Order number: . . . Option: …

UNIVERSITY OF MOHAMED BOUDIAF – MSILA

FACULTY OF MATHMATICS AND COMPUTER SCIENCE
DEPARTEMENT OF COMPUTER SCIENCE

Master in **Computer science**

By

Bechere M'hamed Ayoub

Zamit Zohir

Prediction Model for Forests Fire Spread in M'sila

Under the supervision of

Dr. Tahar Mehenni

Composition of the jury

Kamel Mohamed	University of Msila	President
Mehenni Tahar	University of Msila	Supervisor
Loucif Hamza	University of Msila	Reporter

September 2023

DEDICATIONS

*"We would like to dedicate this thesis to **our parents**, **family**, and **friends**, whose unwavering support, encouragement, and love have been our driving force throughout this journey. Their constant belief in us, even in the face of challenges, has been a source of inspiration and motivation. We are immensely grateful for their guidance and sacrifice, and we owe them everything. This accomplishment would not have been possible without their love and unwavering support. We dedicate this work to them with love and appreciation."*

Bechere M'hamed Ayoub, Zamit Zohir

ACKNOWLEDGMENTS

We would like to express our sincere gratitude to “Dr. Tahar Mehenni”, our supervisor, for their invaluable guidance, support, and encouragement throughout the entire process of conducting this research and writing this thesis.

We are also grateful to the members of the jury, “Kamel Mohamed” and “Loucif Hamza”, for taking the time to review and evaluate our work, and for their constructive feedback and insightful comments.

We would like to extend our thanks to the University of “Mohamed Boudiaf – Msila”, for providing us with the necessary resources and facilities to carry out this research.

TABLE OF CONTENT:

TABLE OF ILLUSTRATIONS:

INTRODUCTION

Natural disasters such as earthquakes, tornadoes, floods, and forest fires pose a significant threat to both the environment and human lives. Among these, wildfires have emerged as one of the most complex and devastating disasters facing our communities. Studies at a global scale have indicated a steady increase in the frequency of fires, leading to catastrophic consequences for both human life and biodiversity. Consequently, governments, donors, and non-governmental organizations have recognized the need for new strategies and solutions to address this crisis. Fire prevention is widely acknowledged as a crucial technique to mitigate the dangerous consequences of wildfires.

In light of these challenges, scientists have devoted significant effort to develop effective strategies and policies for predicting fire incidents and managing their spread. Artificial intelligence and machine learning have emerged as valuable tools in this domain, offering the potential to expedite tasks and reduce human efforts. Consequently, in this thesis, we propose an approach based on supervised machine learning for predicting forests fire spread. Our model is evaluated using data collected from M'sila forest fires over the past few years.

The structure of this dissertation is as follows:

In Chapter 1, we provide a comprehensive examination of forest fires, including their definition, types, characteristics, causes, contributing factors, consequences, and prevention methods. We also present statistics related to this disaster, with a particular focus on the situation in Algeria.

Chapter 2 serves as an introduction to the field of artificial intelligence and machine learning, exploring its various branches and characteristics.

Chapter 3 delves into the details of our study, presenting a thorough description of our approaches and the proposed architecture. We also provide a comprehensive explanation of the operational methodology underlying our proposal.

Finally, in Chapter 4, we analyze the performance results of our proposed models and compare them with other approaches. We conclude the dissertation with a summary of our approach and outline potential avenues for future research.

Overall, this research aims to contribute to the advancement of forests fire spread prediction using machine-learning techniques. By understanding, the characteristics of forests fire spread, leveraging artificial intelligence, and employing a supervised machine learning approach, we seek to enhance our ability to forecast and manage forests fire spread effectively.

CHAPTER 1

FORESTS FIRE

INTRODUCTION

Forests fire, also known as wildfires, are uncontrolled fires that occur in forests, grasslands, and other wildland areas. These fires can be caused by natural causes or by human activities. Forest fires can cause extensive damage to the natural environment. They can also pose a significant threat to human life and property in nearby communities. In this chapter, we will discuss the forest fires, the causes and the factors of this phenomenon.

1. Definitions

1.1. Fire

Fire is a rapid chemical process of combustion, characterized by the release of heat, light, and various gases, typically including carbon dioxide and water vapor. It is a natural phenomenon that occurs when a combustible material combines with oxygen in the presence of heat or an ignition source, resulting in a self-sustaining reaction known as a fire. The process involves the breaking down of complex organic molecules into simpler compounds, accompanied by the release of energy in the form of heat and light. [1]

1.2. Forest

Forests are vast areas of land dominated by trees and other woody vegetation, forming a complex ecosystem. They play a crucial role in maintaining biodiversity, regulating climate, conserving water, providing habitat for numerous species, and supplying valuable resources such as timber and non-timber forest products. Their high density of trees, diverse plant and animal communities, and a variety of ecological interactions characterizes forests. [2]

1.3. Forest fire

Forest fires, also known as wildfire, are uncontrolled and unplanned fires that occur in forested areas or adjacent vegetation. They are characterized by the rapid and extensive burning of trees, shrubs, grasses, and other organic materials within the forest ecosystem. Forest fires can be ignited by natural causes such as lightning strikes or by human activities, including arson or accidental ignition. They can spread quickly, fueled by dry vegetation, strong winds, and favorable weather conditions. [3]

2. Forests role in Forests Fire

Forests play a critical role in forests fire, both in terms of their susceptibility to fire and their ability to influence the spread and severity of fires.

Forests are particularly vulnerable to fire because they are composed of large amounts of dry vegetation and other flammable materials. When these materials become too dry, they can easily catch fire and spread quickly.

2.1. Forest formations

Forest formations refer to distinct types or categories of forests based on their structure, composition, and ecological characteristics. Specific combinations of tree species, vegetation density, canopy structure, and environmental conditions define them. Forest formations can vary globally and are influenced by factors such as climate, soil type, elevation, and historical disturbances. [4]

They can be classified into different types based on their geographic location, dominant tree species, and climate.

2.1.1. Tropical Forests

Tropical forests are a type of forest ecosystem found in the Earth's tropical regions, typically between the Tropic of Cancer and the Tropic of Capricorn. They are characterized by high levels of biodiversity, dense vegetation, and a warm and humid climate throughout the year. Tropical forests are home to a wide variety of plant and animal species, many of which are unique to these regions. [5]

2.1.2. Temperate Forests

Temperate forests are a type of forest ecosystem that occurs in regions with moderate climate and distinct seasons, typically between the Polar Regions and the tropics. They are characterized by a mix of deciduous and coniferous trees and experience seasonal variations in temperature, precipitation, and daylight hours. Temperate forests are known for their diverse plant and animal species, including mammals, birds, reptiles, and amphibians. [6]

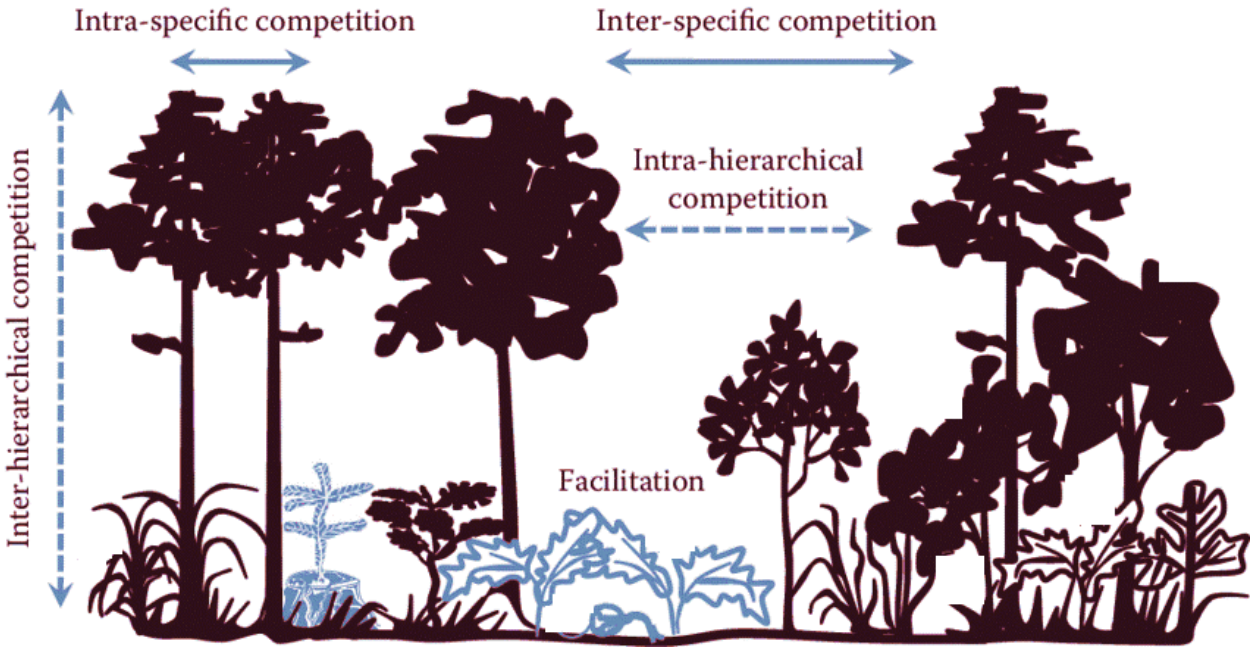

Figure 1.1: The Hierarchical Interaction Levels in Temperate Forest Ecosystems [31]

2.1.3. Boreal Forests

Boreal forests, also known as taiga or snow forests are a type of forest ecosystem that spans the high-latitude regions of the Northern Hemisphere. They are characterized by a cold climate, long and harsh winters, short summers, and a predominance of coniferous trees such as spruce, fir, and pine. Boreal forests are often found in areas with low temperatures and short growing seasons, resulting in slow tree growth and the presence of permafrost in some regions. [7]

Figure 1.2: Illustration of Each Fire Stand Evolves Go Through Boreal Forests [32]

2.1.4. Mediterranean Forests

Mediterranean forests, also referred to as Mediterranean woodlands and scrub, are a type of forest ecosystem found in regions with a Mediterranean climate. These forests are characterized by hot, dry summers and mild, wet winters. They are typically composed of a mixture of evergreen trees, such as oak, pine, cypress, and olive, as well as shrubs and dense undergrowth. [8]

2.1.5. Montane Forests

Montane forests, also known as mountain forests, are a type of forest ecosystem that occurs in mountainous regions at high elevations. They are characterized by cooler temperatures, higher levels of precipitation, and distinct vegetation zones determined by elevation. Montane forests often exhibit a gradient of vegetation types, ranging from coniferous forests at lower elevations to subalpine forests and alpine meadows at higher elevations. [9]

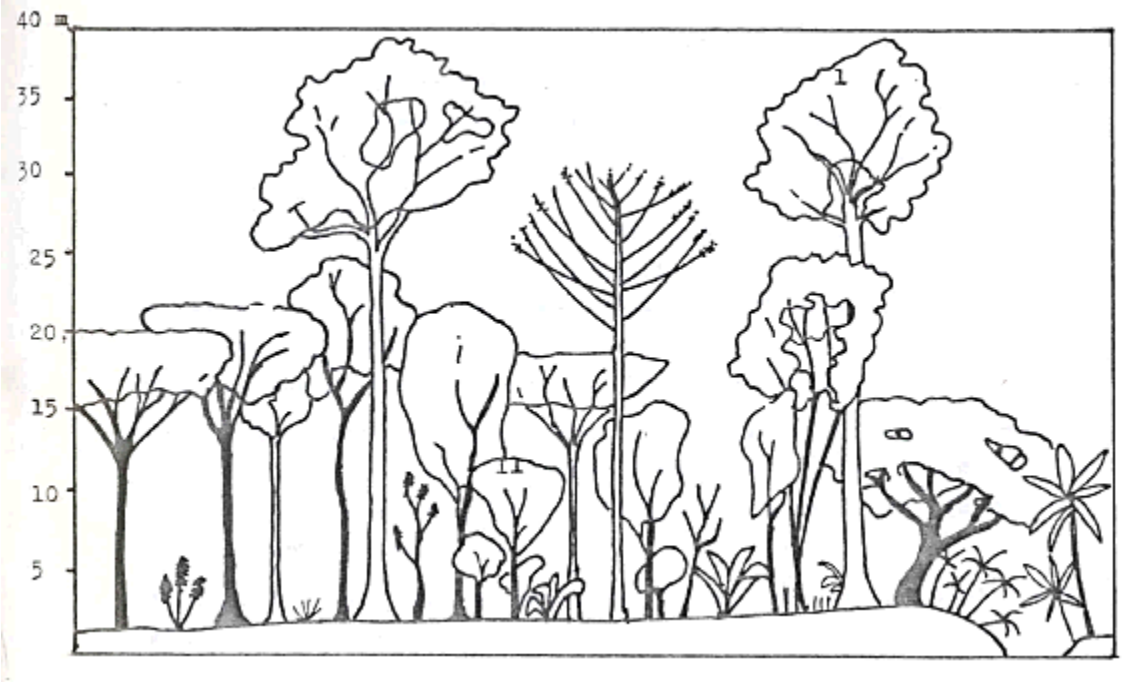

Figure 1.3: Transect through Moist Evergreen Montane Forest [33]

2.2. Small-sized Forest Formations

Small-sized Forest formations refer to forest ecosystems that are characterized by their limited extent or area. These formations can vary in size but generally encompass smaller patches of forested land, often surrounded by other land uses or larger forest ecosystems. They may occur within fragmented landscapes or be remnants of larger forests that have undergone human disturbance or land conversion. [10]

2.2.1. Maquis

Maquis refers to a specific type of Mediterranean shrubland vegetation that is found primarily in the Mediterranean Basin. It is characterized by dense and evergreen shrubs, such as various species of mastic, rockrose, and myrtle, along with scattered small trees like holm oak and cork oak. Maquis vegetation is adapted to the hot, dry summers and mild, wet winters typical of Mediterranean climates. [11]

2.2.2. Garrigue

Garrigue refers to a specific type of Mediterranean vegetation characterized by a sparse and open shrubland ecosystem. It is commonly found in the Mediterranean Basin and is composed of low-growing, drought-resistant shrubs, aromatic herbs, and scattered trees. The dominant plant species in garrigue ecosystems often include lavender, thyme, sage, rosemary, and various species of cistus. Garrigue vegetation is adapted to the hot, dry summers and mild, wet winters typical of Mediterranean climates. [12]

2.2.3. Moorland

Moorland refers to a type of upland habitat characterized by open, treeless expanses of grasses, heather, and low-growing vegetation. It is typically found in cool and wet climates, often in areas with acidic or peaty soils. Moorland ecosystems are common in regions such as the British Isles, Scotland, and parts of northern Europe. They are known for their distinctive landscapes, including rolling hills, bogs, and heathlands. [13]

3. Types of forest fires

Forests fire can be classified into various types based on their cause, location, behavior and can be classified in to:

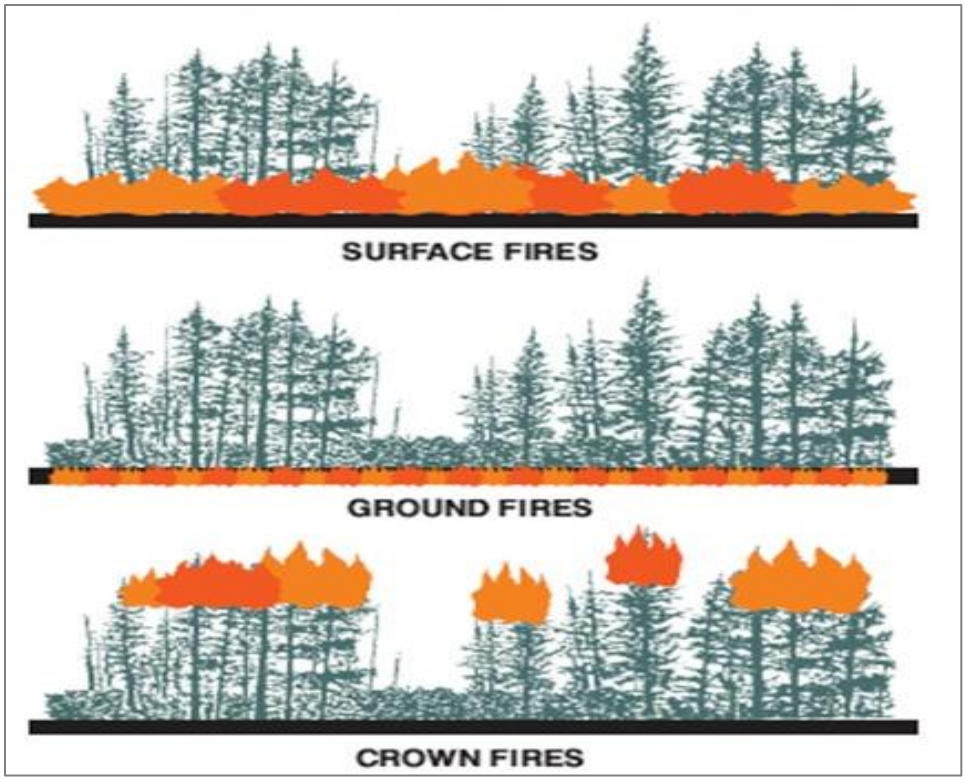

Figure 1.4: Type of Forest Fire [34]

3.1. Surface Fires

Surface fires are a type of wildfire that primarily burns vegetation and other combustible materials on or near the ground surface. Flames that travel through the grass, shrubs, or small trees, consuming the available fuel, characterize them. Surface fires typically have a lower intensity compared to more severe types of wildfires, such as crown fires or ground fires. [14]

3.2. Crown Fires

Crown fires, also known as canopy fires, are a type of wildfire that spreads rapidly through the upper layer or canopy of a forest. Unlike surface fires that primarily burn ground vegetation, crown fires burn through the tree crowns, consuming foliage, branches, and sometimes-entire trees. Intense flames that can generate high heat produce significant amounts of smoke, and cause rapid and extensive damage to the forest ecosystem characterize crown fires. [15]

3.3. Ground Fires

Ground fires, also known as subsurface fires or underground fires, are a type of wildfire that burnt in the organic material and soil layers below the ground surface. Unlike surface fires or crown fires, ground fires typically smolder and burn slowly, consuming the organic matter in the soil, such as leaf litter, roots, and peat. They can be challenging to detect and extinguish, as they often burn deep underground and may go unnoticed until they resurface or cause visible damage. [16]

4. Causes of forest fires

The causes of forest fires can be broadly classified into two categories:

4.1. Natural Factors

Environmental conditions or events that are not caused by human actions or behaviors that contribute to the occurrence or spread of forest fires:

- **Lightning strikes:** Lightning is a common cause of forest fires in areas with dry vegetation. When lightning strikes a tree or the ground, it can ignite a fire that spreads quickly through the surrounding vegetation.
- **Drought and high temperatures:** Drought and high temperatures can cause vegetation to become dry and highly flammable, making it more susceptible to ignition by a spark or flame.
- **Wind:** Strong winds can spread fire quickly and increase its intensity, making it more difficult to control.

4.2. Human Factors

The actions or behaviors of human that can contribute to the occurrence or spread of forest fires:

- **Arson:** Deliberate setting of fires by people, either for personal or criminal purposes, is a significant cause of forest fires.
- **Campfires and outdoor burning:** Unattended campfires or outdoor burning can easily spread and cause forest fires. Improperly discarded cigarettes or fireworks can also start fires.

- **Equipment and machinery:** Sparks from machinery, such as chainsaws or vehicles, can easily ignite dry vegetation and start a forest fire.
- **Power lines:** Power lines can start fires when they become damaged, and the sparks or arcs they produce meet dry vegetation.
- **Land-use practices:** such as deforestation, land clearing, and agriculture can alter the natural fire regime and make forests more susceptible to fires.
- **Climate Change:** Climate change has been linked to the increase in forest fires frequency and severity in many regions. Rising temperatures, changes in precipitation patterns, and other climate-related factors can create conditions that are favorable for forest fires to start and spread.

5. Influential factors in the spread of fire

Several factors can influence the spread of fire, and meteorological conditions and the characteristics of vegetation are two of the most influential factors.

5.1. Meteorological conditions

Meteorological conditions refer to the atmospheric conditions that prevail at a specific time and place, including temperature, humidity, wind speed and direction, precipitation, atmospheric pressure, and other relevant weather variables. These conditions are influenced by factors such as air masses, frontal systems, local topography, and global climate patterns. Meteorological conditions play a significant role in shaping weather patterns, climate, and the occurrence of various weather phenomena. [17]

5.2. The characteristics of vegetation

The characteristics of vegetation refer to the specific traits, features, and attributes of plant life in a particular area or ecosystem. These characteristics can include physical attributes such as height, leaf shape, color, and texture, as well as ecological properties such as growth form, reproductive strategies, adaptation to environmental conditions, and interactions with other organisms. [18]

6. Fire and Its Effects on Forests

Fire is a natural and essential process in many forest ecosystems. It helps to clear out dead and decaying vegetation, promote nutrient cycling, and create new habitats for many plant and animal species. However, fires can also be destructive, especially when they occur outside their natural frequency and intensity.

- **Vegetation Changes:** Fires can lead to changes in the composition and structure of vegetation. Some plant species are adapted to fire and may even require it for regeneration, while others may be negatively impacted or unable to recover.
- **Habitat Modification:** Fires can alter the physical structure of habitats, affecting the availability of resources and habitat suitability for various species. It can create open spaces, promote successional changes, and create diverse ecological niches.
- **Nutrient Cycling:** Fire releases nutrients stored in vegetation and organic matter, making them available for uptake by plants. It can also affect soil chemistry and nutrient availability in the long term.
- **Wildlife Dynamics:** Fires can affect wildlife populations, both directly by causing mortality and indirectly by modifying habitat and food availability. Some species may benefit from post-fire habitats, while others may experience negative consequences.
- **Forest Regeneration:** In some cases, fires can stimulate forest regeneration by initiating seed release, reducing competition, and creating conditions favorable for seed germination and seedling establishment.

7. Methods of prevention

Prevention forest fires is a key to reducing their impact. Many are caused by human's activities such as careless behavior, improper use of fire, and arson. Some of the methods used for preventing forest fires are:

7.1. Fire Prevention Education

Public education campaigns are one of the most effective methods for preventing forest fires. This involves educating the public about the causes and risks of forest fires, as well as promoting responsible behavior in natural environments. This includes simple

actions such as properly extinguishing campfires and disposing of cigarettes and other flammable materials.

7.2. Controlled Burning

Controlled burning is a proactive method of reducing the risk of forest fires. This involves intentionally setting fires under controlled conditions, which can help to reduce fuel loads, decrease the risk of uncontrolled fires, and promote the growth of new vegetation.

7.3. Forest Management:

Proper forest management practices can also help to prevent forest fires. This includes thinning out overgrown forests, creating firebreaks, and removing dead or diseased trees that can serve as fuel for fires.

7.4. Fire Detection and Monitoring

Early detection of forest fires is critical to minimizing their impact. This can be achieved through the use of fire detection systems, such as satellite imagery and sensors, as well as monitoring by trained personnel who can quickly respond to new fires.

7.5. Fire Suppression

When forest fires do occur, rapid and effective suppression is essential to minimizing their impact. This involves deploying trained firefighters, equipment, and aerial support to contain and extinguish fires as quickly as possible.

8. Forest fires in Algeria

In 2021, the total burned area worldwide is 1,113,464 ha, a similar total to that mapped in 2020. The total burned area mapped in North Africa and the Middle East was very similar to 2020 and slightly worse than the long-term average, but with large differences within individual nations. Tunisia had a worse year than 2020, while Libya's season was better. The most affected country in the region was Algeria, accounting for 69% of the total burned area. In Algeria, the total burned area mapped was the highest since 2012. 295 fires were mapped, representing a total burned area of 31,275 ha, two-thirds of which was on agricultural land.

The first fire of the season was mapped in February and the last in November, but 85% of the damage occurred in August. The largest fire of the season exceeded 25,000 ha, and there were 20 other fires over 1,000 ha and 15 over 500 ha.

Land cover	Area burned (ha)	% of total
deciduous forest	20466	15.2
coniferous forest	3171	2.4
mixed forest	105	0.1
Transitional	19711	14.7
Other natural lands	4720	3.5
Agriculture	85934	64.0
Other land cover	6	0.0
Artificial surfaces	161	0.1

Figure 1.5: Distribution of burnt areas (ha) in Algeria by type of land cover in 2021

Conclusion:

The forests fire have been a significant and recurring problem in Algeria, particularly during the dry summer months when the risk of fires is highest. The country's forested areas, including the northern region, have been particularly vulnerable to these fires, which have caused significant damage to the natural environment, property, and infrastructure.

The effects and consequences of forests fire in Algeria are significant, including loss of biodiversity and vegetation, soil erosion, air pollution, property damage, and economic losses. These fires also pose a threat to human lives and safety, as well as disrupting local economies and communities.

Efforts to prevent and control forests fire spread in Algeria have been ongoing, with the government implementing measures such as firebreaks and fire prevention campaigns. However,

there is still a need for increased investment in forest management and prevention strategies to effectively tackle the problem.

In conclusion, the impact of forest fires in Algeria highlights the importance of effective forest management practices and proactive measures to prevent and control fires. Moreover, in the next chapter will define the most important prediction methods and techniques in machine learning.

CHAPTER 2

MACHINE LEARNING

INTRODUCTION

Machine learning is a subfield of artificial intelligence that enables computers to learn from data and make predictions or decisions without being explicitly programmed. It is used to solve complex problems across a wide range of fields, from healthcare to finance to marketing. Machine learning algorithms and techniques are constantly evolving and improving, allowing for more accurate predictions and better decision-making.

In this chapter, we will explore the basics of machine learning, including the importance of data, the steps involved in building a machine-learning model, and some popular algorithms and techniques. We will also delve into some advanced topics, such as deep learning and natural language processing, to highlight the potential of machine learning in solving complex problems.

1. Artificial intelligence

1.1. Definitions

Artificial intelligence (AI) refers to the development and implementation of computer systems or machines that can perform tasks that typically require human intelligence. It involves the simulation of human cognitive processes, such as learning, reasoning, problem solving, and decision-making, by machines.

AI encompasses various subfields, including machine learning, natural language processing, computer vision, robotics, and expert systems. These technologies enable machines to process and analyze large amounts of data, recognize patterns, make predictions, and adapt to changing circumstances. [19]

1.2. Artificial intelligence history

The history of artificial intelligence (AI) dates back to the mid-20th century when researchers began exploring the idea of creating machines that could exhibit human-like intelligence. The field has evolved significantly over the decades, with various milestones and breakthroughs shaping its development. [20]

- **Dartmouth Conference:** (1956): Considered the birth of AI, the Dartmouth Conference brought together leading researchers to explore the possibilities of creating "thinking machines" and laid the foundation for AI as a distinct field of study.

- **Early AI Research** (1950s-1960s): In the following years, researchers focused on developing symbolic AI systems that relied on rule-based programming and logic. Examples include the Logic Theorist, General Problem Solver, and the development of expert systems.

- **Rise and Fall of AI** (1970s-1980s): During this period, AI experienced significant growth, with advancements in areas such as knowledge representation, natural language processing, and expert systems. However, AI faced criticism and a subsequent decline known as the "AI winter" due to high expectations and limited progress.

- **Machine Learning and Neural Networks** (1990s-2000s): The resurgence of AI came with the advent of machine learning algorithms and neural networks. The development of algorithms like backpropagation and the increasing availability of large datasets led to breakthroughs in areas such as computer vision and speech recognition.

- **Big Data and Deep Learning** (2010s-Present): The proliferation of big data and advancements in computational power paved the way for deep learning models, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs). These models have achieved remarkable performance in tasks like image recognition, natural language processing, and game playing.

2. Machine Learning

2.1. Definitions

Machine learning refers to the field of study and practice that involves developing algorithms and models capable of learning from data and making predictions or taking actions without being explicitly programmed. It focuses on developing systems that can automatically improve and adapt their performance based on experience.

Machine learning algorithms analyze large datasets, identify patterns, and build mathematical models that can make predictions or decisions. These algorithms can learn from examples, past experiences, or feedback and they continually refine their models to improve their accuracy and performance. [21]

2.2. Machine Learning History

Machine learning has a rich history that spans several decades, with key developments and milestones shaping its evolution. Here is a brief overview of the history of machine learning [22]

- **Early Foundations** (1950s-1960s): The foundation of machine learning can be traced back to the development of early computing and artificial intelligence. Researchers like Arthur Samuel and Frank Rosenblatt made notable contributions, with Samuel's work on game-playing programs and Rosenblatt's development of the perceptron algorithm.

- **Rule-Based Systems** (1960s-1970s): During this period, researchers focused on developing rule-based systems that could reason and make decisions based on predefined rules. Symbolic AI and expert systems emerged as dominant approaches, with projects like MYCIN and DENDRAL highlighting the potential of rule-based learning systems.

- **Statistical Approaches** (1980s-1990s): The 1980s and 1990s saw a shift towards statistical approaches to machine learning. Researchers explored techniques such as

decision trees, Bayesian networks, and support vector machines. The development of the backpropagation algorithm for training neural networks also occurred during this period.

- **Rise of Big Data** (2000s): The advent of the internet and the proliferation of digital data led to a renewed interest in machine learning. The availability of large datasets facilitated the development of more sophisticated algorithms and techniques. Support for deep learning, ensemble methods, and reinforcement learning grew, leading to significant advancements in various domains.

- **Deep Learning Revolution** (2010s-Present): The past decade has witnessed a revolution in deep learning, a subset of machine learning that focuses on artificial neural networks with multiple layers. The availability of massive computing power, large-scale labeled datasets, and breakthroughs in algorithms have propelled deep learning to achieve exceptional performance in areas such as computer vision, natural language processing, and speech recognition.

3. Supervised learning

Supervised learning is a subfield of machine learning where an algorithm learns from labeled training data to make predictions or decisions. In supervised learning, a dataset is provided that consists of input examples (features) and corresponding output labels. The algorithm's goal is to learn a mapping function that can accurately predict the output labels for new, unseen input examples.

During the training phase, the algorithm analyzes the labeled data and seeks patterns or relationships between the input features and output labels. It builds a model or hypothesis based on these patterns, aiming to generalize from the training data to make accurate predictions on unseen data.

Supervised learning can be further divided into two main categories: classification and regression. In classification, the algorithm learns to assign input examples to predefined

classes or categories. In regression, the algorithm predicts a continuous numerical value based on the input features. [23]

The idea of supervised learning dates back to the 1940s, when researchers began exploring the use of artificial neural networks to model complex systems. Since then, supervised learning has become one of the most widely used and studied areas of machine learning, with applications ranging from image recognition to natural language processing and predictive analytics.

Supervised learning works by first dividing the labeled dataset into a training set and a validation set. The training set is used to train the model, while the validation set is used to evaluate the model's performance on unseen data. The model is typically defined as a mathematical function that maps input features to output labels, and the goal of training is to learn the parameters of the function that best fit the training data.

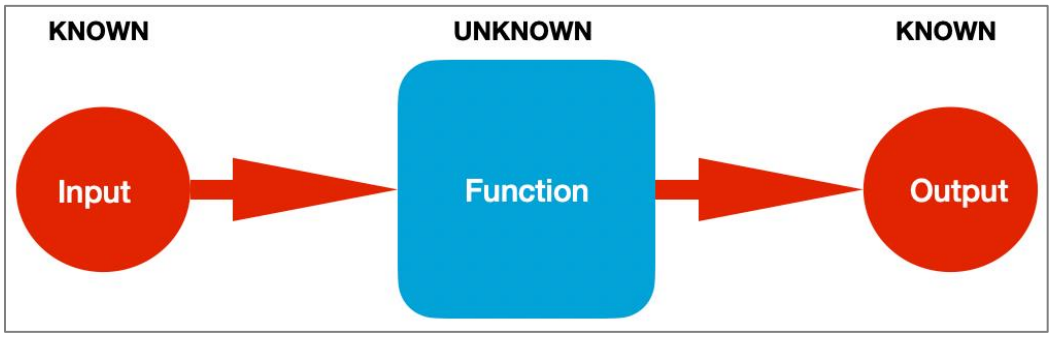

Figure 2.1: The input and output values of Supervised Learning [35]

3.1. Regression

Regression is a subfield that deals with predicting a continuous numerical output or value based on input features. It focuses on estimating the relationship between the input variables and the output variable, allowing the algorithm to make predictions on unseen data.

Regression algorithms aim to learn a mapping function that can capture the underlying patterns or trends in the training data and generalize to make accurate predictions for new input examples.

Unlike classification, where the output is a discrete class or category, regression predicts a continuous value. Examples of regression tasks include predicting house prices based on features like location, size, and number of rooms, or forecasting stock prices based on historical data and relevant indicators.

Regression algorithms can be linear or nonlinear, depending on the nature of the relationship between the input features and the output variable. Linear regression assumes a linear relationship, while nonlinear regression models can capture more complex relationships using techniques such as polynomial regression, decision trees, or neural networks. [24]

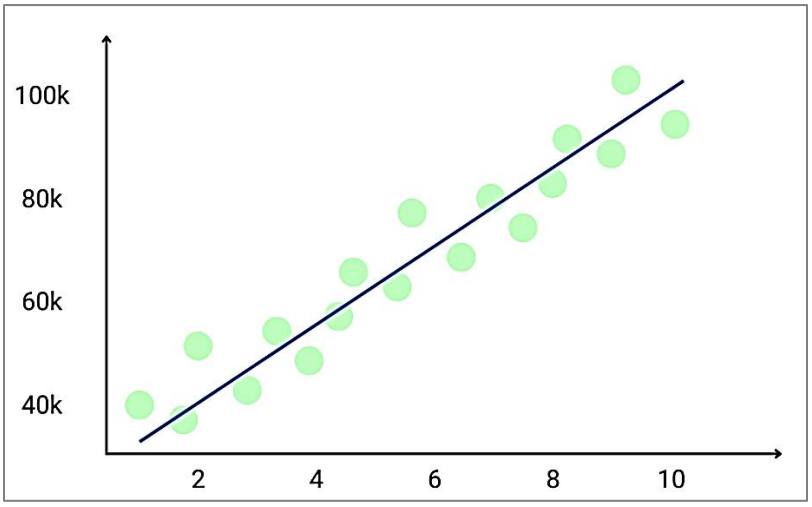

Figure 2.2: Simple linear regression. [36]

3.1.1. Linear regression

Linear regression is used to model the relationship between a dependent variable (also known as the target variable) and one or more independent variables (also known as features or predictors). The goal of linear regression is to find the line of best fit that

describes the relationship between the variables. This line can be used to make predictions about the dependent variable for new values of the independent variable.

Linear regression has its roots in the work of Francis Galton in the 19th century, who developed the concept of regression towards the mean. In the early 20th century, Karl Pearson and Ronald Fisher developed the mathematical framework for linear regression, which is still used today.

Linear regression works by finding the best-fitting line through the data using a method called ordinary least squares (OLS). The OLS method minimizes the sum of the squared differences between the predicted values and the actual values for the dependent variable. This line is called the regression line or the line of best fit.

The formula for linear regression is:

$$\boldsymbol{y = \beta 0 + \beta 1 x1 + \beta 2 x2 + \beta 3 x3 + \cdots + \varepsilon} \quad \textbf{(1)}$$

Where y is the dependent variable, x(x1,x2…,x4) is the independent variable, M(β0, β1, β3,….β4) is the slope of the line, B is the y-intercept and ε is is the error. The slope of the line represents the change in Y for a one-unit increase in x. The y-intercept is the value of Y when x is equal to zero.

A linear regression model has two components: a deterministic part (b1X1+ b2X2 + . . .) and a random part (the error, ε). These two components can be considered as the signal and the noise in the model. If We have only one input variable X, the regression model is the best row that fits the data.

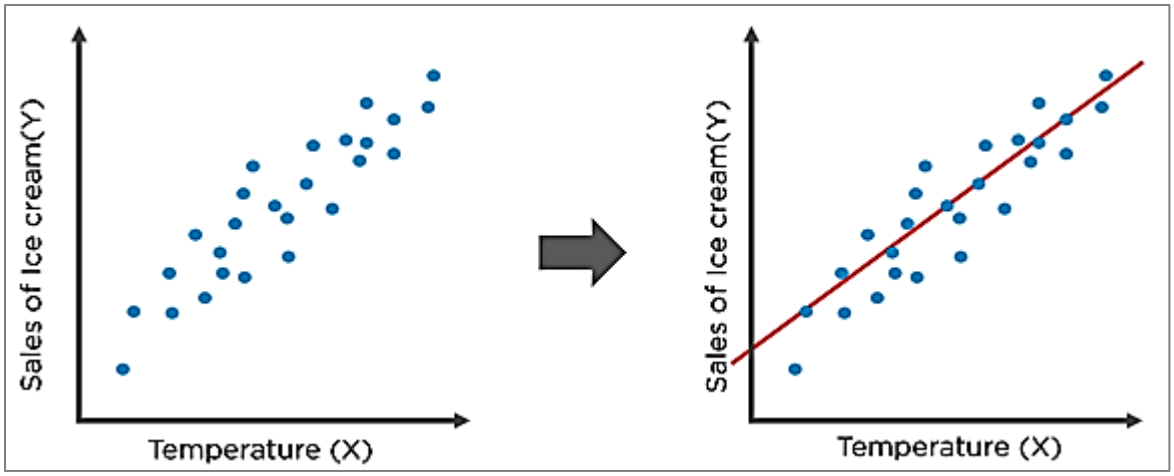

Figure 2.3: Linear Regression [37]

3.1.2. Polynomial Regression

Polynomial regression is a type of regression analysis used to model the relationship between a dependent variable (target) and one or more independent variables (features) when the relationship is not linear but follows a polynomial pattern. In polynomial regression, higher-order polynomial terms are included in the model to capture the non-linearity.

The general formula for polynomial regression is:

$$y = \beta 0 + \beta 1x + \beta 2x^2 + \beta 3x^3 + ... + \beta nx^n + \varepsilon$$

- **Y**: The dependent variable or target variable that you want to predict.

- **X**: The independent variable(s) or feature(s) used for prediction.

- **β0, β1, β2 ... βn**: The coefficients that need to be estimated from the data. These coefficients determine the shape and magnitude of the polynomial terms. β0 is the intercept (the value of y when x is zero), β1 represents the linear coefficient, β2 represents the coefficient for the quadratic term, and so on.

- **x^2, x^3, ... x^n**: These terms represent the polynomial features, where x is raised to different powers (2, 3, ..., n) to capture the non-linear patterns in the data.

- **ε**: The error term, which represents the noise or unexplained variance in the data. It accounts for the discrepancies between the predicted values and the actual observed values.

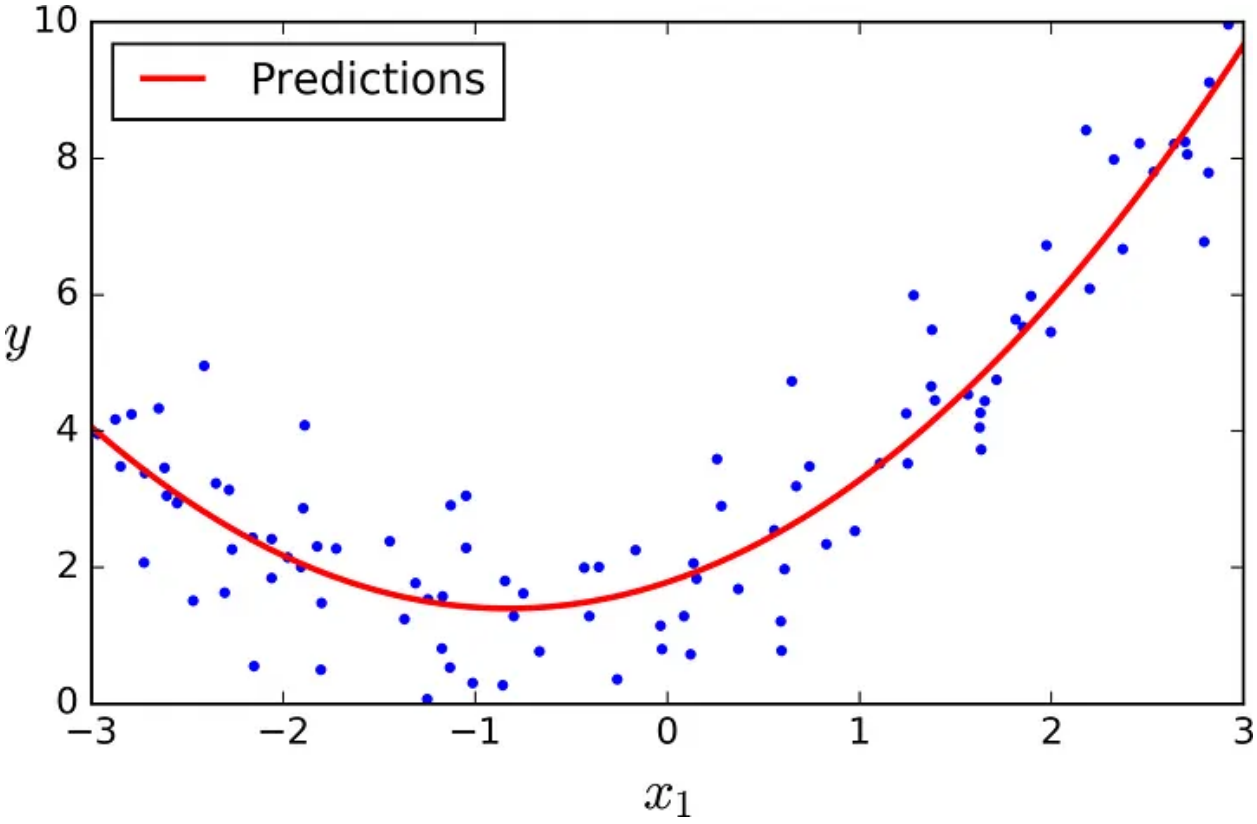

Figure 2.4: Polynomial Regression [64]

The key component of polynomial regression is the creation of polynomial features by raising the independent variable(s) (x) to different powers. These polynomial features are used to model the non-linear relationship.

3.1.3. Decision Trees Regression

Decision tree regression is a machine learning technique used for solving regression problems, where the goal is to predict a continuous numeric output based on input

features. Unlike classification trees that predict categorical labels, decision tree regression predicts numeric values at the leaf nodes of the tree. It works by partitioning the feature space into regions and assigning a constant value to each region.

Like other machine learning techniques, decision tree regression starts with a dataset containing input features and corresponding numeric target values.

To make predictions on new, unseen data, you traverse the decision tree from the root node to a leaf node by following the feature-based decisions at each node. The predicted output for the new data point is the value associated with the leaf node it reaches.

Decision tree regression does not have a single formula like linear regression. Instead, it involves a tree-like structure with nodes and branches, where each node represents a decision based on a feature, and each leaf node represents the predicted output value.

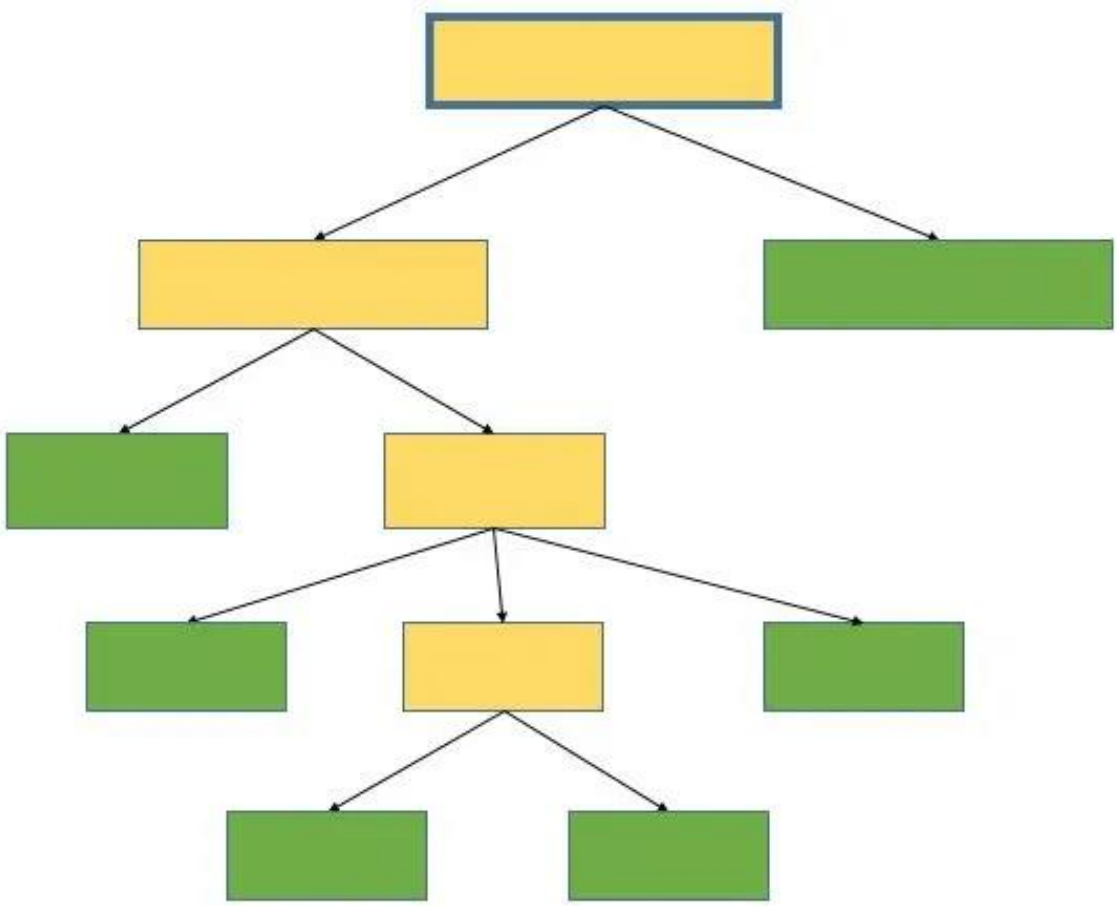

Figure 2.5: Decision Tree for Regression [65]

The performance of the decision tree regression model is typically evaluated using regression metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), or R-squared (coefficient of determination) on a validation or test dataset.

3.1.4. Random Forest Regression

Random Forest Regression is an ensemble machine learning technique used for regression tasks. It is an extension of decision tree regression that combines the predictions of multiple decision trees to provide a more robust and accurate prediction of a continuous numeric output. Random Forests are known for their ability to handle complex relationships, reduce overfitting, and provide insights into feature importance.

Random Forest Regression starts with a dataset containing input features (independent variables) and corresponding continuous target values (dependent variable). The core idea behind Random Forest is to create an ensemble of decision trees.

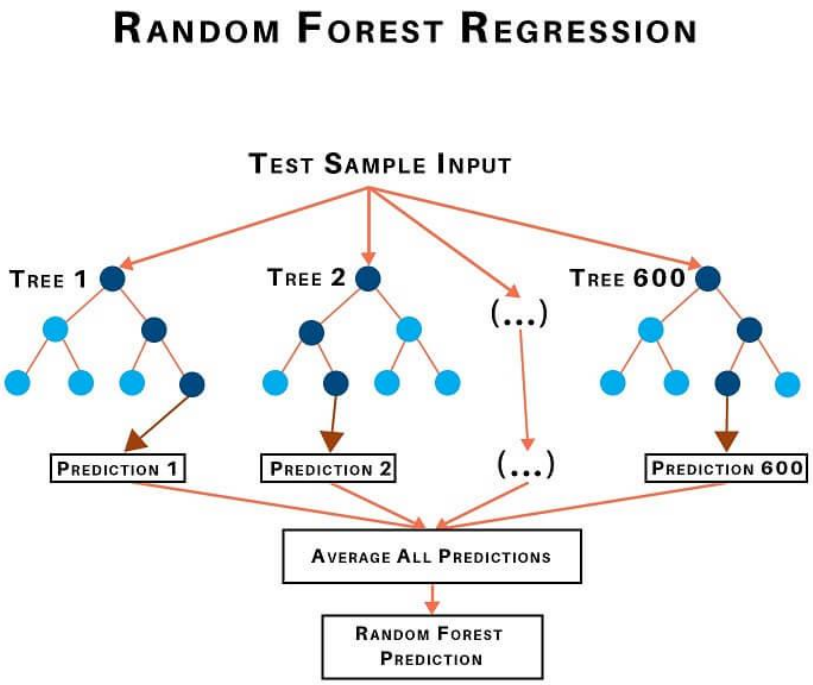

Figure 2.6: Random Forest Regression [66]

Unlike linear regression or polynomial regression, Random Forest Regression does not have a single formula. Instead, it consists of an ensemble of decision trees. Each tree in the ensemble makes individual predictions, and the final prediction is typically an average of these individual predictions.

3.1.5. Ridge Regression

Ridge Regression is a linear regression technique used to address the problem of multicollinearity (high correlation among independent variables) and reduce the impact of overfitting in a linear regression model. It introduces a regularization term to the linear regression equation, which penalizes large coefficient values, leading to a more stable and well-behaved model.

The formula for Ridge Regression is:

Loss function = Sum of squared errors + α * (Sum of squared coefficients)

- **Loss Function**: The loss function represents the objective that Ridge Regression tries to minimize. It is typically the sum of squared errors (similar to ordinary least squares linear regression), which measures the difference between the predicted values and the actual target values.

- **α** (Alpha): Alpha is the hyperparameter in Ridge Regression that controls the strength of the regularization. It is a non-negative parameter that you can tune to strike a balance between fitting the data well and preventing overfitting. A higher α leads to stronger regularization.

- **Sum of Squared Coefficients**: This term, added to the loss function, is the regularization term. It penalizes the model for having large coefficients. The larger the coefficients, the larger this term becomes, encouraging the model to shrink the coefficients towards zero.

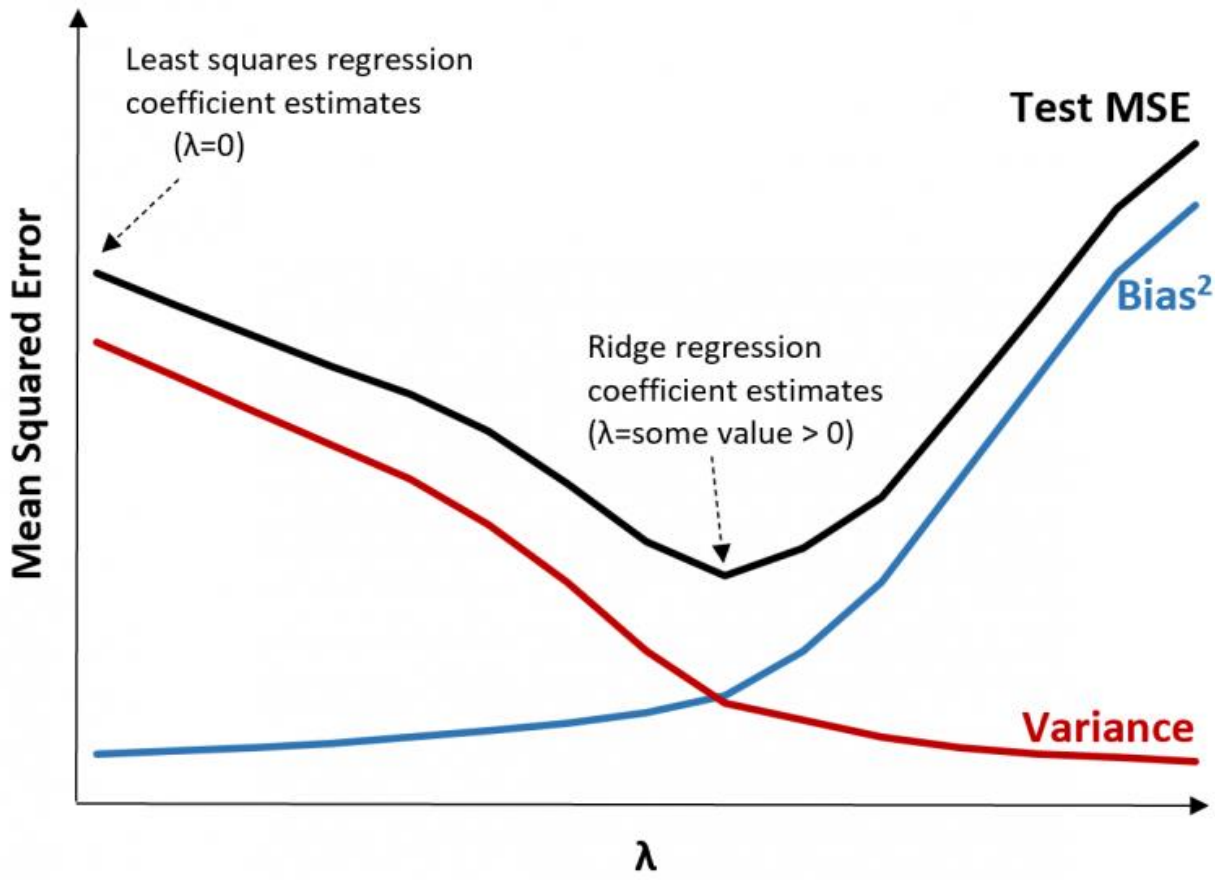

Figure 2.7: Ridge Regression [67]

3.1.6. Lasso Regression

Lasso (Least Absolute Shrinkage and Selection Operator) Regression is a linear regression technique used to address the problem of multicollinearity and perform feature selection by adding a regularization term to the linear regression equation. Similar to Ridge Regression, Lasso Regression helps prevent overfitting, but it has a different regularization term that encourages some coefficient values to be exactly zero, effectively selecting a subset of the most important features.

The formula for Lasso Regression is:

Loss function = Sum of squared errors + α * (Sum of absolute values of coefficients)

- **Loss Function**: The loss function represents the objective that Lasso Regression tries to minimize. Like Ridge Regression, it is typically the sum of squared errors (similar to ordinary least squares linear regression).

- **α** (Alpha): Alpha is the hyperparameter in Lasso Regression that controls the strength of the regularization. It is a non-negative parameter that you can tune to strike a balance between fitting the data well and performing feature selection. A higher α leads to stronger regularization.

- **Sum of Absolute Values of Coefficients**: This term, added to the loss function, is the regularization term specific to Lasso Regression. It penalizes the model for having large coefficients and encourages some coefficients to be exactly zero. In this way, Lasso Regression promotes feature selection.

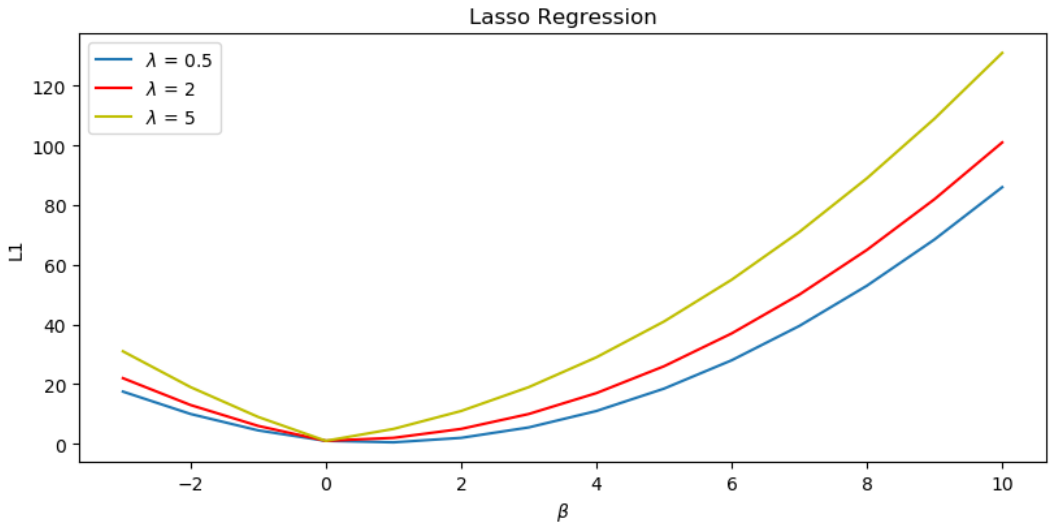

Figure 2.8: Lasso Regression [68]

3.2. The Classification

Classification is a subfield that deals with assigning input examples to predefined classes or categories based on their features. It involves training an algorithm using labeled data, where each input example is associated with a known class or category.

The goal of a classification algorithm is to learn a mapping function that can accurately predict the class labels for new, unseen input examples. The algorithm analyzes the labeled training data to identify patterns or relationships between the input features and the corresponding class labels.

During the training phase, the algorithm builds a model or classifier based on these patterns, aiming to generalize from the training data to make accurate predictions on unseen data. The classifier can then be used to assign class labels to new input examples based on their feature values.

There are various classification algorithms available, including decision trees, logistic regression, support vector machines (SVM), and naive Bayes classifiers. Each algorithm has its own strengths and assumptions, and the choice of algorithm depends on the specific problem and data characteristics. [25]

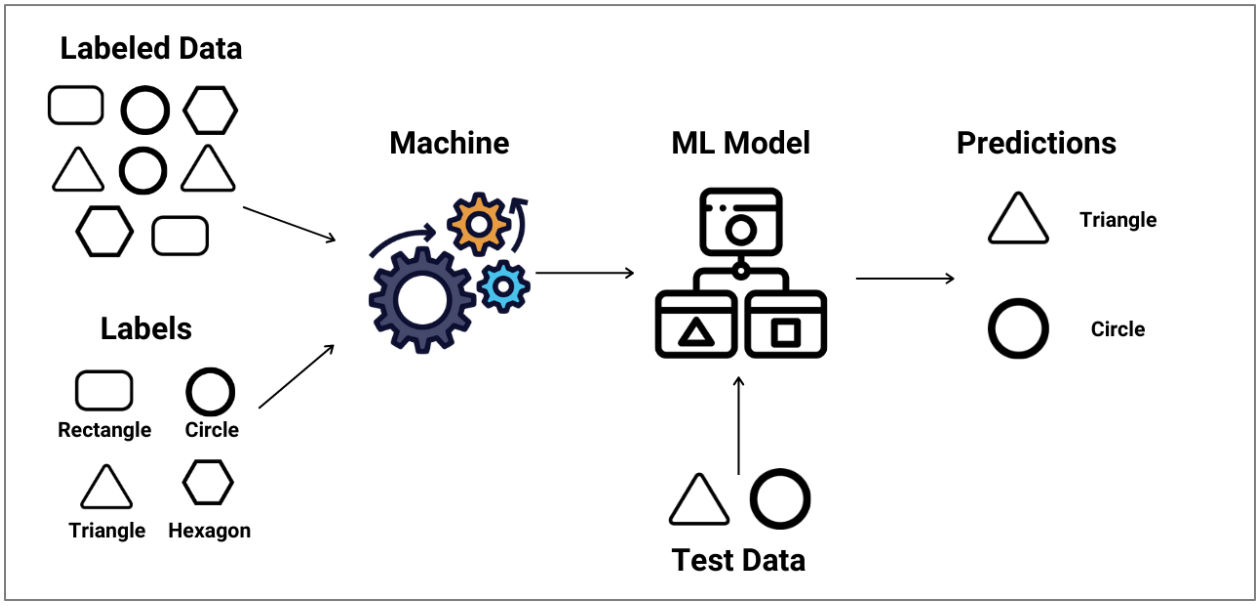

Figure 2.9: Supervised Learning [38]

3.2.1. Logistic Regression

Logistic regression is a widely used classification algorithm in supervised learning. Despite its name, logistic regression is primarily used for binary classification tasks,

where the goal is to predict between two possible classes. It models the relationship between the input features and the probability of belonging to a particular class.

Unlike linear regression, which predicts continuous values, logistic regression predicts the probability of an outcome belonging to a specific class using a logistic function, also known as a sigmoid function. This function maps any real-valued input to a value between 0 and 1, representing the probability of belonging to the positive class.

The logistic regression model works by estimating the optimal coefficients for the input features, which are used to calculate the weighted sum of the features. This weighted sum is then transformed using the logistic function to produce the predicted probability of belonging to the positive class.

The logistic regression formula:

$$\boldsymbol{p(y = 1|x)} = \frac{\mathbf{1}}{\mathbf{1}+\boldsymbol{e}^{-\boldsymbol{z}}} \qquad (2)$$

Here, p(y = 1|x) represents the probability of the positive class given the input features x, and z represents the weighted sum of the features.

Logistic regression has several advantages. It is a simple and interpretable algorithm, providing insights into the relationship between input features and the predicted probability. It can handle both continuous and categorical input features. Logistic regression also performs well when the decision boundary between classes is relatively linear.

However, logistic regression also has limitations. It assumes a linear relationship between the input features and the log-odds of the target variable, which may not hold in complex datasets. It can struggle with capturing complex nonlinear relationships without incorporating additional techniques like polynomial features or interaction terms. Logistic regression is also sensitive to outliers and can be affected by multicollinearity when features are highly correlated.

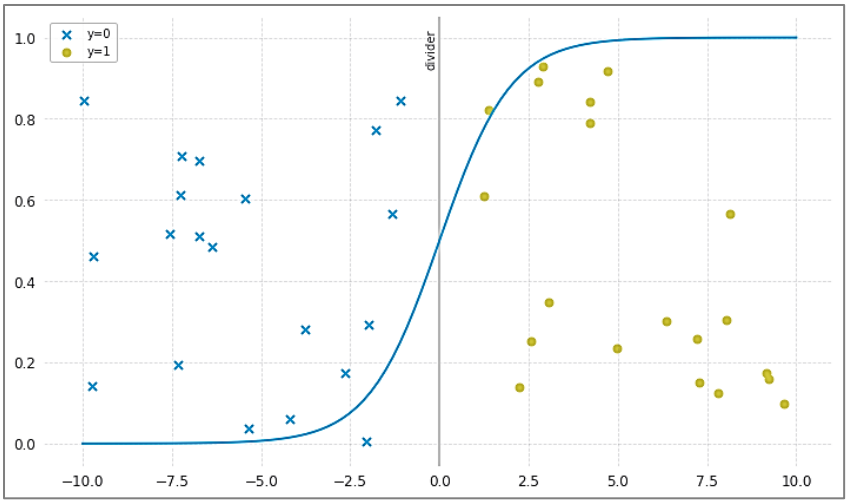

Figure 2.10: Binary Logistic Regression [39]

3.2.2. Decision Trees

Decision trees are a classification algorithm in supervised learning that create a tree-like model of decisions based on the input features. Each internal node of the tree represents a feature or attribute, and each leaf node represents a class or category. Decision trees are intuitive, easy to understand, and can handle both binary and multi-class classification tasks.

The decision tree algorithm works by recursively splitting the dataset based on different features, aiming to maximize the homogeneity or purity of the classes within each subset. The splitting process involves selecting the feature and the splitting criterion that best separates the data points of different classes.

At each internal node, the decision tree algorithm evaluates a feature and applies a split based on a specific condition.

The decision tree format can be represented visually as a tree structure, with nodes representing the features and branches representing the possible values or conditions. Each leaf node corresponds to a class label or a probability distribution over classes.

Decision trees have several advantages. They are easy to interpret and visualize, providing clear insights into the decision-making process. Decision trees can handle both categorical and numerical features, and they are robust to missing values. They can

capture complex relationships between features, including nonlinear interactions. Decision trees also handle irrelevant features well by automatically assigning them less importance.

However, decision trees can suffer from over-fitting, where the model becomes too complex and tailored to the training data, resulting in poor generalization to unseen data. To mitigate over-fitting, techniques like pruning and setting a minimum number of samples per leaf node can be applied. Decision trees are also sensitive to small changes in the data, which can lead to different tree structures.

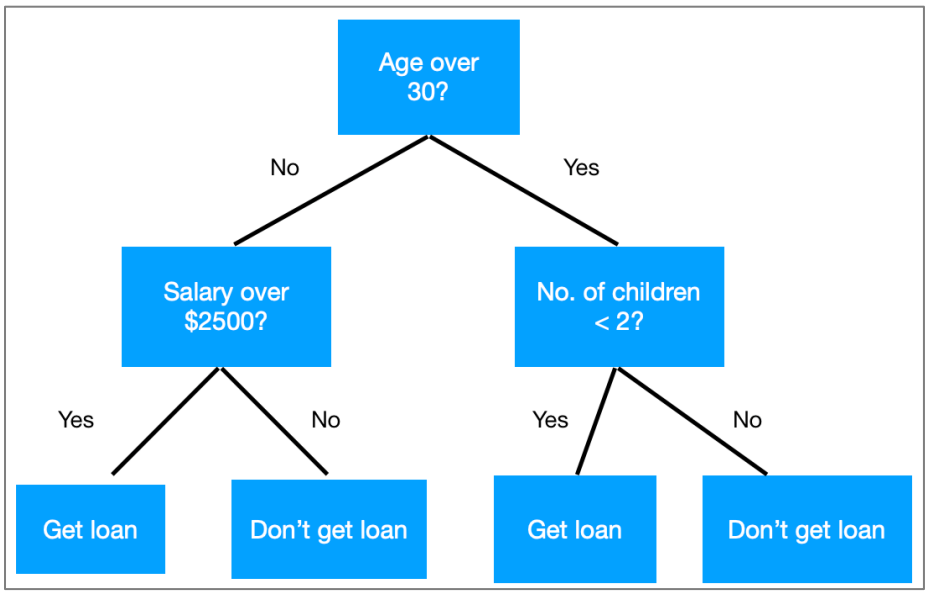

Figure 2.11: Decision Tree [40]

3.2.3. Naive Bayes

Naive Bayes is a classification algorithm based on Bayes' theorem and the assumption of feature independence, known as the "naive" assumption. Despite its simplifying assumption, Naive Bayes has been proven to be effective in many real-world applications, especially in text classification and spam filtering.

The algorithm works by calculating the probability of a data point belonging to a particular class given its features. It calculates the conditional probability of the class given the features using Bayes' theorem:

$$P(Class|Features) = \frac{P(Features|Class) \times P(Class)}{P(Features)} \quad (3)$$

In Naive Bayes, the "naive" assumption assumes that the features are conditionally independent given the class. This means that the presence or absence of one feature does not affect the presence or absence of other features. Although this assumption rarely holds true in real-world data, Naive Bayes can still produce reasonably accurate results in practice.

To apply Naive Bayes, the algorithm first estimates the prior probability of each class, P (Class), based on the distribution of classes in the training data. It then estimates the conditional probability of the features given each class, P (Features Class), by assuming independence among the features. This estimation can be done using different probability models, such as Gaussian Naive Bayes for continuous features or multinomial Naive Bayes for discrete features.

Once the probabilities are estimated, Naive Bayes uses them to calculate the posterior probability of each class given the features and assigns the data point to the class with the highest probability.

Naive Bayes has several advantages. It is computationally efficient and requires a relatively small amount of training data compared to other algorithms. It can handle high-dimensional feature spaces effectively, making it suitable for text classification tasks. Naive Bayes is also robust to irrelevant features and performs well in situations where the independence assumption is approximately met.

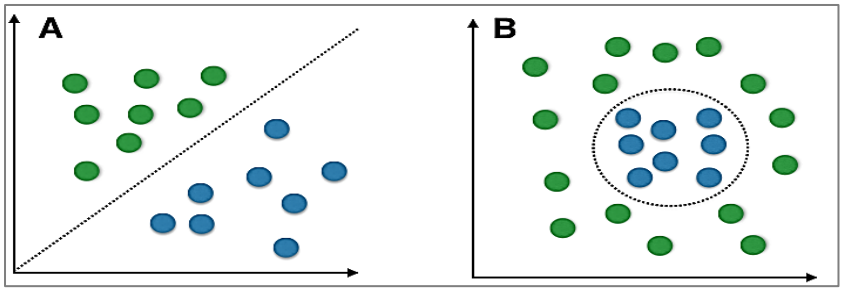

Figure 2.12: Linear (A) vs. non-linear problems (B) [41]

3.2.4. Support Vector Machines

Support Vector Machines (SVMs) are classification algorithms that aim to find an optimal hyperplane that separates data points of different classes in a high-dimensional feature space. SVMs are widely used in various domains, including image classification, text categorization, and bioinformatics.

The key idea behind SVMs is to maximize the margin, which is the distance between the decision boundary (hyperplane) and the nearest data points of each class. The hyperplane that achieves the largest margin is considered the optimal solution. This margin maximization leads to better generalization and robustness of the classifier.

1. **Data Preparation**: The input data is represented as feature vectors in a high-dimensional space. SVMs work well with both linear and nonlinear data, thanks to a technique called the kernel trick, which allows mapping the data into a higher-dimensional space where it may be linearly separable.

2. **Hyperplane Selection**: The SVM algorithm searches for the hyperplane that separates the data points with the maximum margin. The hyperplane is defined by a weight vector and a bias term.

3**. Support Vector Identification**: The algorithm identifies the support vectors, which are the data points closest to the decision boundary. These support vectors play a crucial role in determining the hyperplane and making predictions.

4. **Margin Optimization**: The goal is to find the hyperplane that maximizes the margin while minimizing the classification error. This is achieved through optimization techniques, such as quadratic programming or convex optimization.

SVMs offer several advantages. They can handle high-dimensional feature spaces effectively and are less prone to overfitting compared to other algorithms. SVMs can handle both linearly separable and non-linearly separable data by using different types of kernels, such as linear, polynomial, or Gaussian (RBF) kernels. Additionally, SVMs can capture complex decision boundaries and handle datasets with outliers.

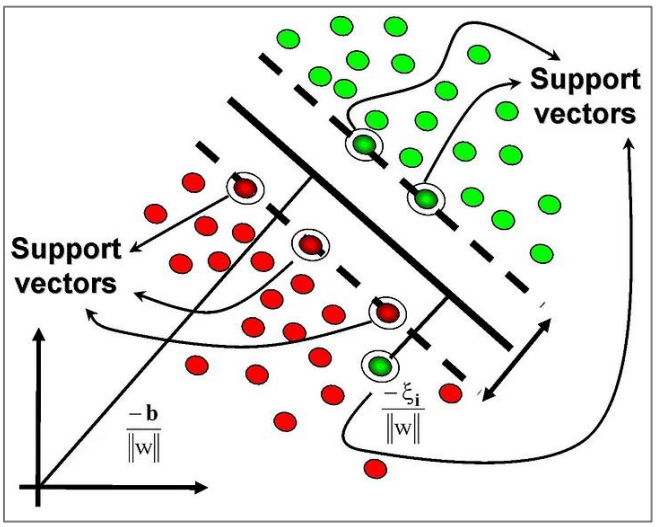

Figure 2.13: Illustration of Support Vector Machine [42]

3.2.5. K-Nearest Neighbors

K-Nearest Neighbors (KNN) is a simple yet effective classification algorithm in machine learning. It is a non-parametric algorithm, meaning it doesn't make any assumptions about the underlying data distribution. KNN is often used for pattern recognition and can handle both classification and regression tasks.

1. **Data Preparation**: The input data is represented as feature vectors in a multidimensional space, where each data point is associated with a class label.

2. **Distance Calculation**: KNN calculates the distance between the new data point (to be classified) and all other data points in the training set. The most common distance metrics used are Euclidean distance and Manhattan distance, but other metrics can also be used depending on the data.

3**. Finding Neighbors**: KNN selects the K nearest data points (neighbors) to the new data point based on the calculated distances. K is a user-defined parameter.

4**. Class Assignment**: For classification, the class label of the new data point is assigned based on the majority vote of the K nearest neighbors. The class that appears most frequently among the K neighbors determines the class label of the new data point. In regression tasks, the output value is typically the average or weighted average of the target values of the K nearest neighbors.

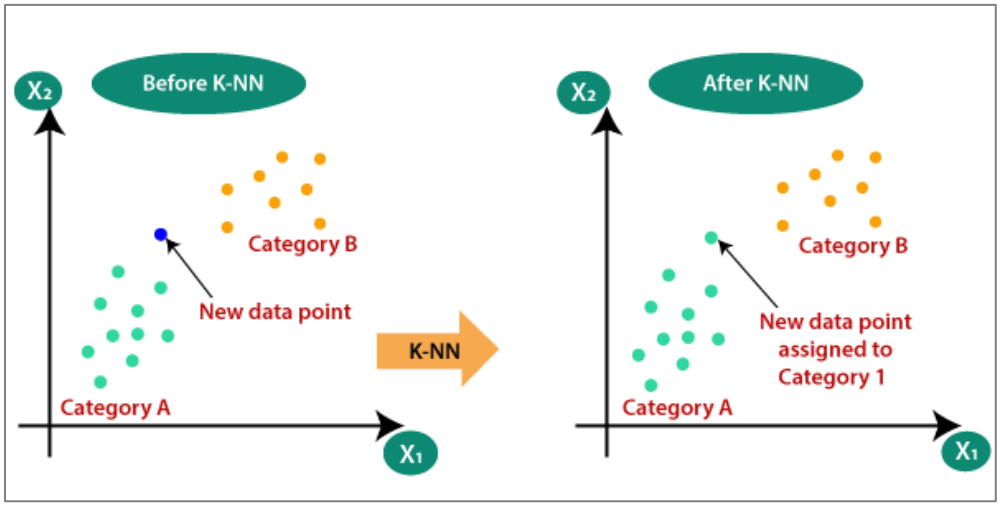

Figure 2.14: K-Nearest Neighbor [43]

KNN has several characteristics that differentiate it from other classification algorithms:

1. **Non-Parametric**: KNN does not assume any underlying data distribution, making it versatile and applicable to a wide range of problems.

2. **Instance-Based Learning**: KNN does not explicitly build a model from the training data but instead stores all the training instances. This makes it easy to adapt to new training data without retraining the model.

3. **Lazy Learning**: KNN postpones the computation until the prediction phase, as it does not perform any explicit training process. This can be advantageous when dealing with dynamic or evolving data.

The strengths of KNN include its simplicity, as it is easy to understand and implement. It can handle multi-class classification problems and can adapt well to complex decision boundaries. KNN is also robust to outliers and can be effective when dealing with imbalanced datasets.

4. Unsupervised learning

Unsupervised learning is a subfield of machine learning where an algorithm learns patterns, relationships, or structures from unlabeled data without any predefined output labels or target variables. Unlike supervised learning, there is no explicit guidance or ground truth provided to the algorithm.

In unsupervised learning, the algorithm explores the data to identify inherent structures or patterns based on the similarities, differences, or distributions within the data itself. It aims to discover meaningful insights or groupings that can help in understanding the underlying nature of the data.

Common tasks in unsupervised learning include clustering, dimensionality reduction, and anomaly detection. Clustering algorithms group similar data points together based on their proximity or similarity. Dimensionality reduction techniques aim to reduce the number of input variables while preserving relevant information. Anomaly detection focuses on identifying data points that deviate significantly from the expected patterns.

Unsupervised learning is particularly useful in exploratory data analysis, pattern recognition, and generating insights from large and complex datasets. [26]

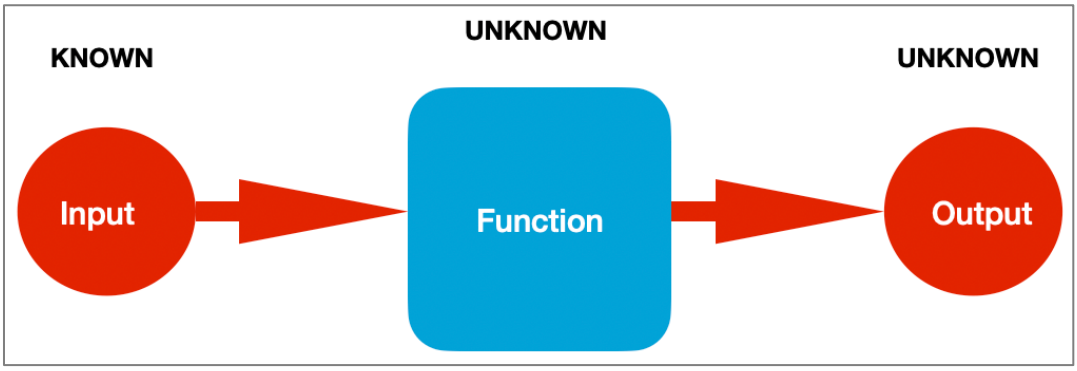

Figure 2.15: The input and output of Unsupervised Learning [44]

4.1. Clustering

Clustering algorithms aim to group similar data points together based on their inherent similarity or proximity in the feature space. The goal is to identify natural groupings or clusters within the data. Examples of clustering algorithms include k-means clustering, hierarchical clustering, DBSCAN (Density-Based Spatial Clustering of Applications with Noise), and Gaussian Mixture Models (GMM).

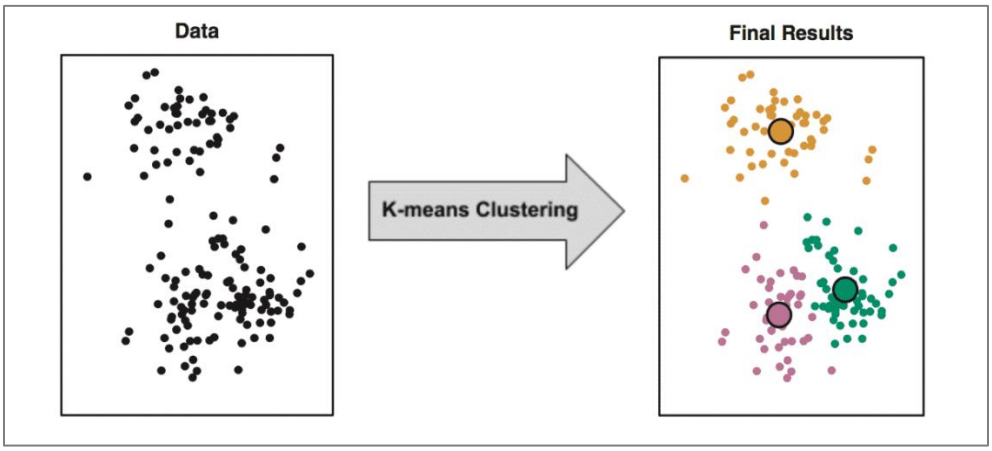

Figure 2.16: K-means clustering [45]

4.2. Dimensionality reduction

Dimensionality reduction techniques are used to reduce the number of input features while retaining the essential information present in the data. They are particularly useful when dealing with high-dimensional data or when visualizing data in lower-dimensional spaces. Principal Component Analysis (PCA), t-SNE (t-Distributed Stochastic Neighbor Embedding), and Linear Discriminant Analysis (LDA) are common dimensionality reduction methods.

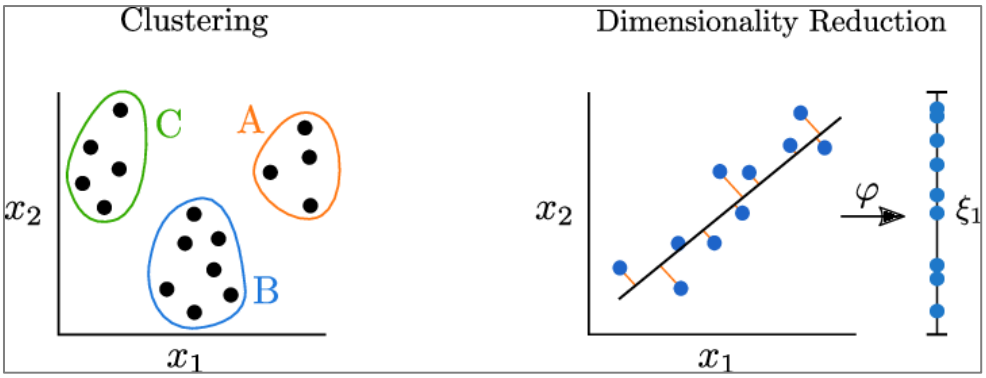

Figure 2.17: Clustering and Dimensionality Reduction [46]

4.3. Anomaly detection

Anomaly detection focuses on identifying rare or unusual data points that deviate significantly from the norm. It is particularly useful for detecting fraud, network intrusions, or any unexpected patterns in data. Techniques for anomaly detection include statistical approaches such as outlier analysis, clustering-based methods, and auto-encoders.

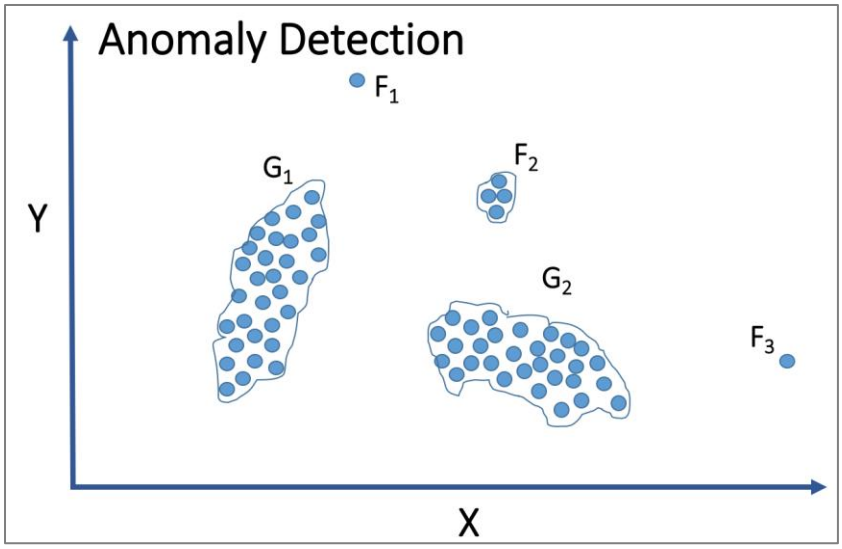

Figure 2.18: Anomaly detection [47]

4.4. Association rule mining

Association rule mining aims to discover interesting associations or relationships between variables in a dataset. It identifies frequent item-sets or co-occurring patterns and generates rules that express relationships between different items. Apriority algorithm and FP-Growth algorithm are popular association rule mining techniques.

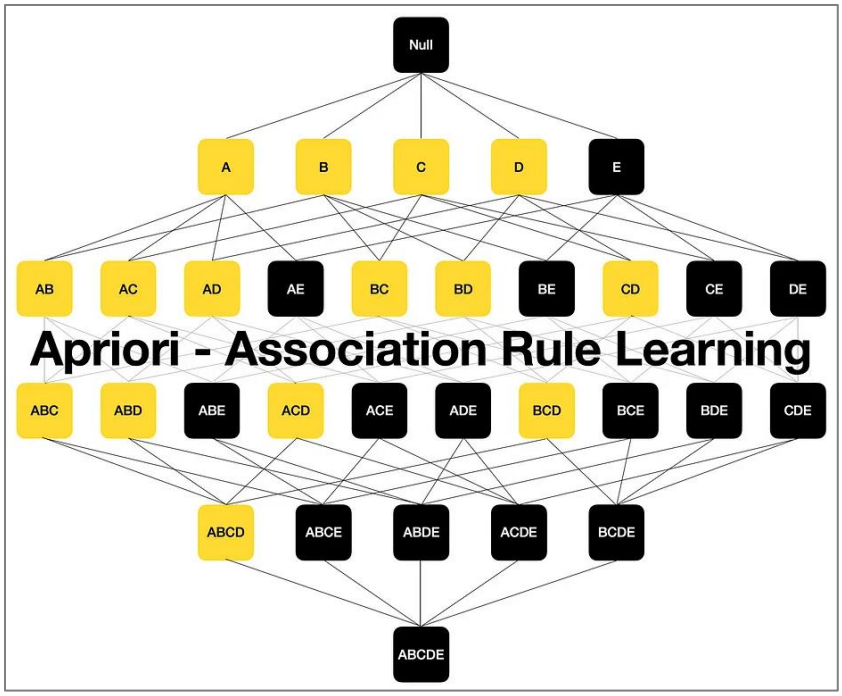

Figure 2.19: Association Rule Learning [48]

4.5. Generative modeling

Generative models learn the underlying data distribution and generate new samples that resemble the original data. These models can be used for tasks such as data synthesis, image generation, and text generation. Popular generative models include Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and Restricted Boltzmann Machines (RBMs).

4.6. Self-organizing maps

Self-organizing maps (SOM) are neural network-based unsupervised learning techniques that map high-dimensional input data onto a lower-dimensional grid. SOMs can be used for visualization, clustering, and discovering topological relationships in the data.

Unsupervised learning does not have a specific formula or equation like supervised learning algorithms. Instead, it employs various statistical and mathematical techniques to uncover patterns and structures within the data.

One challenge of unsupervised learning is the lack of a clear evaluation metric, as there are no explicit target variables to measure against. The assessment of unsupervised learning models often relies on subjective judgment or indirect measures of performance.

Additionally, since unsupervised learning algorithms do not have access to labeled data, they are more susceptible to finding spurious patterns or being influenced by irrelevant or noisy data. Without clear guidance, unsupervised learning may discover patterns that may not necessarily align with meaningful insights or actionable knowledge.

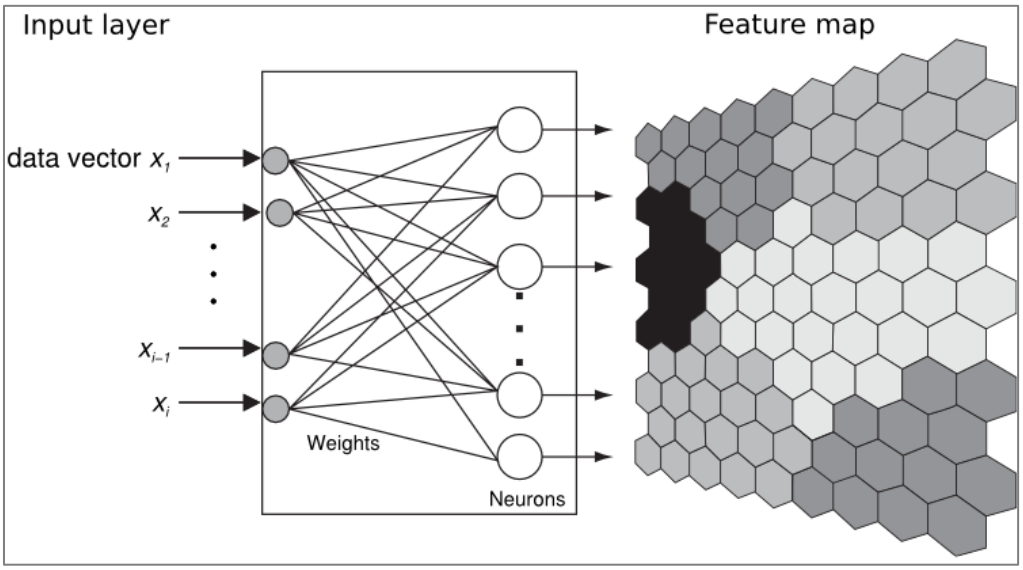

Figure 2.20: Example of Unsupervised Learning [49]

5. Semi-supervised learning

Semi-supervised learning is a machine learning approach that combines elements of both supervised and unsupervised learning. It leverages a small amount of labeled data along with a larger amount of unlabeled data to train a model.

The concept of semi-supervised learning has its roots in the field of active learning, which aims to select the most informative samples for labeling. However, the term "semi-supervised learning" gained popularity in the early 2000s as a distinct area of study.

Semi-supervised learning works by using the labeled data to learn from the available labels, and then utilizing the unlabeled data to capture the underlying patterns or structure of the data. The model learns from the combination of labeled and unlabeled data, and the goal is to improve the predictive performance of the model compared to using only labeled data.

The formula or equation for semi-supervised learning depends on the specific algorithm or technique being used. Various algorithms can be adapted for semi-supervised learning, including those from both supervised and unsupervised learning domains.

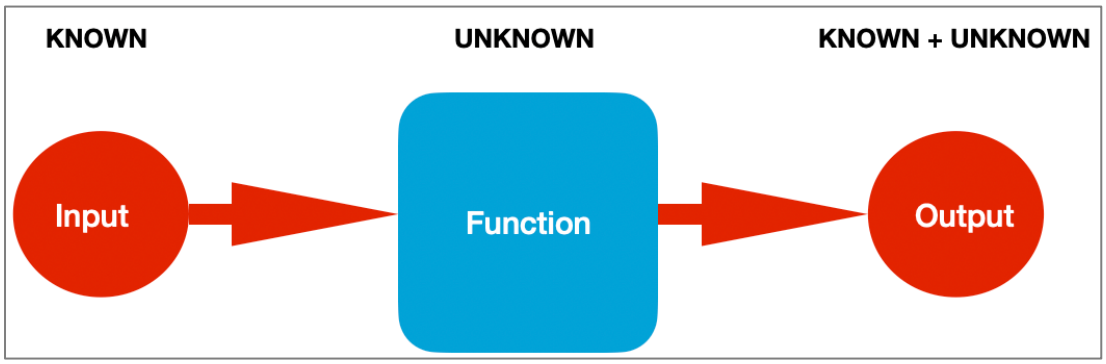

Figure 2.21: The input and output of Semi-Supervised [50]

5.1. Semi-Supervised Techniques

1. **Self-training**: This approach involves training a model using the labeled data and then using the trained model to predict labels for the unlabeled data. The confident predictions are then treated as additional labeled data and used to retrain the model iteratively.

2. **Co-training**: Co-training involves training multiple models on different subsets of features or views of the data. Each model learns from a different view and contributes its predictions to label the unlabeled data. This technique assumes that different views provide complementary information.

3**. Graph-based methods**: Graph-based semi-supervised learning methods construct a graph representation of the data, where nodes represent data points and edges represent

relationships between them. Label propagation algorithms leverage the graph structure to propagate labels from labeled to unlabeled data points.

4**. Generative models**: Generative models, such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), can be extended to semi-supervised learning. The generative models learn the underlying data distribution and utilize the labeled data to guide the generation process.

One challenge of semi-supervised learning is the reliance on the quality and representativeness of the labeled data. If the labeled data is biased or contains errors, it can negatively impact the performance of the semi-supervised learning model. Additionally, the choice of the right combination of labeled and unlabeled data and the handling of the unlabeled data in an effective manner can be challenging.

6. Reinforcement learning

Reinforcement learning is a subfield of machine learning that focuses on how agents can learn to make optimal decisions or take actions in an environment to maximize a reward or minimize a penalty. It is inspired by the concept of how humans and animals learn through trial and error and interaction with their surroundings. [27]

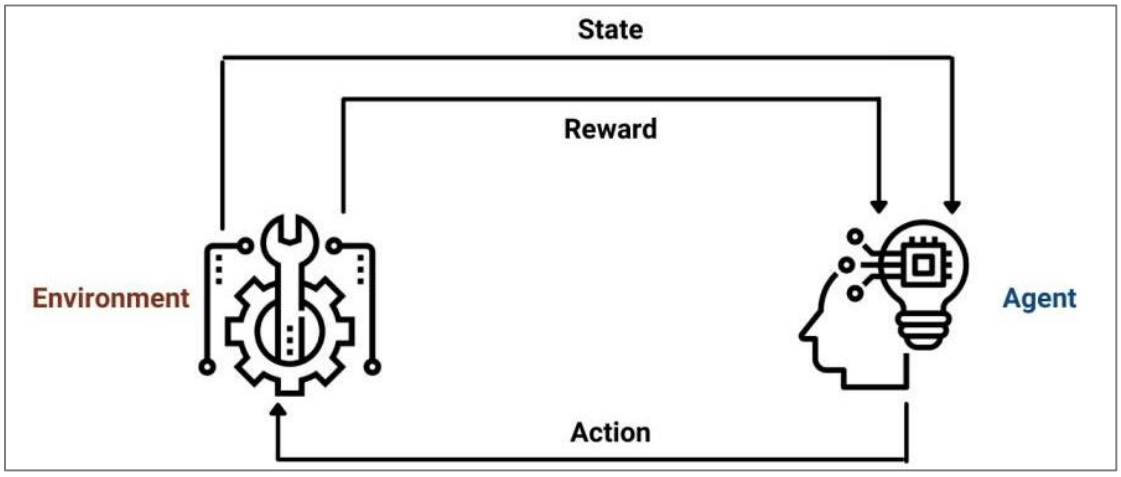

Figure 2.22: Reinforcement Learning [51]

In reinforcement learning, an agent interacts with an environment and learns by receiving feedback in the form of rewards or punishments based on its actions. The agent's goal is to discover a strategy or policy that maximizes the cumulative reward over time. Through a process of exploration and exploitation, the agent learns to take actions that lead to desirable outcomes.

Reinforcement learning typically involves a Markov Decision Process (MDP) framework, where the environment is modeled as a series of states, actions, rewards, and transition probabilities. The agent learns a policy, which is a mapping of states to actions, through various algorithms such as Q-learning or policy gradients.

6.1. Model-free methods

Model-free techniques learn directly from experience without explicitly modeling the dynamics of the environment. Q learning and SARSA (State-Action-Reward-State-Action) are popular model-free algorithms. They learn the values of state-action pairs to guide decision-making.

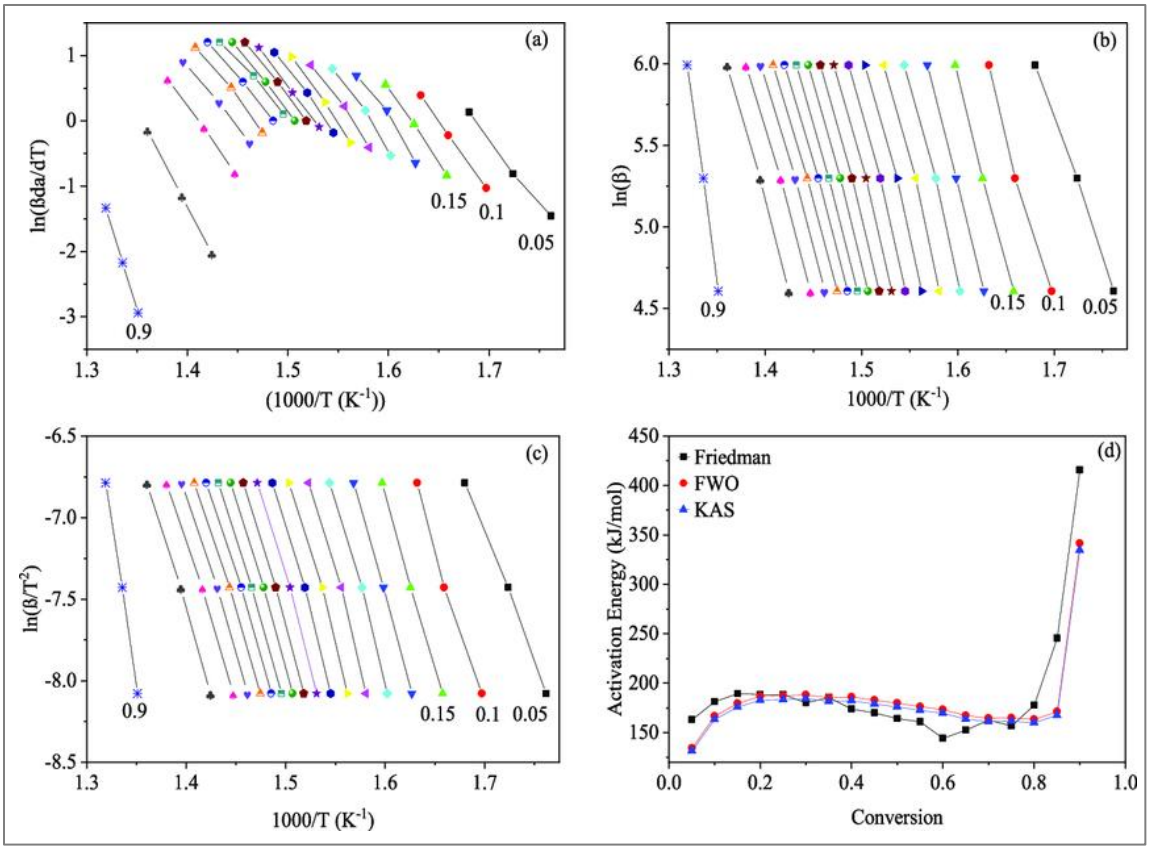

Figure 2.23: Plots of the model-free methods for high heating rates [52]

6.2. Model-based methods

Model-based techniques involve building a model of the environment's dynamics and using it to plan and make decisions. The model represents the transition probabilities and reward functions. Model-based algorithms combine planning and learning to improve decision-making efficiency.

Reinforcement learning has been successfully applied in various domains, including robotics, game playing, autonomous systems, and recommendation systems. It has achieved notable successes such as AlphaGo, which defeated human champions in the game of Go, and OpenAI's Dota 2 bot, which achieved high-level gameplay.

However, one challenge of reinforcement learning is the "exploration-exploitation" trade-off. The agent must explore different actions and states to discover the best policy

while simultaneously exploiting the current knowledge to maximize immediate rewards. Balancing exploration and exploitation can be difficult, especially in complex environments with sparse rewards.

7. Deep learning

Deep learning is a subfield of machine learning that focuses on training artificial neural networks with multiple layers (deep neural networks) to learn and represent complex patterns and relationships in data. It is inspired by the structure and functioning of the human brain's neural networks.

The origins of deep learning can be traced back to the development of artificial neural networks in the 1940s and 1950s. The seminal work of Warren McCulloch and Walter Pitts laid the foundation for understanding the computational capabilities of neural networks. In the 1980s, researchers such as Geoff Hinton, Yann LeCun, and Yoshua Bengio made significant contributions to the field by developing algorithms and techniques for training deep neural networks.

Deep learning works by training neural networks with multiple layers of interconnected nodes, known as neurons. Each neuron applies a nonlinear transformation to its inputs and passes the result to the next layer. The networks learn through a process called backpropagation, where errors are propagated backward through the layers to adjust the network's parameters (weights and biases).

The formula or equation for deep learning depends on the specific architecture and type of neural network being used. However, the fundamental operation in deep learning is the weighted sum of inputs passed through an activation function.

7.1. Convolutional Neural Networks (CNN)

CNNs are widely used for image and video processing tasks. They consist of convolutional layers that extract local features from the input data, followed by pooling layers for spatial down sampling. CNNs have revolutionized image recognition, object detection, and image generation.

7.2. Recurrent Neural Networks (RNN)

RNNs are designed for sequence data processing, where the output of each step is fed back as input for the next step. They are effective in tasks involving time series data, natural language processing, speech recognition, and machine translation.

7.3. Long Short-Term Memory (LSTM) Networks

LSTMs are a type of RNN that can capture long-term dependencies in sequential data. They address the vanishing gradient problem and have been successful in tasks requiring memory, such as language modeling, speech recognition, and sentiment analysis.

7.4. Generative Adversarial Networks (GAN)

GANs consist of two neural networks—a generator and a discriminator—that compete against each other in a game-theoretic framework. GANs have revolutionized generative modeling and are used for tasks like image generation, style transfer, and data augmentation.

Conclusion

In this chapter, we provided an overview of machine learning and its various algorithms. We began by defining machine learning and highlighting its different proposed architectures. Subsequently, we delved into an extensive explanation of supervised learning algorithms.

CHAPTER 3

THE ART STATE & THE PROPOSED APPROACH

INTRODUCTION

Forests fire spread prediction using machine learning has gained significant attention from researchers in recent years. Previous studies have investigated the potential of machine learning algorithms to accurately predict forests fire occurrence, spread, and severity. Effort to build early warning systems based on weather data that reduce forest risks. In this chapter, we will present the previous works relating to the studied subject and present our approach.

1. Methodologies and Approaches

Previous research on forest fire spread prediction using machine learning has employed diverse methodologies and approaches. One common approach is the use of supervised learning techniques, where historical fire data and associated environmental variables are used to train prediction models. These models aim to learn the relationships between the input features and the occurrence, spread, and severity of forest fires. The models are trained using labeled data, where fire events are classified based on their characteristics and environmental conditions.

Another approach explored in previous studies is the use of unsupervised learning techniques. Unsupervised learning aims to identify patterns and clusters in the data without the need for labeled examples. In the context of forest fire prediction, unsupervised learning can be used to identify spatial and temporal patterns in fire occurrences, detect fire hotspots, and group similar fire events based on their characteristics. This approach can provide valuable insights into the underlying patterns and dynamics of forest fires.

In addition to traditional machine learning techniques, some studies have explored the use of advanced methods such as deep learning. Deep learning algorithms, particularly convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have shown promising results in various domains. In the context of forest fire prediction, deep learning

models can capture complex spatial and temporal dependencies in the data, leading to predictions that are more accurate. These models have the ability to automatically learn hierarchical representations from the input features, enabling them to capture intricate patterns and relationships.

2. Performance Evaluation and Findings

Performance evaluation of forest fire spread prediction models has been conducted using various metrics and previous researches has demonstrated the potential of machine learning techniques in improving forest fire spread prediction accuracy and providing valuable insights for fire management

2.1.Canada's Research:

In 2019, in Canada, FireCast, a novel system, was introduced. It is a supervised machine learning algorithm that combines artificial intelligence (AI) techniques with data collection strategies from geographic information systems (GIS).FireCast predicts which areas surrounding a burning wildfire have high-risk of near-future wildfire spread, based on historical fire data and using modest computational resources.

Layer	Operation	Kernel/Pool Size	Feature Maps
1	Avg Pooling	2×2	–
2	Convolution	3×3	32
	Max Pooling	2×2	–
	Dropout	–	–
3	Convolution	3×3	64
	Max Pooling	2×2	–
	Dropout	–	–
4-Out	Dense	–	128
5	Concat	–	136
6-Out	Dense	–	1

Figure 3.1: FireCast model architecture [53]

FireCast provides an image-based output for each day of a testing fire. These images depict sampled Points of Interest (POIs); each assigned a color corresponding to its

predicted likelihood of burning. To generate these images, the model utilizes various visual input layers, a known starting fire perimeter, and 24 hours of atmospheric data specific to the location.

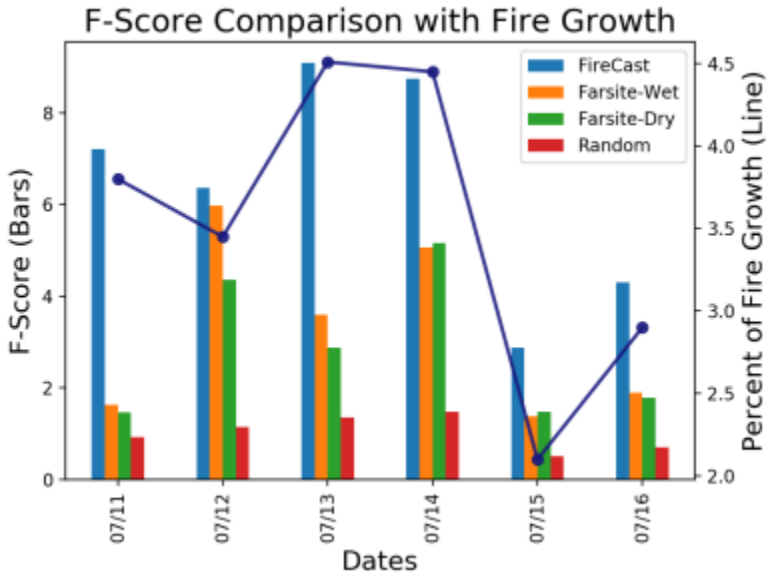

Figure 3.2: Comparing F-scores of FireCast, Farsite, and a random model. The line represents the percent of fire growth. [54]

FireCast performance achieves an average accuracy of 87.7%. In comparison, random predictions yield an accuracy of 50.4%, while Farsite, with wet and dry fuel moisture conditions, achieves accuracies of 67.8% and 63.6%, respectively.

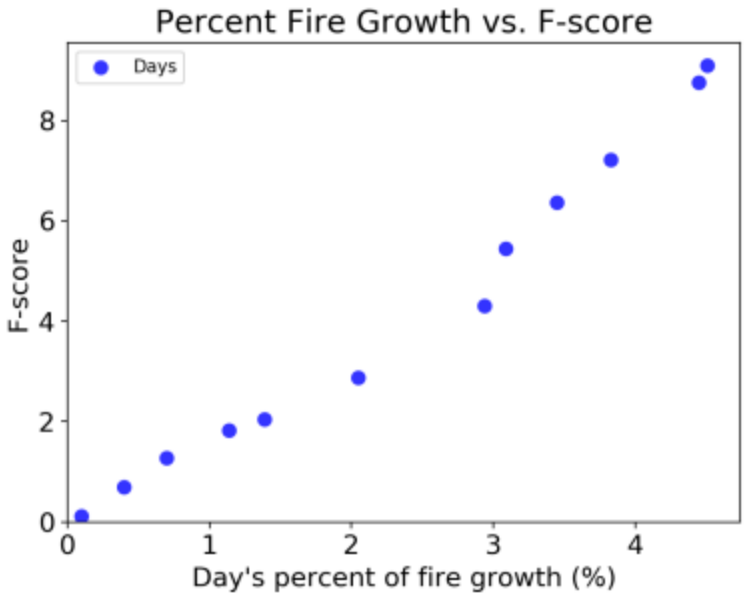

Figure 3.3: Comparisons between the F-score and percent of new burn for two chunks of consecutively mapped days of the testing fire. [55]

For the testing fire, they observe an average recall of 91.1% for FireCast, 50.4% for the randomly assigned pixels, 74.8% and 81.1% for wet and dry condition Farsite models respectively. Thus, FireCast outperforms current fire modeling technologies with respect to recall.

Instead of performing predetermined calculations from manually collected input layers, machine learning within FireCast allows the system to learn important correlations. The reduced number of necessary and variable data.

FireCast	Farsite
Landsat8	Fuel Model
DEM	DEM
Fire Perimeters	Canopy Cover
Weather/Wind	Weather/Wind
	Adjustment File
	Fuel Moisture
	Conversion File*
	Fuel Model File*
	Fire Acceleration File*
	Fire Perimeters*

Figure 3.4: FireCast vs. Farsite input variables. Optional input files. [56]

2.2.United States's research

In 2022 at United States. By collecting the data over the contiguous United States from 2012 to 2020.

PUBLICLY AVAILABLE FIRE DATASETS

Name	Fire information	Coverage	Period	Spatial resolution	Fire spreading	Other variables
FRY [20]	Total burn area	Worldwide	2005 - 2011	500 m	N/A*	-
Fire Atlas [21]	Total burn area	Worldwide	2003 - 2016	500 m	N/A*	-
GlobFire [22]	Active fire	Worldwide	2001 - 2017	500 m	Daily fire maps	-
1.88 million US Wildfires [23]	Active fire	USA	1992 - 2015	Point coordinates	N/A	-
Fire events in Canada [30]	Active fire	Regions in Canada	August 2014	Point coordinates	N/A	Vegetation, Surface Temperature
Next Day Wildfire Spread	Active fire	USA	2012 - 2020	1 km	t and $t+1$ *day* fire maps	Elevation, Wind Direction and Speed, Min. and Max. Temperatures, Humidity, Precipitation, Drought Index, Vegetation, Population density, ERC

Figure 3.5: Publicly Available Fire Datasets [57]

Then split the data between training, evaluating, and testing by randomly separating between 2012 and 2020 according to an 8:1:1 ratio, respectively. Then using Neural Network. Logistic Regression and a Random Forest models.

.

	AUC (PR)	Precision	Recall
Neural Network	28.4	33.6	43.1
Random Forest	22.5	26.3	46.9
Logistic Regression	19.8	32.5	35.3
Persistence	11.5	35.7	27.3

Figure 3.6: Wildfire spreading prediction metrics [58]

The Highest AUC is the neural network at 28.4%, followed by the random forest and then logistic regression. The precision and recall for the neural network on the positive class are 33.6% and 43.1%, respectively. The AUCs of the logistic regression and random forest models— the non-deep learning models—are within 3% of one another and at least 6% lower than the neural network. The logistic regression baseline achieves nearly the same precision as the neural network at 32.5% but has a lower recall at 35.3%. The random forest baseline has a higher recall than the neural network at 46.9% but has a lower precision at 26.3%.

The best precision/recall tradeoff for different resolutions is when the output corresponds to an area that is eight times larger than the input area. This results in an AUC of 66.3%. While the prediction metrics improve as the prediction region size increases, the predictions become less useful from an operational perspective due to less localization of the fire. When predicting at lower spatial resolutions, smaller fires are

Label	Lower Resolution	AUC (PR)	Precision	Recall
Decimal	2x	38.8	48.4	40.8
	4x	52.3	69.1	39.8
	8x	64.5	88.7	24.7
Binary	2x	39.8	40.3	51.9
	4x	53.5	48.8	61.0
	8x	66.3	62.1	63.1
Persistence	2x	19.5	45.6	36.8
	4x	30.6	56.0	46.8
	8x	41.3	64.0	52.5

Figure 3.7: Lower resolution predictions [59]

2.3.Heilongjiang's research

In 2022 at Heilongjiang, china. By collecting the data over the Heilongjiang Province from 2006 to 2019. To build the ANN model.

Variable type	Variable name	Source	Code
Climatic factors	Relative humidity per hour	China Meteorological Data Network data.cma.cn/	Humidity
	Hourly temperature		Temperature
	Precipitation per hour		Precipitation
	Wind speed per hour		Wind speed
	Hourly wind direction		Wind direction
Topographic variables	Slope	Geospatial Data Cloud www.gscloud.cn/	
	Aspect		
	Elevation		
Combustible factor variables	Vegetation cover type	Institute of Botany, Chinese Academy of Sciences www.ibcas.ac.cn/	Vegetation
	Surface water content		TVDI
Land cover type variables	Roads	National Catalogue Service for Geographic Information www. webmap.cn/	Lrdl
	Railways		Lrrl
	Settlements		redl
	Lakes		Hyda
	Ditches		hydl
	Wells		hydp

Figure 3.8: Variables needed to build the model. [60]

The Boruta feature-screening method was used to sort and filter the fifteen extracted variables. Which based on the random forest classification algorithm. Then the dataset was divided into two subsets: one subset comprised 70% of the total data and was used for model training, and the other subset included 30% of the data and was used for data verification and testing.

The final set of variables is divided into two types: confirmed and rejected variables. And K-nearest neighbors (KNN) classification algorithm was used to sort and filter the 16 different variables.

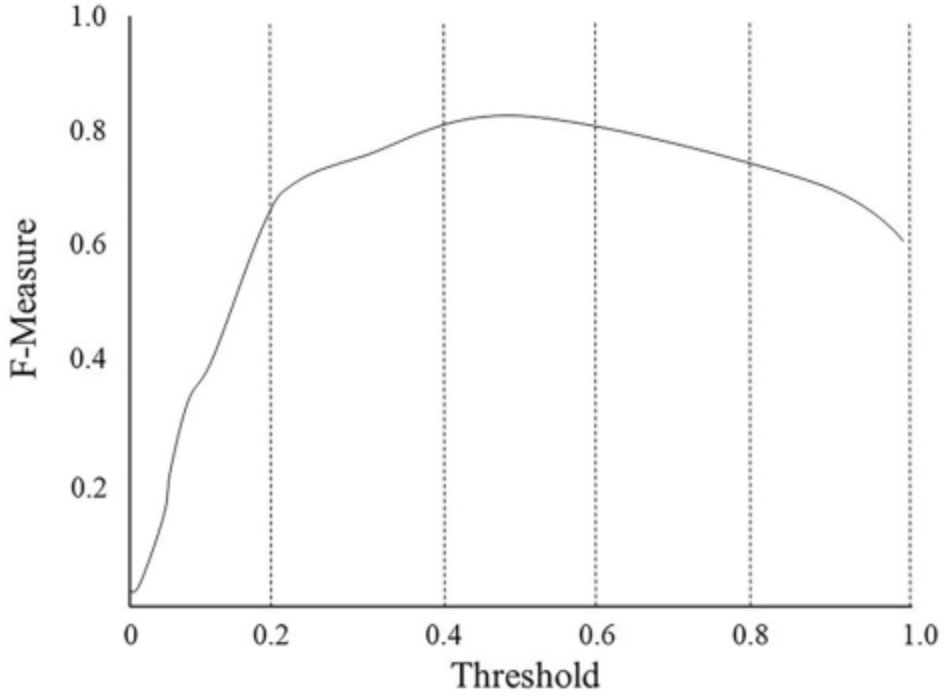

Figure 3.9: The ANN predictions of the average F-measures under different thresholds for 2414 combustion maps. [61]

The Wang Zhengfei's model. Which is considers on the paper the base model for the combination of it with CA. Was included to compare the accuracy of the two models.

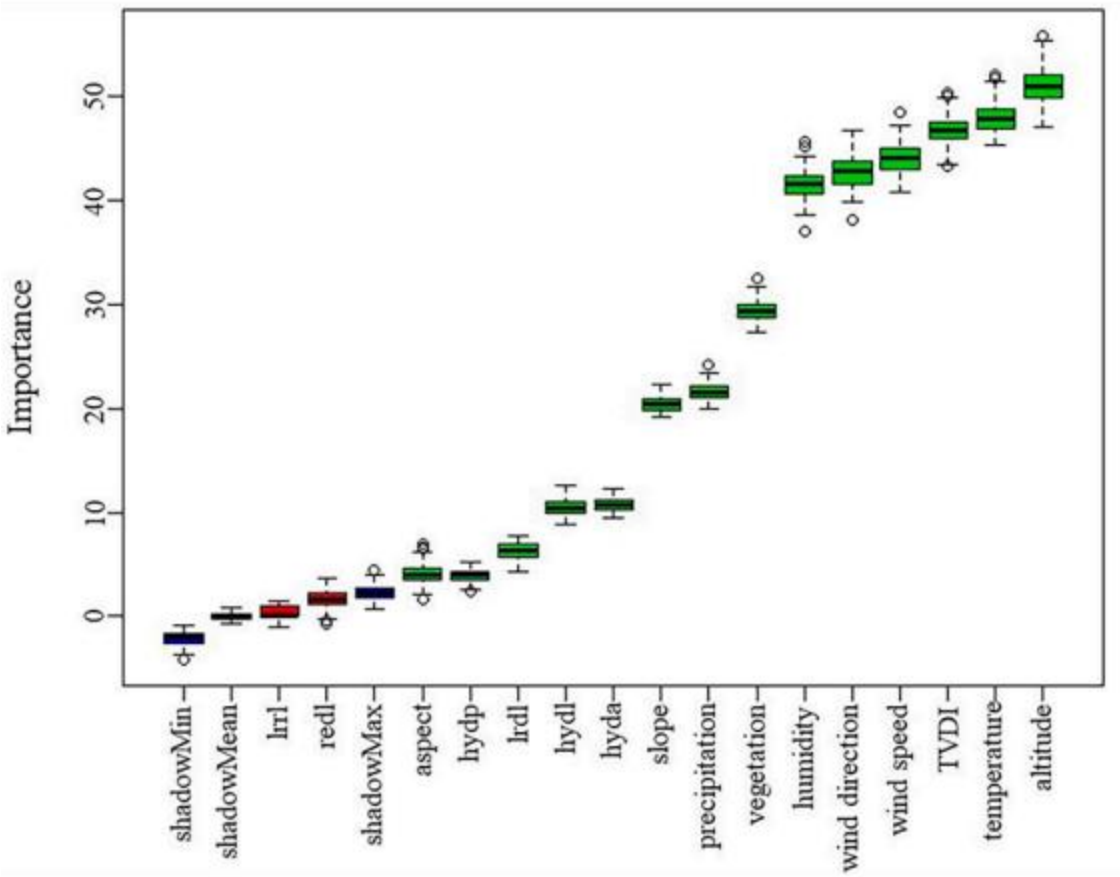

Figure 3.10: Variable sorting results obtained with the Boruta algorithm. [62]

After using Artificial Neural Networks (ANN) model and compare, it to the result of Wang Zhengfei-CA models. They found forest fire prediction results based on the artificial neural network were found to have average accuracy, sensitivity, and F-measure values of 85.02%, 95.26%, and 89.85%, respectively. Moreover, the constructed ANN model had a higher prediction accuracy than the combined models

Model	P	S	F-measure
ANN	0.85	0.95	0.90
CA	0.78	0.83	0.80

Figure 3.11: Performance comparison between the ANN and Wang Zhengfei-CA models. [63]

3. The Place of study

M'sila is one of the interior states, consisting of 15 districts and 47 municipalities. It shares its northern border with the states of Sétif and Bordj Bou Arreridj, while its western neighbors are Bouira and Médéa. To the south, it is adjacent to Djelfa and Biskra states, and to the east, it is bordered by the state of Batna. M'sila experiences a continental climate and lies between hilly and desert terrain. The majority of the state is flat, with elevations ranging from 200 to 300 meters above sea level.

The Wilaya of M'sila is dominated by a continental climate, which is characterized by hot and dry summers and moderately cold rains in winter.

3.1. Conservation of the M'sila forests:

- Creation decree: The conservation of forests in the wilaya of M'sila has been established by Resolution No. 95/333 of October 25, 1995. The state forest estate covers an area: 388,792 ha
- Number of provinces: 22 (M'Sila, Hammam Dalaa, Ouled Derradj, Sidi Aissa, Ben Srour, Bou Saada, Magra, Medjedel, Jebel Messaad, Ain El Hadjel)
- Headquarters: Al-Muwailiha district 1000 cooperative housing

3.1.1. Headquarters

Legal capacity: an institution of an administrative nature Under the supervision of the Ministry of Agriculture and Rural Development (General Directorate of Forests). At the state level, the domain of Al-Ghabi covers an area estimated at: 388,792 hectares, or 22.5% of the total area of M'sila.

The conservation headquarters includes four (04) departments and nine (09) offices which are:

- **Departments:**
  - Administration and resources
  - Department of Asset Management
  - The Wealth Expansion and Wealth Protection Department
  - Bureau of Forest Resources Development.

- **Offices:**
  - The Office of Regulation and Forest Police.
  - Studies and Programs Office.
  - Bureau of Inventory, Preparedness and Forest Products.
  - Plant and animal protection service: Office of Fire and Disease Prevention and Control.
  - Office of Protected Species, Hunting and Fishing Activities.
  - Land Reclamation Office.
  - Office of Human Resources Management
  - Budget and resources management office.

4. Methodology and data preparation

We obtained the data relating to the various fires of M'sila from the extracts of reports established by the Conservation of the forests of the wilaya. This data spanned a period from 2011 to May 2023 that included 83 wildfires.

5. Implementation

To evaluate and test the performance of the proposed approach, we will first have to go through the implementation stage. In this section, we describe the different tools and programming languages used.

5.1. Programming language

Nowadays, there are several programming languages and each language has its own characteristics. Among its languages, our choice focused on Python.

5.1.1. Python

Python is a widely used programming language known for its simplicity, versatility, and extensive libraries and frameworks. It has gained popularity in various fields, including data science and machine learning.

5.2. Development environment

- **PyCharm** (2023.1.2):

PyCharm is an integrated development environment (IDE) used for programming in Python. It provides code analysis, a graphical debugger, an integrated unit tester, integration with version control systems, and supports web development

- **TensorFlow(2.10.0):**

TensorFlow is a popular open-source machine learning frame- work developed by Google. It provides a comprehensive ecosystem of tools [28], libraries, and resources for building and deploying machine-learning models.

- **Sklearn(1.2.1):**

Scikit-learn (sklearn) is a comprehensive machine learning library for Python [29], providing a wide range of algorithms and tools for various machine learning tasks, such as classification, regression, clustering, and dimensionality re- duction.

- **NumPy(1.24.1):**

NumPy is a Python library for numerical computing that enables efficient and high-performance operations on multidimensional arrays [30], essential for tasks such as scientific computing, data analysis, and machine learning.

- **Pandas(1.5.3):**

Pandas is a widely used open-source library for data manipulation and analysis in Python [31]. It offers a user-friendly and efficient way to handle structured data using its Data-Frame data structure. With Pandas, you can easily clean, filter, transform, and aggregate data. It also provides robust tools for data visualization and seamless integration with other libraries. Pandas is a popular choice among data scientists and analysts for working with tabular data in Python.

6. Proposed models

Machine learning models have proven to be effective in predicting the burned area of forest fires. In this case section, we will explore three different models to predict forest fire spread area: The **Linear regression** model, **Polynomial Regression** model and **Random Forest Regression** model.

Conclusion:

Overall, the proposed models provided valuable insights into predicting the burned area of forest fires, each with its own strengths and limitations. While the ANN model excelled in capturing complex relationships, the Random Forest Classifier model highlighted its robustness and interpretability, and the SVM model demonstrated its effectiveness in handling non-linear relationships. The choice of model depends on the specific requirements of the task and the desired trade-offs between accuracy, interpretability, and computational complexity.

CHAPTER 4

EXPERIMENTAL RESULTS

INTRODUCTION

To evaluate the performance of the proposed models for predicting the burned area of forest fires, we conducted extensive experiments and obtained insightful results. In this section, we present the experimental results along with corresponding figures to illustrate the effectiveness of each model.

1. Methodology

1.1. Dataset Description

The dataset covers an extensive time period, encompassing a full year of fire spread. This duration enables us to conduct a comprehensive analysis of seasonal variations, long-term trends, and the impact of external factors on fire spread. With a prominent focus on accuracy and completeness, the dataset contains a wide range of variables that capture vital aspects of fire spread. These variables include:

- **Tree type**: The "Tree type" column can help assess how different tree species react to environmental conditions, including fire susceptibility. Certain tree species may be more prone to fire damage, while others may have fire-resistant properties. Understanding the tree types present in the dataset can be crucial for predicting fire behavior and its effects on various species.

- **Temperature**: Temperature measurements are vital for fire risk assessment. Higher temperatures can increase the likelihood of wildfires, and monitoring temperature

- Changes over time can provide insights into seasonal variations in fire risk.

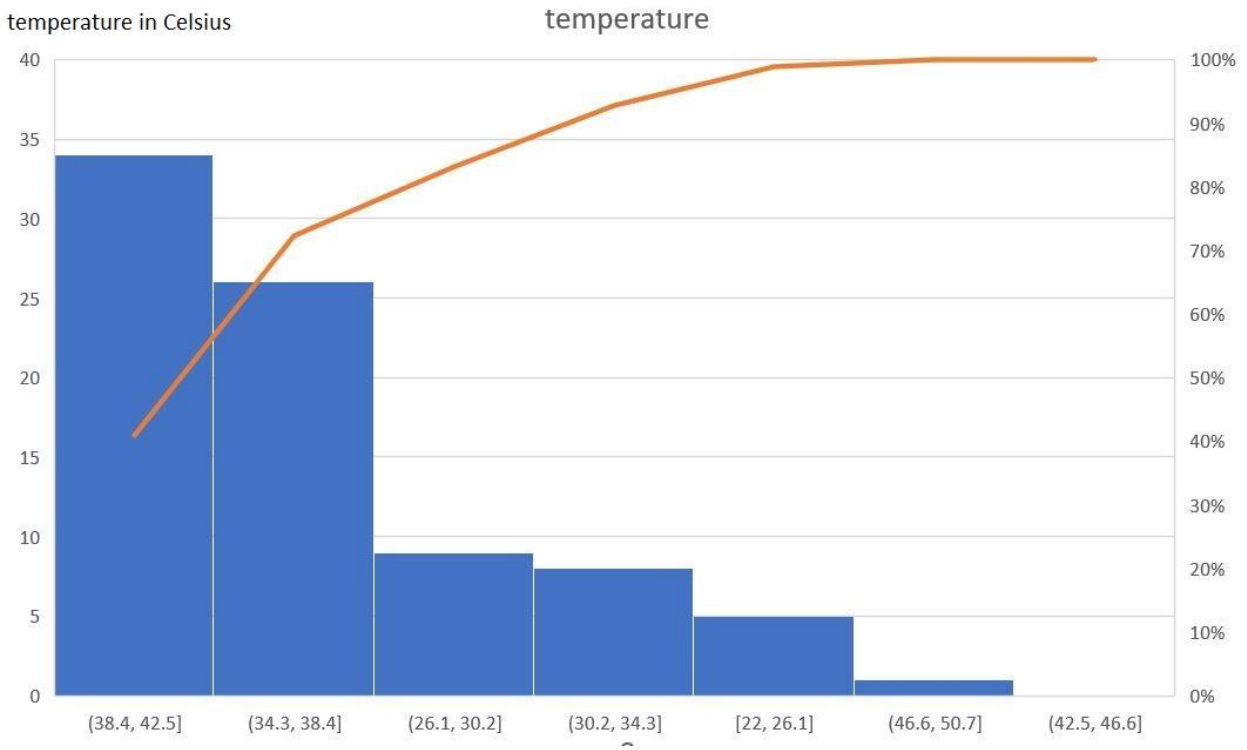

Figure 4.1: Temperature

- **Vents**: Wind measurements are essential for fire management. Strong winds can rapidly spread fires, so understanding wind patterns in relation to the location of trees can help predict fire spread and develop effective fire containment strategies.

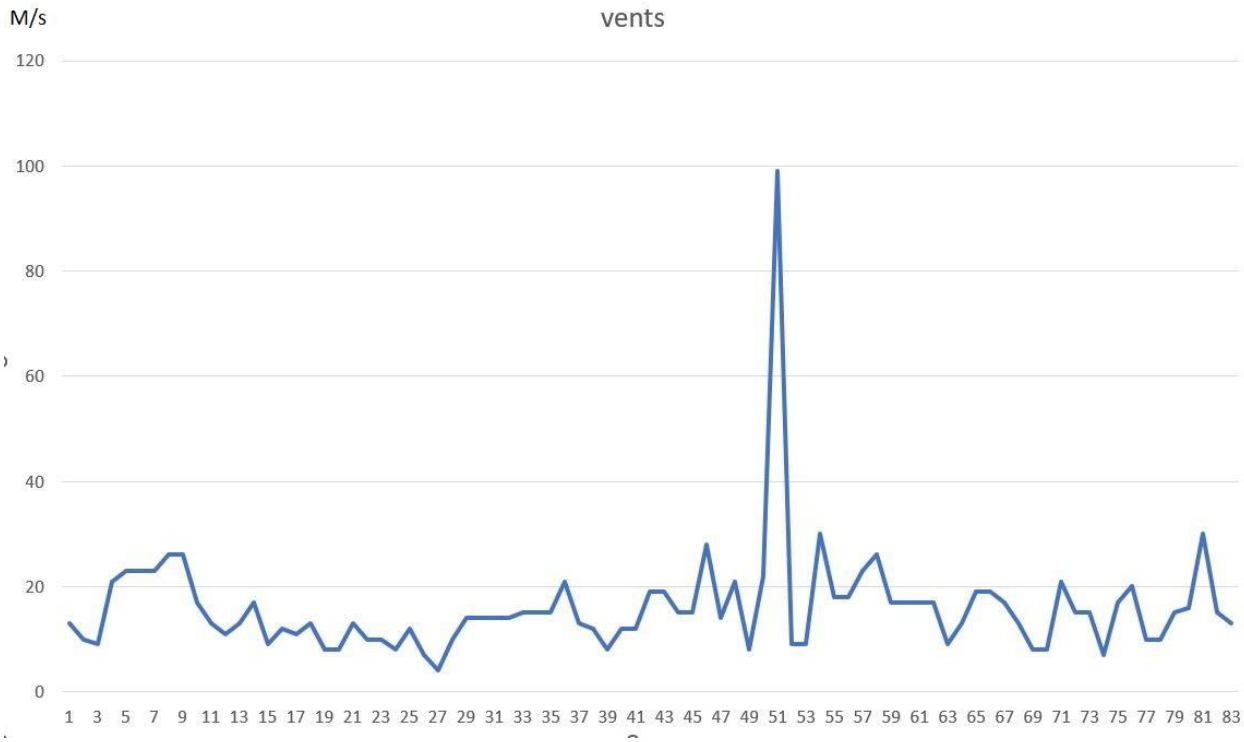

Figure 4.2: Vents

- **Humidity**: Humidity levels are significant in fire risk analysis. Lower humidity levels can contribute to dry conditions, which can increase the risk of wildfires. Monitoring humidity levels can aid in assessing the potential for fire ignition and spread.

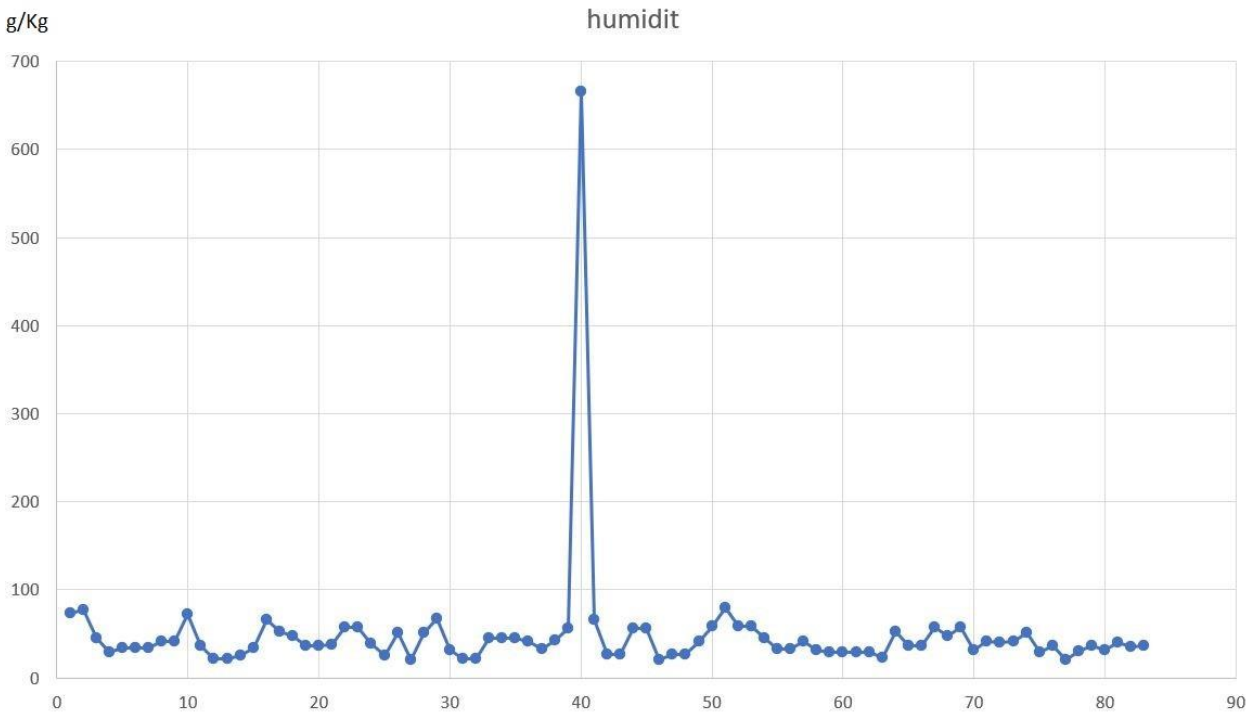

Figure 4.3: Humidity

- **Rate of Fire Spread**: This column measures the speed at which a fire is spreading one hour after it starts. It provides a crucial metric for assessing the immediate impact of fire on the environment. The rate of fire spread is influenced by various factors, including temperature, wind, humidity, surface conditions, and the types of trees present. Tracking this rate can help in understanding how rapidly a fire can propagate under specific conditions, aiding in fire management and mitigation strategies. With this new column, you can analyze how the rate of fire spread correlates with temperature, wind, humidity, surface conditions, tree types, and other variables in your dataset. This information is essential for predicting and managing wildfires effectively.

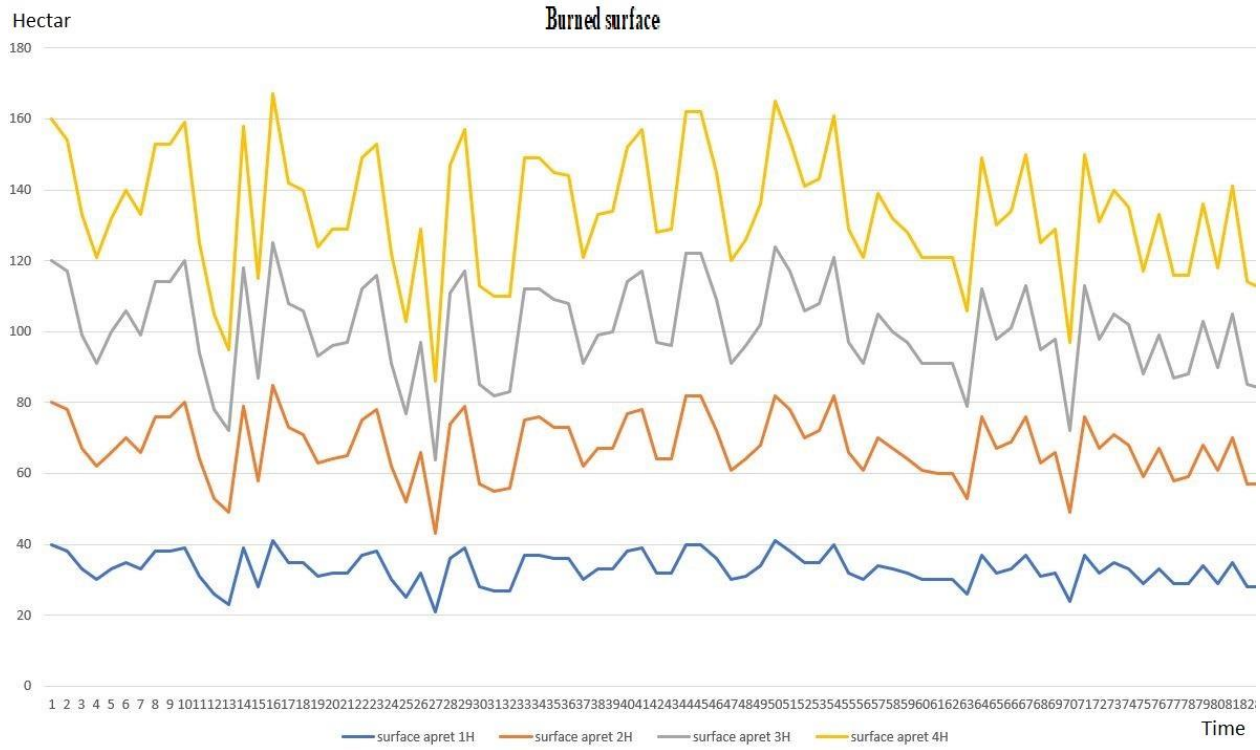

Figure 4.4: Burned surface

1.2.Sample from the dataset

daira	commune	nome foret	date	heure	type	temprqture	vents	prcipitations	humidit	surface apret 1H	surface apret 2H	surface apret 3H	surface apret 4H
hammam dalaa	hammam dalaa	dreat	09-06-11	14:53	foret	30	13	0	73	40	80	120	160
hammam dalaa	hammam dalaa	dreat	11-06-11	21:45	autres	25	10	0	77	38	78	117	154
hammam dalaa	hammam dalaa	dreat	06-07-11	21:45	alhalfa	35	9	0	45	33	67	99	133
magra	magra	righa dahra	07-07-11	14:30	maquis	37	21	0	29	30	62	91	121
hammam dalaa	hammam dalaa	dreat	18-07-11	17:00	maquis	38	23	0	34	33	66	100	132
hammam dalaa	hammam dalaa	dreat	18-07-11	17:00	broussailles	38	23	0	34	35	70	106	140
hammam dalaa	hammam dalaa	dreat	18-07-11	17:00	alhalfa	38	23	0	34	33	66	99	133
hammam dalaa	hammam dalaa	dreat	21-07-11	17:00	foret	39	26	0	42	38	76	114	153
hammam dalaa	hammam dalaa	dreat	21-07-11	17:00	maquis	39	26	0	42	38	76	114	153
hammam dalaa	hammam dalaa	dreat	30-09-11	15:13	maquis	26	17	15	72	39	80	120	159
hammam dalaa	hammam dalaa	el fedj	15-06-12	20:30	maquis	35	13	1	36	31	64	94	125
medjedel	medjedel	kef manaa	28-06-12	12:30	broussailles	38	11	0	22	26	53	78	105
ain el hadjel	ain el hadjel	el mergeb	30-06-12	09:00	broussailles	32	13	1	22	23	49	72	95
medjedel	medjedel	kef manaa	18-06-13	18:30	foret	40	17	0	26	39	79	118	158
magra	dehahna	souk ouled	19-06-13	15:00	maquis	39	9	0	34	28	58	87	115
bousaada	el hamel	barrage vert	24-06-13	17:10	foret	40	12	0	66	41	85	125	167
bousaada	el hamel	barrage vert	27-06-13	16:00	autres	37	11	1	52	35	73	108	142
hammam dalaa	hammam dalaa	dreat	31-07-13	19:20	foret	37	13	1	48	35	71	106	140
hammam dalaa	hammam dalaa	dreat	01-08-13	09:30	foret	42	8	0	36	31	63	93	124

Figure 4.5: Sample from the dataset

1.3.Relationships and connections

The relationship between wildfire spread and environmental factors such as wind, temperature, and humidity is fundamental to wildfire science and management. These factors collectively shape the behavior and trajectory of wildfires, influencing their speed and intensity.

Wind is a dominant driver of wildfire spread. Strong winds can propel flames and embers over considerable distances, making wildfires challenging to control. Wind direction further dictates the path of a fire, affecting the areas it endangers. Wind-driven fires tend to expand rapidly and unpredictably, posing substantial threats to both property and lives. Monitoring wind speed and direction is pivotal for evaluating potential fire trajectories and rates of spread.

Temperature directly affects fire dynamics. Elevated temperatures accelerate the drying of vegetation, increasing the risk of ignition. Higher temperatures can also lead to more aggressive fires that advance swiftly. Conversely, lower temperatures can decelerate fire growth, affording firefighters and emergency responders more control. Consistent monitoring of temperature patterns aids in anticipating periods of heightened fire risk and aids in strategically allocating firefighting resources.

Humidity levels regulate vegetation moisture content, significantly affecting wildfire susceptibility. Reduced humidity can desiccate plants, rendering them more prone to ignition. Dry fuels ignite readily and burn intensely. Conversely, higher humidity dampens fire behavior, rendering fires less aggressive and more manageable. Ongoing humidity monitoring offers valuable insights into potential ignition and fire spread risks.In summary, the intricate relationship between wildfire spread and environmental factors, including wind, temperature, and humidity, is pivotal. Comprehending how these variables interact is critical for predicting wildfire behavior, executing effective firefighting tactics, and mitigating the impacts on ecosystems and communities. Firefighters, meteorologists, and fire scientists closely scrutinize these factors to assess fire risk and act proactively to safeguard lives and property.

1.4.Data pre-processing

Data preprocessing is a vital step in data analysis that involves cleaning, transforming, and organizing the data before analysis. It aims to improve data quality, address missing values and outliers, and ensure the data is in a suitable format for analysis. By performing data preprocessing, analysts can enhance the accuracy and reliability of their results and derive meaningful insights from the data

1.4.1. Data Cleaning

In order to display the missing data, we write the following code:

```
print(data.DataFrame({'missing:': data.isnull().sum()}))
```

Missing data results**:**

We have 3 missing data in the humidit column.

We have 5 missing data in the vents column.

Since the missing data is not large, we will delete the row that contains missing data by writing the following code:

```
data = data[data[humidit].notna()]
data = data[data[vents].notna()]
```

In order to avoid analysis bias, we remove the duplicated lines we write the following code:

```
 data = data.drop_duplicates()
```

in order to discover Noise Data, we write the following code:

```
plt.figure()
data[['vents','temprqture','humidit']].boxplot()
plt.title("Noise Data")
plt.show()
```

And by doing that process, we achieved the following result:

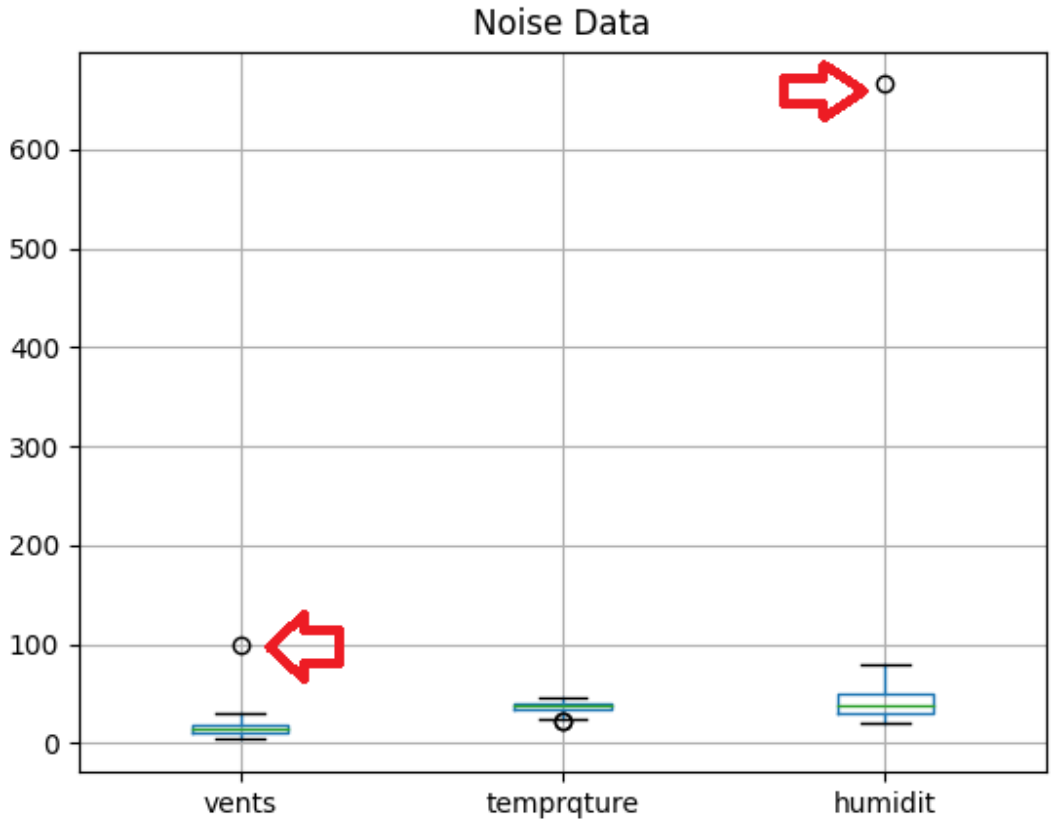

Figure 4.6: Noise Data Results

We notice that there are two outliers in Route column and NumOfPass column, to know these values, we write the following code:

```
NoiseData1 = data[data.humidit > 200]
NoiseData2 = data[data.vents > 80]
print(NoiseData1.humidit)
print(NoiseData2.vents)
```

In order to handle the abnormal values, we write the following code:

```
data.at[NoiseData1.index[0], 'humidit'] = 66
data.at[NoiseData2.index[0], 'vents'] = 9
```

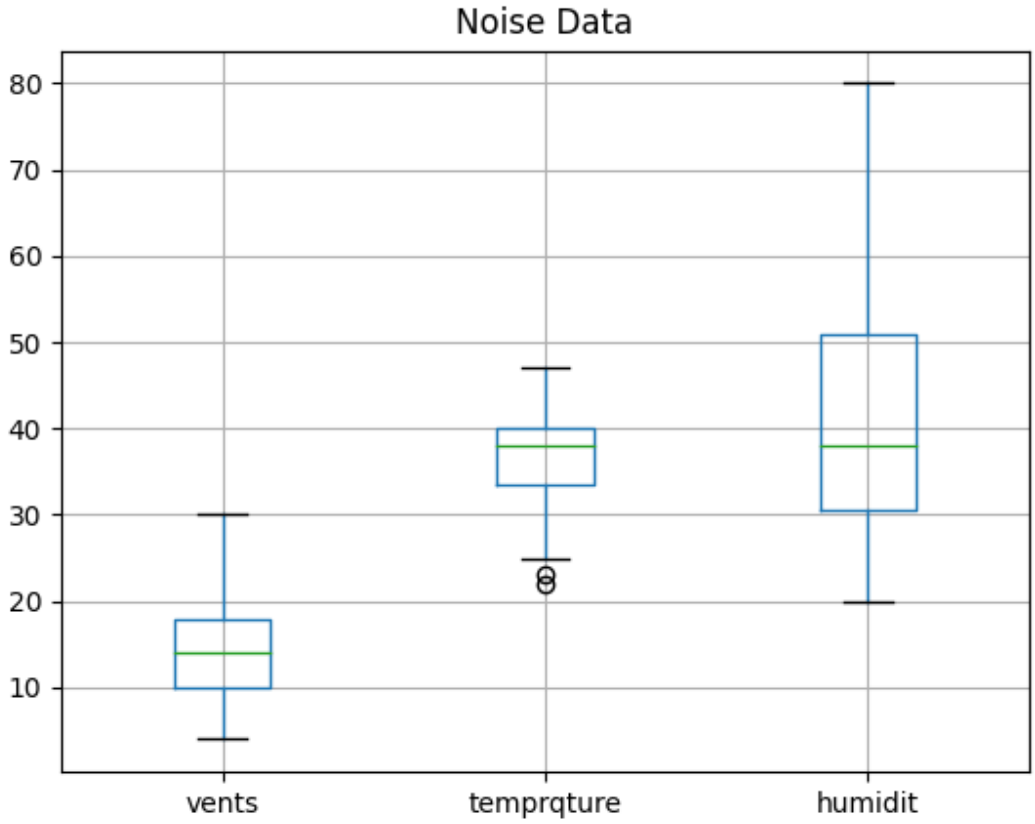

Figure 4.7: Noise Data after Cleaning

1.4.2. Data Selection

We selected a dataset about on fire spread from conservation of the M'sila forests. This dataset provides valuable information for analyzing the fires spread, and we delete the data we don't need with the following code:

```
data.drop(['daira'], axis=1, inplace=True)
data.drop(['commune'], axis=1, inplace=True)
data.drop(['nome foret'], axis=1, inplace=True)
data.drop(['date'], axis=1, inplace=True)
data.drop(['heure'], axis=1, inplace=True)
data.drop(['prcipitations'], axis=1, inplace=True)
```

1.4.3. Data Partitioning

Data partitioning is a critical step in model development, allowing for effective evaluation and validation. By splitting the dataset into different partitions, such as a training set, validation set, and test set, we can assess the performance and generalization capabilities of the model.

In our study, we allocated 80% of the dataset to the training set, and the remaining 20% to the test set. The larger proportion assigned to the training set ensures that the model has ample data to learn from and capture underlying patterns. The validation set is used to fine-tune the model's parameters and assess its performance during the training process. Finally, the test set, which remains unseen by the model during training, provides an unbiased evaluation of its ability to generalize to new, unseen data.

Below is the partition code:

```
X = data[['vents', 'temprqture', 'humidit']]
y = data['surface apret 4H']
```

The following display shows the training data and testing:

	vents	temprqture	humidit	target	
0	13	30	73	0	160
1	10	25	77	1	154
2	9	35	45	2	133
3	21	37	29	3	121
4	23	38	34	4	132
..	...	...	...		...
78	15	40	37	78	136
79	16	34	32	79	118
80	30	28	40	80	141
81	15	28	35	81	114
82	13	28	36	82	112

2. Regression models

Regression models are a class of statistical models used in data analysis and machine learning to explore the relationship between one or more independent variables (often called predictors or features) and a dependent variable (often called the target or outcome). The primary goal of regression analysis is to understand how changes in the independent variables are associated with changes in the dependent variable. This understanding can help in making predictions and explaining the observed data.There are various types of regression models, each suited to different types of data and research questions. Some common types of regression models include:

2.1.Linear Regression

```
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=42)

model = LinearRegression()
model.fit(X_train, y_train)
y_pred = model.predict(X_test)


mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)
print(f"Mean Squared Error: {mse}")
print(f"R-squared (R²): {r2}")

# Create a scatter plot of the actual vs. predicted values
plt.scatter(y_test, y_pred)
plt.xlabel("Actual")
plt.ylabel("Predicted")

# Plot the regression line
plt.plot([min(y_test), max(y_test)], [min(y_pred), max(y_pred)],
color='red', linewidth=2)
plt.show()
```

2.1.1. Model Description

Linear regression is a statistical method used for modeling the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship and aims to find the best-fit line to make predictions and understand the connections between variables.

2.1.2. Evaluation

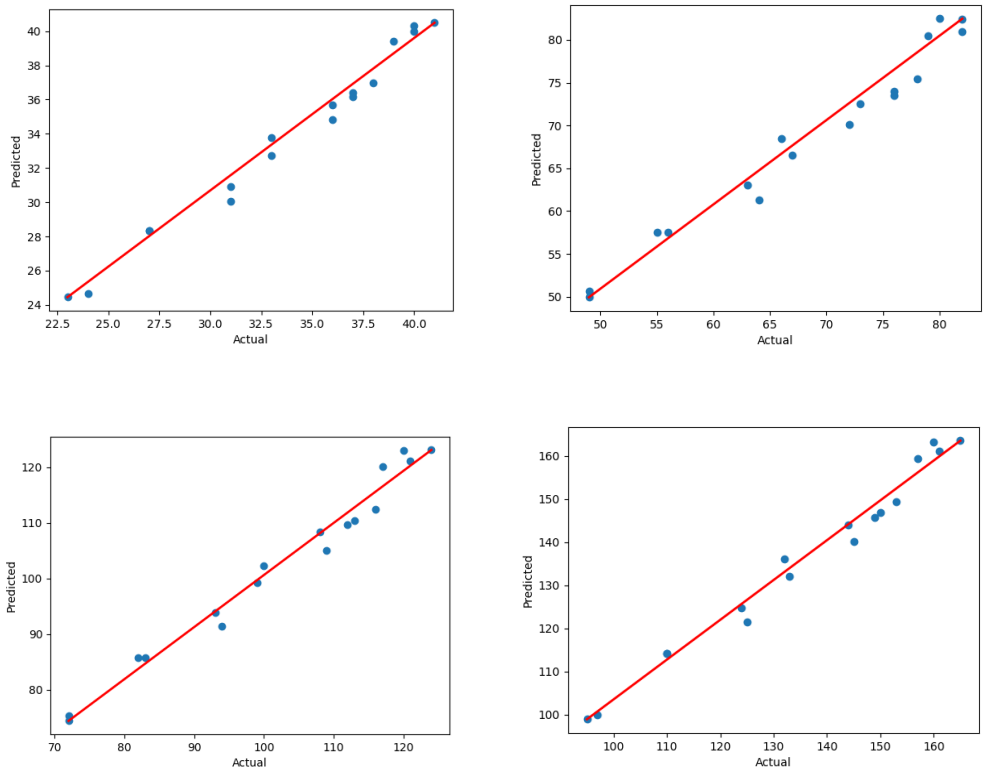

Figure 4.8: Evaluation of Linear Regression

Hypothesis 1 (H1):

H1 demonstrates a highly accurate model with a low Mean Squared Error (MSE) of 0.696 and a strong R-squared (R^2) value of 0.978. The low MSE indicates minimal prediction errors, while the R^2 value suggests that 97.75% of the variance in the dependent variable is explained by the independent variables. This model provides robust and reliable predictions.

Hypothesis 2 (H2):

H2 results in a reasonably low MSE of 3.304 and a solid R-squared (R^2) value of 0.972. The MSE implies modest prediction errors, and the R^2 value indicates that 97.19% of the

variation in the dependent variable is accounted for by the independent variables. This model offers a strong relationship between predictors and the target variable.

Hypothesis 3 (H3):

H3 exhibits a MSE of 6.546 and an impressive R-squared (R^2) value of 0.976. The MSE represents relatively small prediction errors, and the R^2 value shows that approximately 97.59% of the dependent variable's variation can be attributed to the independent variables. This model demonstrates a robust fit.

Hypothesis 4 (H4):

H4 displays a MSE of 9.821 and an outstanding R-squared (R^2) value of 0.979. The MSE, though slightly higher, still signifies reasonable prediction accuracy. The R^2 value of 0.979 implies that around 97.95% of the variance in the dependent variable is explained by the independent variables. This model provides a strong relationship between predictors and the target variable, with exceptional explanatory power.

In summary, all four hypotheses indicate the effectiveness of the linear regression models, with varying levels of prediction accuracy and explainability. These results offer valuable insights into the quality and suitability of the models for their respective analyses or prediction tasks.

2.2.Polynomial Regression

```
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=42)

# Create polynomial features
degree = 2  # Set the degree of the polynomial
poly_features = PolynomialFeatures(degree=degree)
X_train_poly = poly_features.fit_transform(X_train)
X_test_poly = poly_features.transform(X_test)

# Create and train a Polynomial Regression model
poly_model = LinearRegression()
poly_model.fit(X_train_poly, y_train)
y_pred = poly_model.predict(X_test_poly)

# Calculate Mean Squared Error and R-squared (R²)
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)
print(f"Mean Squared Error: {mse}")
print(f"R-squared (R²): {r2}")
```

```
# Visualize the polynomial fit
plt.scatter(X_test['temprqture'], y_test, label="Actual",
color='blue')
plt.scatter(X_test['temprqture'], y_pred, label="Predicted",
color='red')
plt.xlabel("temprqture")
plt.ylabel("Surface Apret 4H")
plt.legend()

# Sort X_test and y_pred by X_test['vents'] for smoother curve
plotting
sorted_indices = np.argsort(X_test['temprqture'])
X_test_sorted = X_test.iloc[sorted_indices]
y_pred_sorted = y_pred[sorted_indices]

plt.plot(X_test_sorted['temprqture'], y_pred_sorted,
color='green', linewidth=2, label="Polynomial Fit")
plt.legend()

plt.show()
```

2.2.1. Model Description

Polynomial regression is an extension of linear regression that models non-linear relationships between variables. It fits a polynomial equation to the data, allowing for curved relationships to be captured. The degree of the polynomial determines the complexity of the curve, and careful selection is needed to avoid overfitting or underfitting. This model is used when linear regression cannot adequately describe the data's behavior.

2.2.2. Evaluation

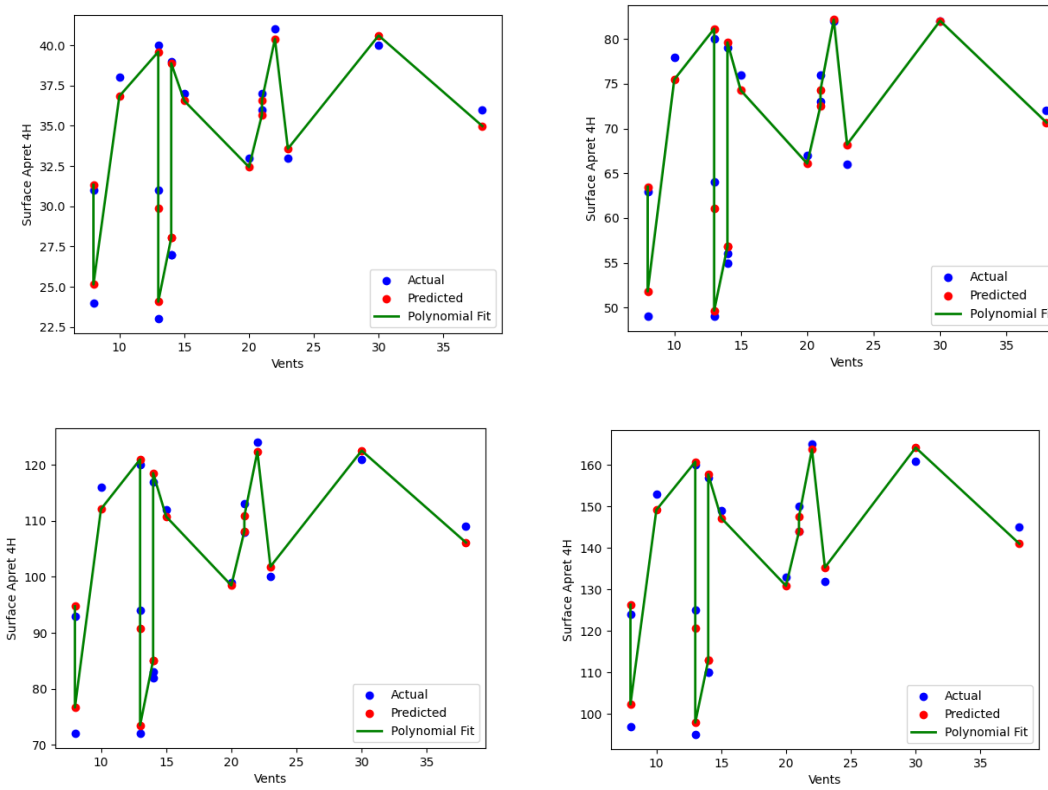

Figure 4.9: Evaluation of Polynomial Regression

Hypothesis 1 (H1):

In H1, the polynomial regression model demonstrates exceptional performance with a low Mean Squared Error (MSE) of 0.614 and an impressive R-squared (R^2) value of 0.980. The low MSE indicates minimal prediction errors, while the high R^2 value suggests that approximately 98.01% of the variance in the dependent variable is explained by the polynomial relationship. These results indicate a strong and reliable fit, with minimal errors in prediction.

Hypothesis 2 (H2):

H2 yields a reasonably low Mean Squared Error (MSE) of 2.515 and a solid R-squared (R^2) value of 0.979. The MSE represents moderate prediction errors, while the R^2 value indicates that around 97.86% of the dependent variable's variation is accounted for by the polynomial equation. This model provides a robust relationship between the predictors and the target variable.

Hypothesis 3 (H3):

In H3, the polynomial regression model exhibits a MSE of 5.460 and a strong R-squared (R^2) value of 0.980. The MSE signifies relatively small prediction errors, and the R^2 value shows that approximately 97.99% of the variation in the dependent variable is attributed to the polynomial relationship. This demonstrates a solid model fit and an effective representation of the relationships among the variables.

Hypothesis 4 (H4):

H4 displays a MSE of 8.543 and an outstanding R-squared (R^2) value of 0.982. The MSE, though slightly higher, still represents reasonable prediction accuracy. The R^2 value of 0.982 implies that approximately 98.21% of the variance in the dependent variable is explained by the polynomial equation. This model provides a strong relationship between predictors and the target variable, with exceptional explanatory power.

In summary, all four hypotheses indicate the effectiveness of the polynomial regression models in explaining the relationships between the independent and dependent variables. They exhibit varying degrees of prediction accuracy and explainability, with H4 standing out as particularly robust in its performance. These results offer valuable insights into the quality of the models and their suitability for the intended analysis or prediction tasks.

2.3.Random Forest Regression

```
# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=42)

# Create and train a Random Forest Regression model
n_estimators = 100  # You can adjust the number of estimators as
needed
random_forest_model =
RandomForestRegressor(n_estimators=n_estimators,
random_state=42)
random_forest_model.fit(X_train, y_train)
y_pred = random_forest_model.predict(X_test)

# Calculate Mean Squared Error and R-squared (R²)
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)
```

```
print(f"Mean Squared Error: {mse}")
print(f"R-squared (R²): {r2}")

# Create a scatter plot of the actual vs. predicted values
plt.scatter(y_test, y_pred)
plt.xlabel("Actual")
plt.ylabel("Predicted")

# Add a diagonal line representing perfect predictions
min_val = min(min(y_test), min(y_pred))
max_val = max(max(y_test), max(y_pred))
plt.plot([min_val, max_val], [min_val, max_val], color='red',
linestyle='--', linewidth=2, label="Perfect Predictions")

plt.legend()
plt.show()
```

2.3.1. Model Description

Random Forest Regression is a powerful machine learning technique that combines the predictions of multiple decision trees to improve accuracy and reduce overfitting. It is flexible, robust, and suitable for various regression tasks. This ensemble method provides stable and accurate predictions, making it valuable for data analysis and prediction tasks across different domains.

2.3.2. Evaluation

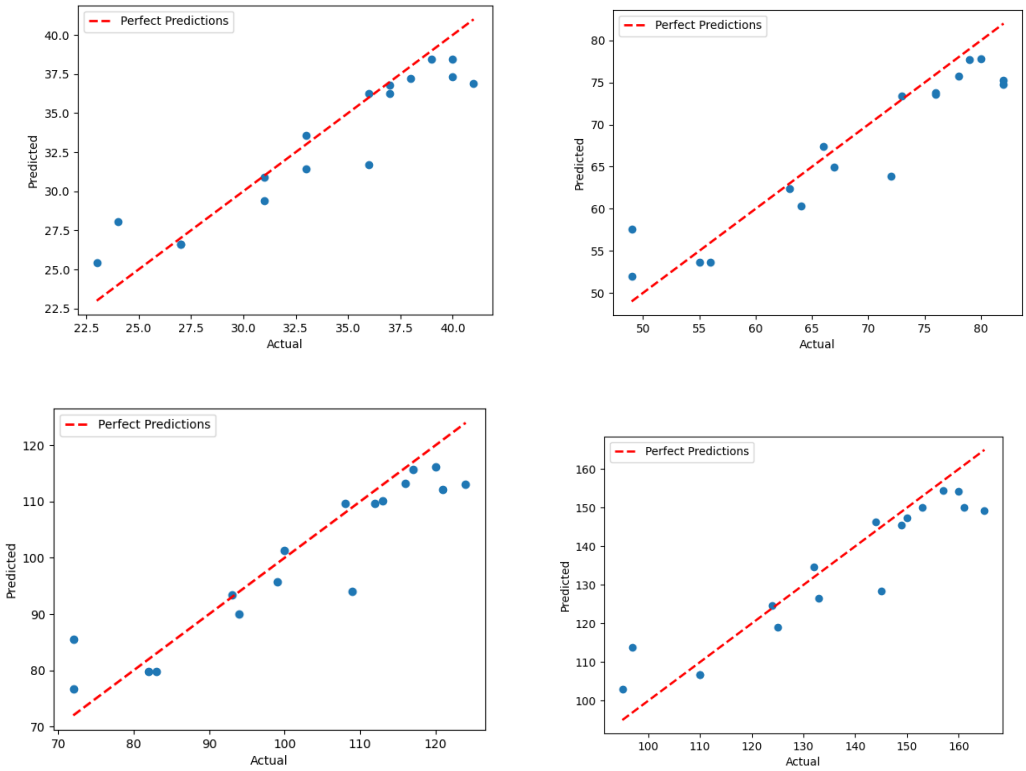

Figure 4.10: Evaluation of Random Forest Regression

Hypothesis 1 (H1):

In H1, the Random Forest Regression model demonstrates reasonable performance with a Mean Squared Error (MSE) of 4.392 and an R-squared (R^2) value of 0.858. The MSE represents moderate prediction errors, and the R^2 value indicates that approximately 85.79% of the variance in the dependent variable is explained by the model. While not perfect, these results suggest a fairly strong fit and an ability to capture the relationships between predictors and the target variable.

Hypothesis 2 (H2):

H2 yields a higher Mean Squared Error (MSE) of 17.446 and an R-squared (R^2) value of 0.852. The MSE indicates relatively larger prediction errors compared to H1, and the R^2 value suggests that around 85.17% of the dependent variable's variation is accounted for by the model. Despite slightly increased errors, this model still provides a robust representation of the relationships.

Hypothesis 3 (H3):

In H3, the Random Forest Regression model exhibits a higher MSE of 41.763 and an R-squared (R^2) value of 0.846. The MSE reflects larger prediction errors, and the R^2 value shows that approximately 84.62% of the variation in the dependent variable is attributed to the model. While prediction errors are more pronounced, this model maintains a solid representation of the underlying relationships.

Hypothesis 4 (H4):

H4 displays a further increased MSE of 68.434 and an R-squared (R^2) value of 0.857. The MSE, though higher, still signifies reasonable prediction accuracy. The R^2 value of 0.857 implies that approximately 85.68% of the variance in the dependent variable is explained by the model. This model provides a strong relationship between predictors and the target variable, despite slightly larger prediction errors.

In summary, all four hypotheses indicate the effectiveness of the Random Forest Regression models in explaining the relationships between the independent and dependent variables. They exhibit varying degrees of prediction accuracy and explainability, with H1 standing out as particularly robust in its performance. These results offer valuable insights into the quality of the models and their suitability for the intended analysis or prediction tasks.

3. Results

3.1. Random Forest Regression

H1: The Random Forest Regression model in H1 exhibits reasonable performance with moderate prediction errors (MSE = 4.392) and explains approximately 85.79% of the variance in the dependent variable ($R^2 = 0.858$).

H2: H2 shows a similar trend with slightly larger prediction errors (MSE = 17.446) but still provides a robust representation of relationships ($R^2 = 0.852$).

H3: In H3, the model has larger prediction errors (MSE = 41.763) but maintains a solid representation ($R^2 = 0.846$).

H4: H4 displays a further increase in prediction errors (MSE = 68.434) but maintains a strong relationship ($R^2 = 0.857$).

3.2. Polynomial Regression

H1: The Polynomial Regression model in H1 demonstrates exceptional performance with low prediction errors (MSE = 0.614) and explains approximately 98.01% of the variance in the dependent variable ($R^2 = 0.980$).

H2: H2 maintains solid performance with moderate prediction errors (MSE = 2.515) and strong explanatory power ($R^2 = 0.979$).

H3: In H3, the model exhibits larger prediction errors (MSE = 5.460) but maintains a robust representation ($R^2 = 0.980$).

H4: H4 continues to perform well with reasonable prediction errors (MSE = 8.543) and strong relationship ($R^2 = 0.982$).

3.3.Linear Regression

H1: The Linear Regression model in H1 demonstrates good performance with moderate prediction errors (MSE = 0.696) and explains approximately 97.75% of the variance in the dependent variable ($R^2 = 0.977$).

H2: H2 maintains solid performance with moderate prediction errors (MSE = 3.304) and strong explanatory power ($R^2 = 0.972$).

H3: In H3, the model exhibits larger prediction errors (MSE = 6.546) but maintains a robust representation ($R^2 = 0.976$).

H4: H4 continues to perform well with reasonable prediction errors (MSE = 9.821) and a strong relationship ($R^2 = 0.979$).

3.4.Comparison

Among the three models, Polynomial Regression consistently outperforms the others in terms of both Mean Squared Error (MSE) and R-squared (R^2) values. It consistently achieves the lowest MSE and the highest R^2, indicating the best overall fit and predictive power.

Random Forest Regression and Linear Regression perform reasonably well but have larger prediction errors and slightly lower R^2 values compared to Polynomial Regression.

Based on these results, Polynomial Regression is the best-performing model among the three for the given dataset and hypotheses.

It's important to note that the choice of the "best" model also depends on the specific context of the analysis and the trade-offs between prediction accuracy and model complexity. However, in this comparison, Polynomial Regression stands out as the top performer.

Conclusion

In this chapter, we applied and compared Linear Regression, Polynomial Regression, and Random Forest Regression models to predict forest fire spread.

We found that Polynomial Regression consistently outperforms the other models, providing more accurate predictions with lower MSE and higher R^2 values.

This research demonstrates the potential of AI-driven regression models for improving forest fire spread prediction, which can be invaluable for mitigating the impact of forest fires and safeguarding our environment. Further research may explore the integration of real-time data and other advanced machine learning techniques to enhance prediction accuracy even further.

CONCLUSION

In this thesis, we have examined the significance of early detection in minimizing the impact of wildfires, highlighting the need to reduce material and human losses.

In the first chapter, we presented a comprehensive examination of forest fires, including their definition, types, characteristics, causes, contributing factors, consequences, and prevention methods. We also present statistics related to this disaster, with a particular focus on the situation in Algeria.

The second chapter focused on the role of computer science, specifically artificial intelligence, in developing solutions for effective wildfire detection and risk reduction. We provided a comprehensive overview of artificial intelligence and machine learning concepts relevant to this field.

In the third chapter, we presented our proposed approaches and architecture for wildfire prediction. Our approaches employed three pre-trained algorithms: Linear regression, Polynomial Regression and Random Forest Regression. We selected these algorithms based on their widespread usage and high accuracy compared to other supervised learning approaches.

Through our experimentation, we demonstrated the efficacy of our models in predicting forest fires spread. Notably, Polynomial Regression model outperformed the others.

Overall, our research highlights the potential of machine learning algorithms in effectively predicting forest fires spread. By leveraging these techniques, we aim to enhance early detection capabilities and contribute to reducing the devastating impacts of wildfires.

REFERENCES:

[1] Britannica. (n.d.). Fire. In Encyclopædia Britannica.
https://www.britannica.com/science/fire-combustion

[2] Food and Agriculture Organization of the United Nations. (2012). Global Forest Resources Assessment 2010: Definitions and Methodology. Rome.
http://www.fao.org/docrep/013/i1757e/i1757e00.htm

[3] National Interagency Fire Center. (n.d.). Wildfire Information.
https://www.nifc.gov/fire-information/

[4] UNECE and FAO. (2015). Forest Types and Classification. In Manual on Definitions and Classifications of Forests (p. 9-20).
https://www.unece.org/fileadmin/DAM/timber/publications/SP-27.pdf

[5] World Wildlife Fund (WWF). (n.d.). Tropical Rainforests.
https://www.worldwildlife.org/habitats/tropical-rainforests/

[6] National Geographic Society. (n.d.). Temperate Forest. In National Geographic Society Encyclopedia. https://www.nationalgeographic.org/encyclopedia/temperate-forest/

[7] Canadian Encyclopedia. (n.d.). Boreal Forest.
https://www.thecanadianencyclopedia.ca/en/article/boreal-forest/

[8] WWF. (n.d.). Mediterranean Forests.
https://www.worldwildlife.org/ecoregions/pa1215/

[9] UNESCO. (n.d.). Montane Forests and Associated Wetlands. In World Heritage Papers No. 13 - Mountains: Sources of Water, Sources of Knowledge (p. 78).
https://unesdoc.unesco.org/ark:/48223/pf0000263105/

[10] Food and Agriculture Organization (FAO).
https://www.fao.org/3/Y1997E/y1997e1m.htm

[11] Rivas-Martínez, S., et al. (2011). Worldwide Bioclimatic Classification System. Global Geobotany, 1, 1-634. http://www.globalbioclimatics.org/book/worldwide_bioclimatics.pdf

[12] Blondel, J., Aronson, J., & Bodiou, J. (2010). The Mediterranean Region: Biological Diversity in Space and Time. Oxford: Oxford University Press.

[13] Joint Nature Conservation Committee (JNCC). (2013). Upland Habitats.

https://jncc.gov.uk/our-work/upland-habitats

[14] National Park Service. (n.d.). Fire Regimes: Surface Fire.

https://www.nps.gov/articles/fire-regimes-surface-fire.htm

[15] National Park Service. (n.d.). Fire Regimes: Crown Fire.

https://www.nps.gov/articles/fire-regimes-crown-fire.htm

[16] U.S. Forest Service. (n.d.). Fire Effects: Ground Fires.

https://www.fs.fed.us/science/cenfir/ground-fires.shtml

[17] Glossary of Meteorology. (n.d.). American Meteorological Society.

https://glossary.ametsoc.org/wiki/Meteorological_conditions

[18] Larcher, W. (2003). Physiological Plant Ecology: Ecophysiology and Stress Physiology of Functional Groups (4th ed.). Springer.

[19] Russell, S., Norvig, P., Davis, E., & Genesereth, M. (2021). Artificial Intelligence: A Modern Approach (4th ed.). Pearson.

[20] Russell, S., Norvig, P., Davis, E., & Genesereth, M. (2021). Artificial Intelligence: A Modern Approach (4th ed.). Pearson.

[21] Mitchell, T. M. (1997). Machine Learning. McGraw-Hill.

[22] Jordan, M. I., & Mitchell, T. M. (2015). Machine Learning: Trends, Perspectives, and Prospects. Science, 349(6245), 255-260.

[23] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer.

[24] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer.

[25] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer.

[26] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer.

[27] Sutton, R. S., & Barto, A. G. (2018). Reinforcement Learning: An Introduction (2nd ed.). The MIT Press.

[28] Riau Forest Fire Prediction using Supervised Machine Learning - IOPscience https://iopscience.iop.org/article/10.1088/1742-6596/1566/1/012002

[29] Sustainability | Free Full-Text | Mapping Forest Fire Risk Zones Using Machine Learning Algorithms in Hunan Province, China (mdpi.com)
https://www.mdpi.com/2071-1050/15/7/6292
[30] Forests | Free Full-Text | Forest Fire Probability Mapping in Eastern Serbia: Logistic Regression versus Random Forest Method (mdpi.com)
https://www.mdpi.com/1999-4907/12/1/5
[31] 2 Overview of the hierarchical interaction levels in temperate forest...
https://www.researchgate.net/figure/Overview-of-the-hierarchical-interaction-levels-in-temperate-forest-ecosystems-and-the_fig2_281813304
[32] (2) Facebook
https://www.facebook.com/SecondNatureMB/photos/a.191404680880179/3926629827357627/?type=3
[33] FAO - Forestry - Workshop on Tropical Secondary Forest Management in Africa: Reality and Perspectives https://www.fao.org/3/j0628e/J0628E51.htm
[34] Introduction to Froestry (slideshare.net)
https://www.slideshare.net/parrc/introduction-to-froestry
[35] Supervised, Unsupervised and Semi-supervised Learning (enjoyalgorithms.com)
https://www.enjoyalgorithms.com/blogs/supervised-unsupervised-and-semisupervised-learning
[36] Supervised, Unsupervised and Semi-supervised Learning (enjoyalgorithms.com)
https://www.enjoyalgorithms.com/blogs/supervised-unsupervised-and-semisupervised-learning
[37] Supervised Learning Algorithms: Linear Regression – Query
https://www.query.ai/resources/blogs/supervised-learning-algorithms-linear-regression/
[38] Supervised, Unsupervised and Semi-supervised Learning (enjoyalgorithms.com)
https://www.enjoyalgorithms.com/blogs/supervised-unsupervised-and-semisupervised-learning
[39] Logistic Regression in machine learning - Pianalytix: Build Real-World Tech Projects
https://pianalytix.com/logistic-regression-in-machine-learning-2/
[40] Decision Tree, Random Forest and XGBoost on Arduino (eloquentarduino.github.io)
https://eloquentarduino.github.io/2020/10/decision-tree-random-forest-and-xgboost-on-arduino/
[41] Naive Bayes and Text Classification (sebastianraschka.com)
https://sebastianraschka.com/Articles/2014_naive_bayes_1.html
[42] Learning Data Science: Day 11 - Support Vector Machine | by Haydar Ali Ismail | Medium

https://haydar-ai.medium.com/learning-data-science-day-11-support-vector-machine-8ef06da91bfc

[43] K-Nearest Neighbors (KNN) – Theory (datasciencelovers.com)

http://www.datasciencelovers.com/machine-learning/k-nearest-neighbors-knn-theory/

[44] Supervised, Unsupervised and Semi-supervised Learning (enjoyalgorithms.com)

https://www.enjoyalgorithms.com/blogs/supervised-unsupervised-and-semisupervised-learning

[45] Machine Learning Crash Course, Part II: Unsupervised Machine Learning | Leverege

https://www.leverege.com/blogpost/machine-learning-course-iot

[46] Two main applications of unsupervised learning: clustering and... | Download Scientific Diagram (researchgate.net)

https://www.researchgate.net/figure/Two-main-applications-of-unsupervised-learning-clustering-and-dimensionality-reduction_fig2_344783581

[47] DataVisor https://www.datavisor.com/blog/rules-engines-learning-models-and-beyond/

[48] Apriori Algorithm for Association Rule Learning — How To Find Clear Links Between Transactions | by Saul Dobilas | Towards Data Science

https://towardsdatascience.com/apriori-algorithm-for-association-rule-learning-how-to-find-clear-links-between-transactions-bf7ebc22cf0a

[49] Example of unsupervised learning: self organizing map. | Download Scientific Diagram (researchgate.net)

https://www.researchgate.net/figure/Example-of-unsupervised-learning-self-organizing-map_fig2_326557847

[50] Supervised, Unsupervised and Semi-supervised Learning (enjoyalgorithms.com)

https://www.enjoyalgorithms.com/blogs/supervised-unsupervised-and-semisupervised-learning

[51] Supervised, Unsupervised and Semi-supervised Learning (enjoyalgorithms.com)

https://www.enjoyalgorithms.com/blogs/supervised-unsupervised-and-semisupervised-learning

[52] FireCast: Leveraging Deep Learning to Predict Wildfire Spread | IJCAI

https://www.ijcai.org/Proceedings/2019/636#:~:text=FireCast%20combines%20artificial%20intelligence%20%28AI%29%20techniques%20with%20data,historical%20fire%20data%20and%20using%20modest%20computational%20resources.

[53] Next Day Wildfire Spread: A Machine Learning Data Set to Predict Wildfire Spreading from Remote-Sensing Data (arxiv.org) https://arxiv.org/abs/2112.02447

[54] Simulation of forest fire spread based on artificial intelligence – ScienceDirect https://www.sciencedirect.com/science/article/pii/S1470160X22001248

[55] Machine Learning — Prediction Algorithms — Polynomial Regression — Part 4 | by Ekrem Hatipoglu | Medium https://medium.com/@ekrem.hatipoglu/machine-learning-prediction-algorithms-polynomial-regression-part-4-6c62b4240b53

[56] Decision Tree for Regression — The Recipe | by Akshaya Sriram | Analytics Vidhya | Medium https://medium.com/analytics-vidhya/decision-tree-for-regression-the-recipe-74f7628b8a0

[57] 9 Types of Regression Analysis (in ML & Data Science) | FavTutor https://favtutor.com/blogs/types-of-regression

[58] Introduction to Ridge Regression – Statology https://www.statology.org/ridge-regression/

[59] Unfolding the Maths behind Ridge and Lasso Regression https://www.analyticsvidhya.com/blog/2020/11/lasso-regression-causes-sparsity-while-ridge-regression-doesnt-unfolding-the-math/

Abstract:

In this study, three machine-learning algorithms (Linear regression, Polynomial Regression, and Random Forest Regression) were explored for predicting forest fires spread. A comprehensive dataset consisting of environmental and weather factors influencing forest fires was collected and used to train and test the models. Performance metrics such as accuracy, precision and recall score were used to evaluate the models. The results showed that all three algorithms performed well, but the Polynomial Regression Model achieved the highest accuracy. This study emphasizes the effectiveness of machine learning in forest fire spread prediction, particularly the superiority of the Polynomial Regression Model, and highlights the importance of leveraging advanced techniques for mitigating the impact of forest fires and protecting ecosystems.

الملخص:

في هذه الدراسة، تم استكشاف ثلاثة خوارزميات لتعلم الآلة (الانحدار الخطي, الانحدار متعدد الحدود والانحدار العشوائي للغابات) لتوقع انتشار حرائق الغابات. تم جمع مجموعة بيانات شاملة تتضمن عوامل بيئية وجوية تؤثر على حوادث الحرائق في الغابات واستخدمت لتدريب واختبار النماذج. تم استخدام معايير الأداء مثل والنقاوة والدقة العالية وإعادة الاستدعاء لتقييم النماذج. أظهرت النتائج أن جميع الخوارزميات الثلاثة قد قدمت أداء جيدًا، ولكن نموذج الانحدار متعدد الحدود حقق أعلى دقة. تسلط هذه الدراسة الضوء على فعالية تعلم الآلة في توقع انتشار حرائق الغابات، وخاصة تفوق نموذج الانحدار متعدد الحدود ، وتبرز أهمية استغلال التقنيات المتقدمة للحد من تأثير حرائق الغابات وحماية النظم البيئية.