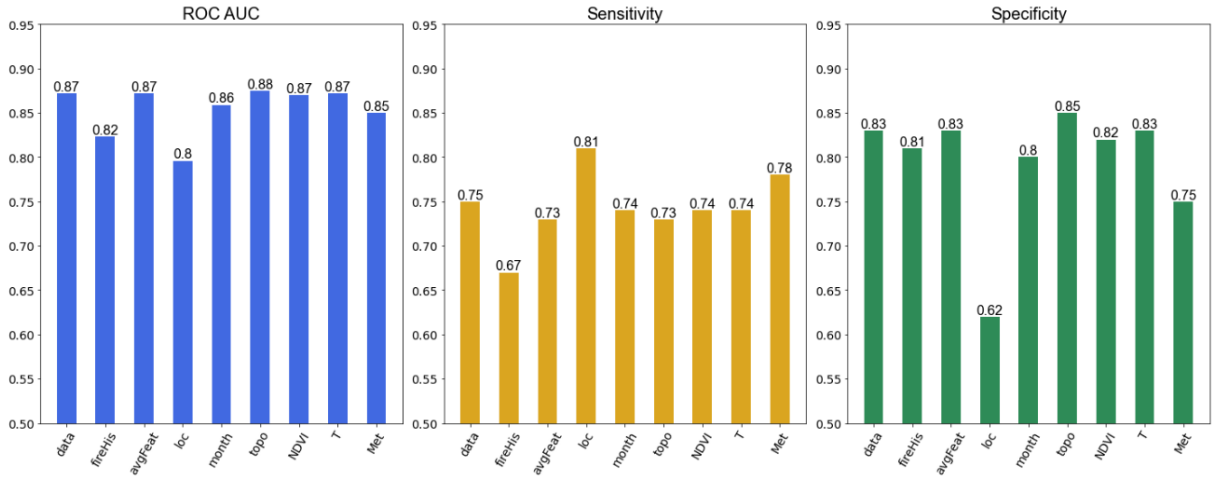



Figure 5.5: Ablation Study for XGBoost.

The presented tables in 5.6 and 5.7 showcase the results of an ablation study conducted on forest fire data using Random Forest (RF) and XGBoost (XGB) models. Each table represents the performance metrics for the models when different attributes are removed from the dataset. The tables provide insights into the importance of individual attributes and their impact on the models' prediction capabilities. By examining the changes in metrics such as ROC AUC, sensitivity, specificity, correct ignitions, and missed ignitions, we can gain valuable insights into the significance of each attribute and the overall performance

| Key | Value |
|---|---|
| data | complete dataset |
| fireHis | Fire History |
| AMT | Average Maximium Temperature |
| AWS | Average Wind Speed |
| TP | Total Precipitation |
| avgFeat | AMT, TP, AWS |
| loc | location (Grid_id) |
| month | month |
| topo | slope, elev, asp |
| NDVI | Normalized Difference Vegetation Index |
| T | Temperature at 2m |
| Met | MaxT, Ws, Prcp |

Table 5.5: Abbreviations

| Data | ROC AUC | Sensitivity | Specificity | Correct Ignitions | Missed Ignitions |
|---|---|---|---|---|---|
| data | 0.869 | 0.66 | 0.89 | 2077 | 1043 |
| fireHis | 0.824 | 0.56 | 0.90 | 1736 | 1384 |
| avgFeat | 0.875 | 0.68 | 0.88 | 2136 | 984 |
| loc | 0.809 | 0.74 | 0.73 | 2308 | 812 |
| month | 0.859 | 0.68 | 0.86 | 2122 | 998 |
| topo | 0.878 | 0.64 | 0.91 | 2005 | 1115 |
| NDVI | 0.872 | 0.67 | 0.89 | 2086 | 1034 |
| T | 0.871 | 0.67 | 0.89 | 2100 | 1020 |
| Met | 0.845 | 0.89 | 0.60 | 2774 | 346 |

Table 5.6: Random Forest Feature Importance

of RF and XGB models in predicting forest fire occurrences.

1. Significance of $fireHis$ attribute: - After removing the $fireHis$ feature, both RF and XGB models show comparable ROC AUC scores. However, XGB demonstrates better sensitivity by correctly predicting more fire ignitions and having fewer missed ignitions compared to RF. This indicates that XGB is more effective in capturing the patterns and characteristics of fire events, making it a preferable choice for forest fire prediction.

2. Significance of $avgFeat$ attribute: - For RF, the removal of the $avgFeat$ feature leads to significant improvements across all performance metrics. The ROC AUC score increases, indicating better overall model performance. Moreover, sensitivity improves, resulting in

| Data | ROC AUC | Sensitivity | Specificity | Correct Ignitions | Missed Ignitions |
|---|---|---|---|---|---|
| data | 0.872 | 0.75 | 0.83 | 2325 | 795 |
| fireHis | 0.823 | 0.67 | 0.81 | 2096 | 1024 |
| avgFeat | 0.872 | 0.73 | 0.83 | 2283 | 837 |
| loc | 0.796 | 0.81 | 0.62 | 2518 | 602 |
| month | 0.859 | 0.74 | 0.80 | 2295 | 825 |
| topo | 0.875 | 0.73 | 0.85 | 2263 | 857 |
| NDVI | 0.870 | 0.74 | 0.82 | 2312 | 808 |
| T | 0.872 | 0.74 | 0.83 | 2319 | 801 |
| Met | 0.850 | 0.78 | 0.75 | 2432 | 688 |

Table 5.7: XGBoost Feature Importance

more accurate fire ignition predictions and fewer missed ignitions. This highlights that $avgFeat$ does not help RF to make better predictions. Conversely, the removal of $avgFeat$ has minimal impact on XGB, suggesting that average features have a relatively better influence on XGB's performance compared to RF.

3. Significance of *loc* attribute: - The *loc* attribute plays a crucial role in both RF and XGB models. Removing this attribute significantly decreases the ROC AUC score for both models, indicating a substantial drop in overall predictive performance. Additionally, the removal of *loc* leads to increased sensitivity. However, this increase in sensitivity comes at the expense of very low specificity, suggesting that the *loc* attribute is essential for distinguishing non-fire instances. Therefore, retaining the *loc* attribute is crucial for accurate forest fire prediction.

4. Significance of *month* attribute: - Removing the *month* attribute has a minimal effect on RF's ROC AUC score, indicating that the model's overall performance remains relatively stable. However, it slightly improves the number of correct ignition predictions and decreases the number of missed ignitions, suggesting that the *month* attribute may have limited influence on RF's fire prediction capabilities. On the other hand, the removal of *month* in XGB results in a slight decrease in all performance metrics, indicating that the *month* attribute plays a more significant role in the XGB model's ability to accurately predict forest fires.

5. Significance of *topo* attribute: - The *topo* attribute demonstrates its importance in

both RF and XGB models. In RF, removing the *topo* attribute leads to a decrease in correct ignition predictions, despite a slight improvement in the ROC AUC score. This implies that topographical data contribute to RF's ability to accurately classify fire events. Similarly, in XGB, the removal of *topo* results in a decrease in correct ignition predictions, suggesting that topographical information aids XGB in making more accurate fire predictions. Therefore, retaining the *topo* attribute is beneficial for both models.

6. Significance of $NDVI$ attribute: - Removing the $NDVI$ attribute shows varying effects on RF and XGB models. For RF, the removal of $NDVI$ yields the best performance, as indicated by an improved ROC AUC score. This suggests that the $NDVI$ data has limited contribution to RF's predictive capabilities and may even introduce noise into the model. In contrast, for XGB, the removal of $NDVI$ leads to a slight decrease in all performance metrics, indicating that the $NDVI$ attribute aids XGB in making more accurate fire classifications. Therefore, the impact of $NDVI$ differs between the two models, with RF benefiting from its removal and XGB relying on it for improved performance.

7. Significance of $T$ attribute: - The impact of the $T$ attribute varies between RF and XGB models. Removing the $T$ attribute improves the overall performance of the RF model, suggesting that this attribute does not contribute significantly to RF's fire prediction capabilities. On the contrary, for XGB, the presence of the $T$ attribute maintains the same ROC AUC and specificity as the baseline model. However, there is a slight drop in sensitivity and correct ignition predictions when $T$ is removed, indicating that the $T$ attribute assists XGB in accurately predicting fire ignitions. Therefore, the influence of $T$ differs between the two models, with RF showing no significant impact and XGB benefiting from its inclusion.

8. Significance of $Met$ attribute: - The $Met$ attribute plays a critical role in both RF and XGB models. Removing the $Met$ attribute reduces the ROC AUC score for both models, indicating a decrease in overall predictive performance. Although the removal leads to more correct ignition predictions and higher sensitivity, it also results in very low specificity, implying a high false positive rate. This highlights the importance of meteorological data in accurately identifying non-fire instances. Therefore, retaining the $Met$ attribute is crucial for achieving reliable forest fire prediction.

In summary, the ablation study provides valuable insights into the impact of different features on the performance of RF and XGB models for forest fire prediction. These observations demonstrate the varying importance of each feature and their effects on different performance metrics. The results highlight the significance of the location, average features, topography, month, NDVI, temperature, and meteorological data in improving the models' fire prediction capabilities.

## 5.4 Computation time analysis

In the comparison of CPU time between Random Forest (RF) and XGBoost (XGB) for different ablation datasets as represented in Tab. 5.8, it was observed that the CPU time for RF was generally lower than that of XGB, except for the Temperature row. This discrepancy in CPU time can be attributed to several factors related to the algorithms and their implementation. The CPU time was measured for the training of the model was measured using the Python time command.

RF is an ensemble-based algorithm that builds multiple decision trees in parallel, utilizing features randomly sampled at each node. This parallel nature allows RF to take advantage of multi-core processing and distribute the workload efficiently, resulting in faster computation and lower CPU time. While it is generally believed that XGB is faster due to its optimization techniques and parallelization, our observations indicate otherwise.

| Data | RF CPU Time | XGB CPU Time |
|---|---|---|
| data | 43.8 s | 1min 21s |
| fireHis | 38.7 s | 1min 21s |
| avgFeat | 34.3 s | 1min 20s |
| loc | 36.1 s | 1min 28s |
| month | 41.8 s | 1min 23s |
| topo | 38.7 s | 1min 21s |
| NDVI | 35.5 s | 1min 12s |
| T | 56.2 s | 38 s |
| Met | 13.6 s | 38.8 s |

Table 5.8: CPU time for Random Forest and XGBoost

One potential explanation for the lower CPU time of RF could be the inherent parallelization of RF, which allows it to take advantage of multiple CPU cores more efficiently. On the other hand, XGB is a gradient-boosting algorithm that builds decision trees sequentially, where each tree learns from the mistakes of the previous tree. This sequential nature limits the inherent parallelism of XGB during the training process, making it relatively slower than RF in terms of CPU time. Additionally, RF may benefit from optimized implementation in software libraries, resulting in faster computations.

It is worth noting that the exception observed in the Temperature row, where XGB had lower CPU time than RF, could be attributed to the characteristics of the specific dataset or the complexity of the feature being ablated. XGB might have been able to exploit certain patterns or optimize its computations more effectively in that particular scenario, resulting in faster CPU time.

Further analysis is required to fully comprehend the factors influencing the relative computational efficiency of RF and XGB. Nonetheless, these findings challenge the common assumption that XGB is consistently faster and emphasize the importance of considering various factors that can impact CPU time in different scenarios.

## 5.5 Challenges encountered during development

In this section, we discuss the various challenges encountered during the development process of our wildfire prediction model. These challenges encompassed multiple aspects, including spatial resolution alignment, dataset availability, imbalance ratio of the data, and the management of a large volume of data. We delve into each of these challenges, providing insights into the complexities faced and the strategies employed to overcome them. By addressing these challenges, we aimed to ensure the reliability and effectiveness of our model, laying the foundation for robust experimentation and reliable results.

### 5.5.1 Spatial Resolution Changes

This section focuses on the issues related to spatial resolution encountered during the data collection phase and its impact on the imbalance ratio of the dataset. Initially, a spatial resolution of 1 km by 1 km was chosen for the data collection process. Consequently, all the collected data was converted to this resolution and combined, resulting in a substantial dataset of approximately 288 GB. However, the imbalance ratio of the data was found to be extremely high at 84,000:1, posing a significant challenge for subsequent analysis.

Attempts were made to address the class imbalance by applying undersampling techniques; however, due to the large volume of data, the computation required for undersampling was time-consuming. Moreover, the results obtained after undersampling were unsatisfactory, as existing downsampling techniques were not specifically designed to handle such high levels of imbalance. To explore alternative solutions, a thorough analysis of the data was conducted, leading to the realization that increasing the spatial resolution could offer potential benefits.

Increasing the resolution would result in a decrease in the imbalance between fire and non-fire samples. Since fire occurrences are rare events, the number of fires within a cell remains the same regardless of the cell size, whether it is 1 km by 1 km or increased to 10 km by 10 km. However, increasing the resolution would reduce the number of non-fire cells, thereby reducing the class imbalance. It is crucial to strike a balance when increasing the resolution to avoid situations where multiple fire events fall within the same cell, as this would decrease the overall number of fire cells and complicate the imbalance ratio.

Considering the geographical context of the Alberta region and the forested areas, a resolution of 10 km by 10 km was deemed reasonable. This resolution accounts for the area considered while maintaining an appropriate level of detail for the forested regions. Additionally, it was observed that the decided resolution of the final data should not exceed the resolution of any individual dataset being combined. For instance, the meteorological data available had an approximate resolution of 10 km by 10 km. Setting the final resolution at 1 km by 1 km would involve duplicating values from a 10 km by 10 km cell into 100 1 km by 1 km cells. This would diminish the quality of the dataset, as multiple cells would share the same values for meteorological features, posing challenges for the model in accurately classifying fire and non-fire cells. This insight highlighted the importance of ensuring that the final resolution of the data does not fall below the resolution of any individual dataset present.

### 5.5.2 Imbalance Ratio of Data

It is essential to address the significant class imbalance observed in the forest fire prediction dataset. The occurrence of wildfire is inherently rare, while the collection of climatic and other factors is conducted regularly throughout the year for various locations. Consequently, this leads to an imbalanced dataset where the number of non-fire instances significantly outweighs the number of fire instances. The presence of such a substantial class imbalance poses a critical challenge in developing accurate forest fire prediction systems. The available balancing techniques are often insufficient to effectively handle such a vast disparity in class distribution. Even after employing undersampling techniques, the resulting model struggles to achieve satisfactory recall for both fire and non-fire classes. The extent of this imbalance is so pronounced that it almost resembles an anomalous behavior within the dataset.

### 5.5.3 Large Volume of Data

During the course of our work with the dataset encompassing the entire region of Alberta, we initially encountered a substantial volume of data of approximately 288 GB, characterized by a spatial resolution of 1 km by 1 km. However, conducting undersampling techniques and implementing various machine learning methods on this dataset proved to be resource-intensive, demanding computational capabilities beyond our available resources. Subsequently, we devised a solution to address the issue of data volume by increasing the spatial resolution to 10 km by 10 km. This adjustment resulted in a significant reduction in the number of rows per year, by a factor of 100. As a result, we obtained a more manageable dataset size of 2.46 GB comprising samples collected over a span of 18 years. Due to the immense size of the 1 km by 1 km resolution dataset, certain distance-based undersampling techniques, such as centroid undersampling, could not be effectively analyzed as they strained the computational limitations of our system.

### 5.5.4 Availability of data

The availability of meteorological data posed a significant challenge during the modeling process. The data we obtained was sourced from weather stations located throughout the Alberta region. However, the distribution of these weather stations varied across different subregions, resulting in varying densities of stations. Some areas had a higher concentration of weather stations, while others had larger distances between them. To predict forest fires accurately, we required comprehensive data coverage for the entire region at a specific spatial resolution. Since weather stations were not evenly distributed, it necessitated the interpolation of data for locations where no weather station was present. This interpolation process involved both spatial and temporal aspects, wherein data had to be interpolated from the nearest weather stations for each day. Identifying the most appropriate interpolation method for this task presented an additional research challenge. We are grateful to Alberta ACIS [2] for providing us with interpolated weather data, which enabled us to mitigate this challenge to some extent. However, it is important to note that relying solely on interpolated values for forest fire predictions may not capture the true climatic conditions accurately. To achieve more precise predictions, a greater number of weather stations recording actual climatic conditions would be desirable.

# Chapter 6

# Conclusion & Future Work

This chapter marks the culmination of our study on forest fire prediction using heterogeneous data sources and machine learning. In this chapter, we summarize the key findings and contributions of our research, discuss the implications of our results, and provide insights into potential avenues for future exploration. Additionally, we reflect on the importance of our work in addressing the challenges of forest fire prediction and discuss its potential impact on wildfire management strategies. This chapter serves as a comprehensive conclusion to our study, while also setting the stage for further advancements in the field of forest fire prediction.

## 6.1 Future Work

In terms of future research directions, there are several promising avenues to explore based on the findings of this study. Firstly, the dataset used in this study holds potential for the development of models capable of predicting future climatic conditions and their impact on forest fire ignition. By integrating advanced forecasting techniques with the dataset, we can gain valuable insights into the relationships between weather patterns and fire occurrence, enabling more accurate predictions.

Furthermore, there is scope to expand the dataset by incorporating additional types of data sources. For instance, considering the potential impact of anthropogenic factors such as electric wire spread in forests can provide valuable information for enhancing the prediction of human-caused forest fires. Incorporating such data sources can enrich the model's understanding of the various factors contributing to forest fire ignitions, leading

to more robust and comprehensive predictions. Also, the data collection framework used for Alberta can be scaled to collect data for the whole country using a similar approach.

In addition, deep learning models present a promising avenue for future investigation. By leveraging the dataset's time-series nature, models such as Long Short-Term Memory (LSTM) can be employed to capture and analyze temporal patterns in the data. LSTM models have shown success in capturing long-term dependencies, making them well-suited for predicting forest fires based on historical weather conditions. Exploring the potential of LSTM models and other deep-learning approaches can further enhance the performance of fire ignition predictions.

Moreover, it is worth exploring alternative data balancing techniques to better understand the behavior of wildfire data and improve wildfire prediction models. Different undersampling and oversampling methods can be applied to assess their impact on model performance and identify the most effective approach for addressing the imbalanced nature of the dataset. Such investigations can provide valuable insights into the optimal data balancing strategies for forest fire prediction.

Furthermore, the availability of data at higher resolutions can significantly enhance the predictive capabilities of the models. By incorporating data from a greater number of weather stations or other relevant sources, we can capture more fine-grained and localized information, allowing for more accurate predictions of fire occurrences in specific regions. This would enable planners and stakeholders to make proactive decisions and allocate mitigation resources effectively.

Finally, An intriguing avenue to explore is the application of federated learning techniques for forest fire prediction. By adopting federated learning, forest fire prediction models can benefit from the diverse and heterogeneous data collected by various stakeholders. Each participating entity can contribute its local data, encompassing meteorological measurements, topographical features, historical fire records, and other relevant variables. The federated learning framework enables the aggregation of these distributed datasets to build a robust and accurate prediction model.

Overall, these future research directions have the potential to advance the field of forest fire prediction, enhancing our ability to forecast fire occurrences, mitigate their impact, and ultimately safeguard ecosystems and human lives.

## 6.2 Conclusion

In this thesis, we have delved into the challenging problem domain of forest fire prediction, aiming to develop accurate and reliable models using heterogeneous data sources and machine learning methods. Forest fires pose a significant threat to ecosystems, human lives, and the economy, necessitating proactive measures for effective firefighting, resource allocation, and risk assessment. However, the complexity and dynamic nature of forest fires, coupled with the challenges associated with data collection, imbalance in the dataset, and the scarcity of high-resolution weather data for specific regions, have hindered the development of robust prediction models. Through a comprehensive investigation guided by research questions, valuable insights have been gained, significant challenges have been addressed, and substantial contributions have been made to the field.

The exploration began by delving into the underlying factors contributing to forest fire ignition, as highlighted in research question RQ1. Through an extensive literature review and analysis, we identified key factors such as meteorological conditions, biophysical characteristics, topographical attributes, and fire history. This comprehensive understanding served as the foundation for our subsequent research endeavors and played a pivotal role in the development of a robust data collection framework. This framework integrated data from diverse sources, including remote satellites and weather stations, into a unified database. By capturing essential attributes at a high temporal and spatial resolution, we created a comprehensive dataset spanning 18 years and covering a vast region of 661,848 $km^2$ at the spatial resolution of 10 $km$ by 10 $km$ in Alberta. The successful implementation of this data collection framework effectively addresses research question RQ2 and constitutes a significant contribution C1 to this thesis.

One of the major challenges encountered in forest fire prediction is the imbalanced distribution of fire and non-fire samples in the dataset. Research question RQ4 focused on understanding the challenges associated with wildfire data and devising strategies to address them. To tackle the imbalance issue, we devised a comprehensive approach consisting of three steps: spatial resolution modification, data spatio-subsampling, and the implementation of the Near Miss-3 undersampling technique. These techniques effectively mitigated the imbalance and improved the performance of our predictive models. Through thorough experimentation and evaluation, we demonstrated the effectiveness of our approach in handling the high imbalance inherent in forest fire prediction data. By successfully addressing research questions RQ4 and RQ5, we have contributed a valuable solution to the field of imbalanced dataset handling, evidenced by contribution C2.

Machine learning models play a crucial role in predicting forest fire ignitions, and research question RQ6 aimed to identify the most suitable methods for this task. We evaluated various machine learning algorithms, with a particular focus on ensemble models such as XGBoost and Random Forest. The performance evaluation yielded impressive results, with XGBoost emerging as the top-performing model, achieving a ROC-AUC score of 87% and a sensitivity of 75%. These findings effectively answer research question RQ6, showcasing the efficacy of ensemble models in addressing the challenges posed by imbalanced data in forest fire prediction. Furthermore, our ablation study provided valuable insights into the contribution of different attributes, revealing the significance of fire history, grid cell location, and meteorological data in enhancing the models' predictive capabilities. Thus, our research contributions C3 shed light on the importance of attributes selected for accurate forest fire prediction, addressing research questions RQ3, RQ7, and RQ8.

The composition of the dataset and the spatial resolution of the data play pivotal roles in determining the performance and reliability of forest fire prediction models. Our study has demonstrated the substantial impact of modifying the dataset's composition and imbalance ratio on classification performance. However, the limited availability of data at an appropriate resolution for the target region presents a significant challenge. In addressing this, we have utilized interpolated data from nearby weather stations to supplement our dataset, although further advancements can be made. Hence, we emphasize the importance of expanding the network of weather stations within the desired area to enhance data quality and strengthen the predictive capabilities of the models. This development will pave the way for more accurate and reliable forest fire predictions, empowering decision-making processes and facilitating proactive measures for effective fire prevention and management.

In summary, this thesis has made substantial contributions to the field of forest fire prediction by addressing research questions, overcoming challenges, and leveraging heterogeneous data sources and machine learning methods. The comprehensive understanding of underlying factors, the development of a robust data collection framework, the effective handling of imbalanced data, and the selection of appropriate machine learning models have significantly advanced the performance and reliability of forest fire prediction. Our research findings provide valuable insights for proactive risk assessment, robust mitigation strategies, and the preservation of ecosystems. Continued research efforts, focusing on dataset refinement, data resolution enhancement, and the exploration of emerging technologies, will further enhance our ability to forecast and prevent forest fires, ultimately leading to improved safety, resource allocation, and environmental conservation.

# References


[1] Gis-based spatial prediction of tropical forest fire danger using a new hybrid machine learning method. *Ecological Informatics*, 48:104–116, 2018.

[2] Forestry Alberta Agriculture and Rural Economic Development (AFRED). Interpolated weather data since 1901 for alberta townships. Available at https://acis.alberta.ca/acis/township-data-viewer.jsp (accessed 30 August 2022).

[3] Statistics Canada. Corrected representation of the ndvi using historical avhrr satellite images (1 km resolution) from 1987 to 2021. Available at https://open.canada.ca/data/en/dataset/44ced2fa-afcc-47bd-b46e-8596a25e446e (accessed 10 August 2022).

[4] Copernicus. Fires, forests, and the future: A crisis raging out of control? Available at https://wwfeu.awsassets.panda.org/downloads/wwf_fires_forests_and_the_future_report.pdf (accessed 10 May 2023).

[5] Mario Elia, Marina D'Este, Davide Ascoli, Vincenzo Giannico, Giuseppina Spano, Antonio Ganga, Giuseppe Colangelo, Raffaele Lafortezza, and Giovanni Sanesi. Estimating the probability of wildfire occurrence in mediterranean landscapes using artificial neural networks. *Environmental Impact Assessment Review*, 85:106474, 2020.

[6] Keke Gao, Zhongke Feng, and Shan Wang. Using multilayer perceptron to predict forest fires in jiangxi province, southeast china. *Discrete Dynamics in Nature and Society*, 2022:1–12, 06 2022.

[7] Umang Garg, Vineet Kukreti, Rahul Singh Pundir, Mahesh Manchanda, and Neha Gupta. Prediction of turkey forest fire using random forest regressor. In *2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA)*, pages 1–7, 2023.

[8] Haoyuan Hong, Paraskevas Tsangaratos, Ioanna Ilia, Junzhi Liu, A-Xing Zhu, and Chong Xu. Applying genetic algorithms to set the optimal combination of forest fire related variables and model forest fire susceptibility based on data mining models. the case of dayu county, china. *Science of The Total Environment*, 630:1044–1056, 2018.

[9] Piyush Jain, Sean C.P. Coogan, Sriram Ganapathi Subramanian, Mark Crowley, Steve Taylor, and Mike D. Flannigan. A review of machine learning applications in wildfire science and management. *Environmental Reviews*, 28(4):478–505, dec 2020.

[10] Ali Karouni, Bassam Daya, and Pierre Chauvet. Applying decision tree algorithm and neural networks to predict forest fires in lebanon. *Journal of Theoretical and Applied Information Technology*, 63:282–291, 05 2014.

[11] Can Lai, Shucai Zeng, Wei Guo, Xiaodong Liu, Yongquan Li, and Boyong Liao. Forest fire prediction with imbalanced data using a deep neural network method. *Forests*, 13:1129, 07 2022.

[12] Yudong Li, Zhongke Feng, Shilin Chen, Ziyu Zhao, and Fengge Wang. Application of the artificial neural network and support vector machines in forest fire prediction in the guangxi autonomous region, china. *Discrete Dynamics in Nature and Society*, 2020:5612650, Apr 2020.

[13] J. Muñoz Sabater. Era5-land hourly data from 1950 to present. copernicus climate change service (c3s) climate data store (cds). Available at https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-land?tab=form (accessed 2 August 2022), 2019.

[14] Khurram Nadeem, S.W. Taylor, Douglas Woolford, and C. Dean. Mesoscale spatiotemporal predictive models of daily human- and lightning-caused wildland fire occurrence in british columbia. *International Journal of Wildland Fire*, 29, 01 2019.

[15] United States Geological Survey (USGS). (n.d.). Earth Explorer. Usgs earth explorer. Available at https://earthexplorer.usgs.gov/ (accessed 5 August 2022).

[16] Earth Resources Observation and Science (EROS) Center. Usgs eros archive - digital elevation - shuttle radar topography mission (srtm) 1 arc-second global. Available at https://www.usgs.gov/centers/eros/science/usgs-eros-archive-digital-elevation-shuttle-radar-topography-mission-srtm-1 (accessed 5 August 2022).

[17] Government of Alberta. Alberta census boundaries current. Available at https://open.alberta.ca/opendata/gda-4d939041-851b-4848-bd30-44dbf129e16c (accessed 2 June 2023).

[18] Government of Alberta. Wildfire maps and data. Available at https://www.alberta.ca/wildfire-maps-and-data.aspx (accessed 2 June 2023).

[19] The Conference Board of Canada. The economic impact of the fort mcmurray fires. Available at https://www.conferenceboard.ca/product/the-economic-impact-of-the-fort-mcmurray-fires/ (accessed 10 May 2023).

[20] Biswajeet Pradhan, Mohd Dini Hairi Suliman, and Mohamad Awang. Forest fire susceptibility and risk mapping using remote sensing and geographical information systems (gis). *Disaster Prevention and Management*, 16:344–352, 06 2007.

[21] Dedi Rosadi, Deasy Arisanty, and Dina Agustina. Prediction of forest fire using neural networks with backpropagation learning and exreme learning machine approach using meteorological and weather index variables. *MEDIA STATISTIKA*, 14:118–124, 01 2022.

[22] Ranak Thakkar, Varad Abhyankar, Polaka Divya Reddy, and Surya Prakash. Environmental fire hazard detection and prediction using random forest algorithm. In *2022 International Conference for Advancement in Technology (ICONAT)*, pages 1–4, 2022.

[23] Zili Wang, Binbin He, and Xiaoying Lai. Balanced random forest model is more suitable for wildfire risk assessment. In *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, pages 3596–3599, 2022.

[24] WWF and Boston Consulting Group. Fires, forests, and the future: A crisis raging out of control? 2020. Available at https://wwfeu.awsassets.panda.org/downloads/wwf_fires_forests_and_the_future_report.pdf (accessed 10 May 2023).

[25] Wenkai Yan, Jie Ren, Jianyuan Feng, Yi Duan, and Changning Wei. A new forest fire risk rating forecast model based on xgboost. In *2022 International Seminar on Computer Science and Engineering Technology (SCSET)*, pages 227–230, 2022.

[26] Suwei Yang, Massimo Lupascu, and Kuldeep S. Meel. Predicting forest fire using remote sensing data and machine learning, 2021.

[27] J. Zhang and I. Mani. KNN Approach to Unbalanced Data Distributions: A Case Study Involving Information Extraction. In *Proceedings of the ICML'2003 Workshop on Learning from Imbalanced Datasets*, 2003.