POLITECNICO DI TORINO

Master's degree in Mathematical Engineering

Master's Degree Thesis

Wildfire Risk Evaluation using Machine Learning Techniques on Satellite Multispectral Data

Supervisor

Prof. Barbara Caputo

PhD. Fabio Cermelli

Company Tutor

Giuseppe Ruggiero

Candidate

Simonetta Bodojra

A.Y. 2021/2022

Abstract

Since climate change has continued to raise global temperatures, wildfires are becoming a greater hazard with increased severity and frequency. To help fire management agencies better plan their preventive measures and increase the effectiveness of suppression, it is crucial to monitor and map fire-susceptible areas through a deep analysis of the regional fire trends.

The aim of this thesis is to test the ability of supervised machine learning methods to map and measure the risk of fires in the vegetative areas of Sicily in 2021 using data from 2016 to 2020. This task was performed on areas of 200x200 square meters by creating a novel dataset with variables representing potential fire drivers: weather, topography, and fuel.

Fuel type and moisture content are modeled using a collection of spectral indices taking advantage of open source multispectral data from Sentinel-2 mission (ESA). Despite the low spatial resolution, the results are encouraging and this work could represent a starting point to build a solid fire management Italian system.

Table of Contents

Chapter 1

Introduction

Historically, it is acknowledged that forest fires are essential for forest renewal or to reduce build-up of fuel and thus to control future fire intensity. Nevertheless, extensive and frequent fires can cause economic damages and loss of human lives in populated areas and can have negative impact on biodiversity both on air and water quality. Nowadays, they are a major environmental problem which is going to worsen in the next years in light of the current worldwide climate situation. Especially during hot summers, this increasing number of wildfires is the result of prolonged droughts, heatwaves and altered meteorological patterns.

According to EFFIS (European Forest Fire Information System) [1], in 2021 Italy was the European country with the highest number of fires and the second most severely damaged in terms of burned area, after Turkey (Figure 1.1). Indeed, the overall burned area, mapped from 1422 fires, was the biggest in almost ten

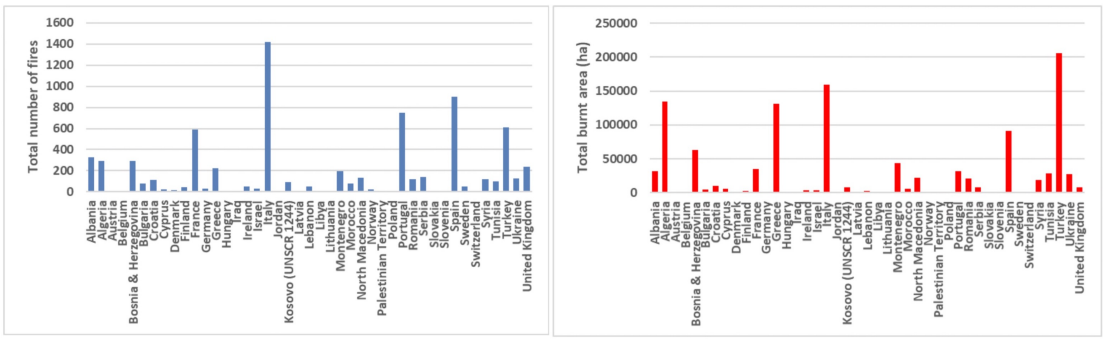

Figure 1.1: European 2021 fires distribution. The plot on the left (blue) shows the distribution of the total number of fires while the one on the right(red) shows the total burnt area(source EFFIS "*Advance Report on Forest Fires in Europe, Middle East and North Africa in 2021*" [1])

Region	**Burnt Area(ha)**	**N2000 Area(ha)**	**Fires**	**Fires(%)**
Sicily	327.042,18	43.519,35	2161	40.79
Calabria	128.944,58	24.661,55	1386	26,16
Sardinia	120.221,39	98,92	246	4,64
Campania	42.006,19	15.951,38	532	10,04
Lazio	32.281,92	6.949,36	428	8,08
Apulia	18.388,40	8.230,34	119	2,25
Piedmont	12.104,71	1.660,11	41	0,77
Basilicata	11.518,13	841,80	131	2,47
Abruzzo	9.453,28	3.863,28	59	1,11
Liguria	6.517,80	339,57	46	0,87
Tuscany	6140,47	903,28	60	1,13
Lombardy	3436,23	365,84	28	0,53
Umbria	1.537,57	219,83	19	0,36
Molise	1.446,64	0	20	0,38
Friuli V. G.	1.118,77	0	7	0,13
Veneto	749,48	1,34	3	0,06
Emilia-Romagna	604,19	50,15	5	0,09
Marche	362,58	5,51	4	0,08
P.A. Trento	50,22	0	3	0,06
Aosta Valley	0	0	0	0
Italy	**723.924,73**	**107.670,61**	**5.298**	**100**

Table 1.1: Italy fires statistics processed by Legambiente from 2008 to 2021. The total burnt area per region is comprehensive of the Natura2000 areas, which are highlighted in the second column. The first fires column describes the absolute number of fires, while the second provides the percentage per region over the total Italian fires. These values are underestimated as all fires under 30 hectares are missing (source [2]).

years with 159.537 acres and in the summer months of July and August, 90% of the damage was done. In addition to that, Italy was afflicted by 49 fires larger than 500 hectares with the largest (in Sardinia) was over 13.000 hectares and 32 of these 49 big fires occurred in Sicily. The Natura2000 zones, which are protected locations for rare and threatened species as well as other rare natural habitat types, suffered 16% of all fires in 2021.

The data shown in Table 1 are a result of Legambiente's processing of EFFIS data between 2008 and 2021. They show that a total of over 723.924 hectares of land were burned as a consequence of 5298 fires, which is almost as much as the entire region of Umbria. Around the 45% of the vegetative surface that was burned throughout these fourteen years is accounted for by Sicily alone. The first three

regions in Table 1, Sicily Calabria and Sardinia result in slightly under 80% of the total burned area, whereas Campania, Lazio, and Apulia added to these three areas result in over 90% of the total surfaces crossed by fire. Therefore, only six regions' territories account for more than 90% of all Italian surfaces that are burned in these years.

The causes of wildfires in Italy, thoroughly described in the most recent Legambiente report "*Italia in Fumo*" [2], are mainly human-related. Fires, for example, might result from outdated or improper agronomic methods, such as burning pruning scraps and stubble. These fires frequently happen each year and have terrible effects on soil biodiversity and organic matter.

Since weather patterns and land cover characteristics are not uniformly distributed throughout the Italian peninsula, the majority of academic studies examine regional assessments: impact of 2021 large fires was studied using remote sensing techniques in [3] and [4]; fire trend monitoring using spectral indexes was explored in [5] for Campania; in [6] fire perimeters available by the European Forest Fire Information System (EFFIS) were used to map burnt area in Sicily using different types of remote sensing data. It is clear that further research must be conducted in order to take a step toward an automated monitoring system.

This thesis explores the possibility of estimating weekly wildfire hazard in Sicily using a unique dataset and supervised machine learning techniques. The fire environment triangle, which consists of weather, topography, and fuel, is the foundation around which the dataset is constructed as these three factors have complete control over a wildfire's origin and spread. While data about the weather and topography are easy to find, it is different for fuels. Sampling fuel data is a costly procedure that needs to be performed in-situ and it is rarely carried out. Because of this, fuel data are frequently incomplete or outdated despite being essential to make a meaningful monitoring. This work makes use of spectral indexes computed from multispectral open-source images to describe different characteristics of fuels, from moisture content to chlorophyll production, with an update period equal to the availability of Sentinel-2 images (around 2/3 images for month considering a cloud coverage of 20%).

The analysis focuses on the southern part of Sicily which is divided into small 200x200 metre zones that are labelled as burned or unburnt thanks to fire data provided directly by Regione Sicilia. The dataset is made up of four different types of information that together depict the weekly state of each area:

- Temporal variables to report seasonality vegetation tendencies;
- Static variables to describe topographical characteristics;
- Historical variables to provide information of past fuel status;

- Forecasting variables to give info about potential future weather trends.

In order to make the model as general as possible, spatial information have been intentionally omitted. This allows future work improvements by leaving open the possibility of expanding the study area, initially, including entire Sicily and later scaling up with other regions.

The filling of the training and the test set follows the same pattern that will try to preserve seasonality and fire regime information. The training set will have wildfires 2016 to 2020 events, while 2021 will be the test. This study examined ML approaches for wildfire risk mapping, including random forest (RF), support vector machine (SVM), k-nearest neighbor (KNN), and gradient boosting classifier (GB), combined with different approaches to deal with the natural class imbalance of burnt and unburnt areas.

This work may have real-world practical applications as it demonstrates that it is possible to develop a framework that can handle raw data of different formats and convert it into tabular data. The dataset may carry way more information more than using a single source, and it is really simple to expand. For example when a new multi-spectral image is produced, it can be automatically processed and added to the dataset. In fact, with some additional computing and storage power, an analogous procedure may be effectively scaled up and automated.

1.1 Thesis outline

The thesis will develop along five different chapters along with the introduction. Chapter 2 will provide a general overview of the background context with the purpose of introducing the domain knowledge required to understand this work. Chapter 3 will describe the area of interest and its fire regime, the data of topography, weather and fuel, with a detailed description of each spectral index used, and the feature engineering process developed to build the dataset. Chapter 4 presents all the methods behind the machine learning sub-framework used with special attention to methods that deals with data imbalance. Chapter 5 presents all the experiments and the results obtained. The final Chapter 6 focuses on conclusions, limitations, and potential future studies.

Chapter 2

Background work

The sections below provide a general overview of the background context with the purpose of acquiring enough domain knowledge to create the dataset.

The physical and chemical explanation of the event is provided in section 2.1, along with a brief historical introduction and some terminology definitions. In order to carry out the feature extraction to build the dataset, it is crucial to have a strong understanding of which variables are the most crucial and how they have been used in literature. In order to introduce multispectral remote sensing and the spectral indexes, which will be the features utilised to define fuel attributes, some spectroscopy concepts are reviewed in section 2.2. Finally, a brief summary of the work that has been done on Earth Observation wildfire monitoring is provided in the final section 2.3.

2.1 Wildfire modelling

Wildfire is a complex phenomena that combines chemical kinetics and heat transfer. It has been faced with a wide range of approaches from purely physics to statistics that has been investigated since the 1920s. This section will help the reader to better understand the basis that lay the foundations for what has been done in this thesis. Along with the historical overview, a brief description of the physics and the chemistry behind fires will be provided. Finally, some clarifications will be done on fire terminology

2.1.1 An historical overview

The idea that understanding of a fire could be gained through theoretical analysis of the factors that could influence fires, started to raise in the 1920s, but it was not until the 1940s that the first physical model came out.

The pioneer of fire modeling was a mechanical engineer named **Wallace Fons** [7] who worked for the United States Forest Service. He modelled fuel as a discrete set of particles, where each particle could heat neighboring fuel particles up to ignition temperature. With this idea, he noticed that the rate of fire spread is controlled by three main characteristics: how long it takes fire to ignite, by the type of fuel and by how far apart the fuel particles are. This was a simplified model that was validated through laboratory experimentation and it had quit good results, but most importantly, it laid the basis for the modelling research.

It is reasonable to state that the birth of fire modelling coincided with the end of WWII, during which fire was widely used as a weapon and it was necessary to find a way to suppressing it. Actually, after the war, authorities were convinced that the next war would also be a fire war so considerable effort was expended exploring the effects of mass bombing (such as occurred in Dresden or Hamburg, Germany) and the collateral incendiary effects of nuclear weapons [7]. The United States Forest Service became actively involved in nuclear blast tests, employing the country's best fire scientists. This war-inspired research helped to discover fundamental knowledge about fire.

So, the late 1960s saw an explosion of research publications and several countries released new fire models, such as Australia, Russia, and Canada. Like Fons's model, many of the newer models were physical and based on the laws of combustion, and heat transfer. However all these model were not self-determining and they were not free from empirical components. Many specific properties of fuels or gases should be provided ahead to these models to work, properties that can only be measured through experimentation. For this reason, a new wave of fire models started to raise, made possible by the experiments and data-gathering of the previous decades: empirical and semi-empirical models. [8]

One of the most important semi-empirical fire models was published in 1972 by **Dick Rothermel**. His equations were suitable for so many wildfires that the Forest Service implemented them in the first release of the National Fire Danger Rating System (NFDRS), which initially simply consisted of lookup tables. Firefighters manually filled in the wind and slope angle to determine the speed and direction of the spread of the fire using paper and pencil. The NFDRS is computerised today, but it continues to be based on Rothermel's revolutionary equations. In reality, the Rothermel model serves as the foundation for every fire model now being utilised in the field.

Another theoretical approach born thanks to the increased computational power in 1996 by **Terry Clark** [9]: fire spread models *coupled with numerical atmospheric models*. The coupling could be done with computational fluid dynamics (CFD) which simulate turbulent airflow at a very high resolution or with 3D numerical weather prediction models. In both cases, this model allows fire to interact with the atmosphere as it does in the real world.

In general wildfire models are composed by a certain number of equations whose solution usually gives values for rate of spread , flame height of fuel consumption. Following this definition, as in [10], wildland fire mathematical models may be classified:

- *According to nature of equations*:
  - *Theoretical models.* They are based on the physics and chemistry of combustion. They can be further divided into models that attempts to represent only the physics and models that attempts to represent both the physics and chemistry of fire spread[11];
  - *Semi-empirical models* are often based on simple physical idea like the conservation of energy principles without making difference between different heat transfer mechanisms nor consider the combustion chemical processes[8];
  - *Empirical models* are purely statistical models which are based upon observation and experiment and not on theory[8];
  - *Mathematical models.* They use a mathematical approach by implementing mathematical ideas that appear similar to the spread of fires [12];
  - *Simulation and GIS models.* They implement other types of models in a simulation rather than modelling context and their primary function is to convert one dimensional models to two dimensions and then propagate a fire perimeter across a modelled landscape [12].

Author	**Year**	**Country**	**Type**	**Fire Model**
W. Fons [13]	1946	USA	Theoretical	Surface
McArthur [14]	1966	Australia	Empirical	Surface
Van Wagner [15]	1977	Canada	Semi-Empirical	Crown
Rothermel [16]	1972	USA	Semi-Empirical	Surface
Rothermel [17]	1991	USA	Semi-Empirical	Crown
Grishin [18]	1997	Russia	Theoretical	Surface
Linn [19]	1997	USA	Theoretical	Surface
Tarifa [20]	1965	USA	Semi-Empirical	Spotting
Albini [21]	1979	USA	Theoretical	Spotting
Frandsen [22]	1987	Canada	Theoretical	Ground

Table 2.1: Most important fire models. Current fire monitoring systems incorporate powerful calculation tools with some of the most important research models.

Figure 2.1: Types of wildland fires. Ground fires occur in deep accumulations of dead vegetation; surface fires burn only surface litter and duff; crown fires burn trees up their entire. Each type of fire should be modelled in a different way as they all have very different characteristics.

- *According to physical system modelled*:
  - *Surface fire models.* Bushes, tiny trees, and anything with a height of less than 2 m form the physical system for these fire, which are the simplest to extinguish and do the least harm to the fores.[23];
  - *Crown fire models.* The strata of surface and aerial plants that make up the physical system are higher vegetation. These are the wildfires that are most intense and hazardous.[23];
  - *Spotting models.* The physical system is made up of firebrands or flaming material that is moved away from main fire perimeter by the convection column. It is a phenomenon that is mostly related to huge wildland fires, and it can cause extremely dangerous scenarios. For example, firefighters might become caught between two fire fronts, or in an urban wildland, it could be the cause of building fires.;
  - *Ground fire models.* Humus, peat, and other similar types of dead vegetation that have dried up sufficiently to burn compose the physical system. Although they burn very slowly, these fires can become challenging to completely extinguish or suppress. Sometimes, especially amid a protracted drought, these fires might smoulder underground all winter until resurfacing in the spring [23].

2.1.2 Physics and chemistry of combustion

Fire is the complicated combination of energy released due to chemical reactions and the transport of that energy to surrounding unburnt fuel. The energy is released as heat and it allow the fuel to reach the ignition temperature and burn.

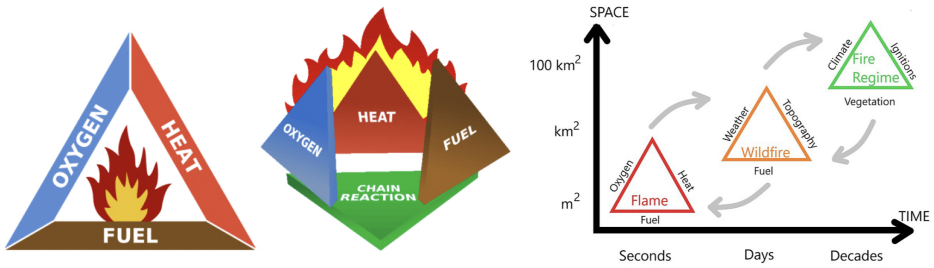

Figure 2.2: Fire descriptive figures. Left: Fire triangle for non-flaming combustion; Centre: fire tetrahedron for a sustained fire; Right: Multi-scale fire triangles

This idea can be summed up by the so-called *fire triangle* (Figure 2.2). It is a simple model to understand the necessary ingredients for a non-flaming combustion comprised of three elements:

- *Oxygen* to sustain combustion;
- *Heat* to let the substance reach the ignition temperature;
- *Fuel* to have something burnable.

However, combustion alone is not enough to cause a fire. Actually, once a fire has started, it needs to be sustained by exothermic *chain reactions* that allows it to continue burning. So a sustained fire is described by a *fire tetrahedron* (Figure 2.2): a fire is stopped when one of the four elements is removed.

In the context of wildland fire, the fire triangle can be scaled up to apply to fire spread over landscapes, as can be seen in the right figure (Figure 2.2). A wildland fire is the resultant of the environment in which the fire is burning and the fire triangle can be scaled up to apply to fire spread over landscapes and recurrence of fire over time. A wildland fire is controlled by three elements, the so-called *fire environment triangle*: fuel, topography and weather.

Topography describes the shape of the land surface and it is a combination of elevation, slope (steepness of the land) and features like canyons, valleys, rivers, etc. Elevation controls fuel distribution and condition as it may lead to different rainfall patterns, temperature, relative humidity and so on. Steepness directly influences the rate of spread, indeed fires burning upslope preheat and dry potential fuel, which can lead to faster spread rates. Finally, features of topography also influence the wind flow behavior: for example, complex terrain can result in gap flow scenarios, where the wind accelerates in gaps between mountain tops.

Weather play an important role to determine the conditions of fuels and how a fire can spread. Indeed, temperature, precipitation, and relative humidity affect the moisture content of fuels: low relative humidity and high temperatures result in extremely dry fuels that are susceptible to ignition. At the time of ignition, wind can be the dominant factor in fire spread as it increases contact between flames and fuels and facilitate transport of heat through convection, creating the conditions for a quicker spread in the wind direction.

Fire behaviour is also strongly influenced by both vegetative and structural *fuel* characteristics. Its present is determined by land use and the material and its moisture content determine the ignition potential.

Chemistry of combustion

The chemistry of combustion involved in wildland fire is complex for two main reasons: the great amount of different fuels and the range of conditions over which combustion can occur [11]. As different types of solids heat up, they will behave in different ways: some solids first change into liquid before they form fuel vapour and burn; other vapour directly upon heating. In general, solid organic materials do not burn in flaming combustion directly, but must first be decomposed by heat and chemical reactions into various combustible. This process of decomposition caused by heat is called *pyrolysis*.

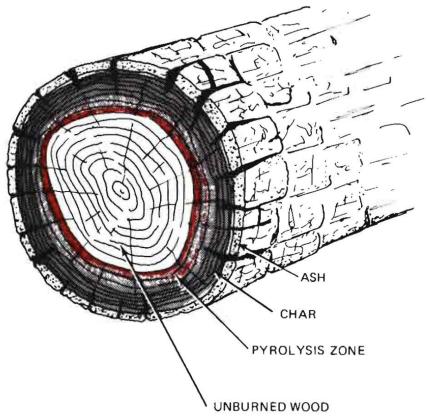

Figure 2.3: Wood pyrolysis. When a piece of wood is ignited, the pyrolysis, but as the combustion continues, a char layer forms on the surface and deepens as the pyrolysis penetrates into the wood. The char burns slower rate than that of the wood following inward to to form grey ash. Unlike wood, the char burns directly without being pyrolyzed into gases. These steps can be seen if we examine a section of burning wood [24].

Wildland fuel is generally composed of *live and dead plant material* consisting primarily of leaf litter, twigs, bark, grasses, and shrubs. Typical forest fuel primarily consists of cellulose, hemicellulose, and lignin in varying proportions, as well as extractive and mineral substances.

So, the overall process of pyrolysis of fuels is believed to proceed as follows:

1. *Dehydration*: the temperature is raised from ambient temperature until the fuel looses its moisture;

2. *Fuel thermal degradation*: the temperature continue to raise and the plant material starts its decomposition process. Hemicellulose is the first to decompose at 200–350°C [25], yielding predominantly volatile products such as carbon dioxide, carbon monoxide and condensable vapours. Then cellulose undergoes its degradation producing reactive gases that react with the oxygen of the air with a temperature above 300°C [25]. With a temperature of 200–450°C[25], also the decomposition of lignin starts. All gaseous decomposition products of these substance are released into the atmosphere and subsequently cause fuel ignition as well as support the combustion.

There are two types of thermal degradation reaction: *volatilisation* and *char formation*. The first is an endothermic reaction whose major product is the levoglucosan, which is often used as a chemical tracer for biomass burning in atmospheric chemistry studies. Instead the production of char, which is a generic term for carbonized solid fuels, is an exothermic reaction and it is the char that reacts directly with the oxygen on the surface during non-flaming combustion.

Several kinetic models have been proposed for the description of the endothermic and exothermic reactions that occur during pyrolysis, they are typically divided into two groups: models based on the global decomposition of wood and models based on the breakdown of wood's constituents, namely hemicellulose, cellulose, and lignin.

Physics of combustion

The physics involved in the combustion of wildland fuel and the behaviour of wildland fires is complicated and highly related on the condition in which a fire is burning. The primary physical process in a wildland fire is that of heat transfer. There are three modes of *heat transfer*:

- *Conduction*is a process by which energy caused by molecular agitation, heat is spontaneously transmitted from a hotter to a cooler body.;

- *Convection* is the transfer of heat by the movement of a gas or a liquid, even if in wildfires liquid phases are really rare. However, the movement of a gas

can be modelled using fluid dynamics laws for a continuous medium as the molecules or particles of a gas are assumed to be continuous and act as a fluid rather than a collection of particles;

- *Radiation* is electromagnetic radiation emitted from a hot source generated by the thermal motion of particles in matter.

All these three methods are usually operating at the same time: radiation and convection can transfer heat to the fuel surface and conduction can transfer heat into the interior fuels. In low wind conditions, the dominating process is that of radiation, but in conditions where wind is not insignificant, it is convection that dominates. However, it is not reasonable to assume one works without the other and thus both mechanisms must be considered.

Heat transfer by conduction is a slower process and has minor direct consequences in the rate of spread of a wildland fire. Indeed, conduction carries heat through fuels and can raise the temperature of fuels to the point that they ignite. Conduction can preheat and dry larger fuels that are touching each other and may increase the duration those fuels burn by promoting the internal transfer of heat if flammable vegetation is abundant and continuous.

As seen in the chemistry section, gases are heated during the pyrolysis which cause a reduction in density and a increase in the *buoyancy*: it results in the gas rising. This moving upward can cause two main drawbacks:

- ignite the leaves and branches of trees and plants above the fire;
- turbulence can be caused in the flow.

The first case is a form of solid phase transport also called advection. The floating embers, sometimes known as firebrands, might fall in regions that have not yet burned and spark smaller flames. This behaviour, known as *spotting*, can cause the fire to spread quickly.

Instead, *turbulence* acts to mix the heated gases with unburnt solid phase fuels or ambient air, to increase flame immersion of fuel and it could have an effect on the movement of firebrands and other solid phase combustion products, causing spot fires to spread downwind of the main burning front.

Solid phase transport can happen in two different cases. The first is through *advection*, so transport of solid materials by convective fluxes, as indeed the vertical convective currents can also lift burning materials.

Even if convection, solid phase transport and conduction are important, radiation is generally considered as the dominant heat transfer mechanism. The source of the radiation emits energy in all directions until it comes into contact with something that absorbs it. The substance's molecular activity is increased by the absorbed radiation, raising its temperature and the amount of heat it contains. Thermal

emission from burning fuel surfaces and flames is the main radiation source in a flame.

2.1.3 Fire management terminology

Commonly the terms *danger*, *hazard* and *risk* are used as synonyms, probably because in many languages they are translated with one word, but they are technically different. Basing on accurate definitions provided in [26] using FAO (i.e Food ad Agriculture Organization of the United Nation) glossary, the three terms could be described as follows:

- *Fire hazard*: it is the fire component regarding the fuels available for burning, so a ratio between the amount of fuel available for combustion over all fuels available in that area. This ratio is strictly correlated to fuel moisture content as fuels with high moisture content are difficult to ignite;
- *Fire danger*: it expresses the difficulty of controlling an active fire. It is both related to human factors, as burning area accessibility, and to weather and topography conditions, as strong wind or steep slopes. It considers more factors than the fire hazard, which relay only on the amount of burning fuels;
- *Fire risk*: it is the probability of a fire to spread in a specific situation and the damage it might produce. It can be expressed with a simple formula:

$$Risk = P_{ign} \cdot P_{prec} \cdot V \tag{2.1}$$

  where P_{ign} is the probability of ignition, while P_{prec} is the probability that the fuel allows the ignition and V is a measure for the expected loss due to the fire. The ignition can be caused by a natural event, such as lightning, or by an unintentional or intentional human action. Indeed, the evaluation of human component requires an understanding of the way in which human activities are related to fire occurrence.

A comprehensive evaluation of fire damage would require an analysis that is out of the scope of this thesis, therefore from now on only the terms fire danger and fire hazard will be used.

2.1.4 Fire danger rating systems

While dozens of variables have been pointed to as important drivers, the majority fall into the three categories of climate, fuel, and socioeconomic. As already stated, climate change is expected to have a strong impact on forest fire risk because of

issues like warming temperatures, increased vapor pressure deficit, and drought severity. Key problems related to fuels include abnormally dry fuel conditions, bark beetle outbreaks, and fuel accumulation due to a legacy of fire suppression. In the socioeconomic category, features that are routinely pointed to as contributors to the wildfire crisis are the rise in the recreational use of forests, population density, and the number of people living in the wildland-urban interface.

An effective way to summarize all the steps needed to manage these extreme events with efficient response is the *disaster management cycle* [27]. It illustrates how to reduce losses, to react during and immediately following a fire and to achieve a rapid and effective recovery, with the following steps:

1. *Prevention and Mitigation*: preventive activities carried out to reduce the probability of a disaster occurrence and, consequently, to reduce its damage;

2. *Preparedness*: plans, initiatives, rules to develop response capabilities;

3. *Response*: efforts to minimize the hazards created by the fire;

4. *Rehabilitation and Recovery*: strategies to support short and long term recovery.

Each phase of this cycle involves a different group of stakeholders, each bearing its own set of responsibilities, interests, and needs. For example, fuel conditions are significant to efficiently clear-cut forest to limit fire spread in the mitigation phase, to deploy resources in advance in the preparedness phase and to decide attack strategies in the response phase[26].

Figure 2.4: The Disaster Management Cycle. This scheme breaks down the different aspects of disaster management, from prevention to preparedness, from response to recovery. It is essential to provide efficient and punctual hazards handling.

Being able to classify fire danger is the key idea to build strategies for each stage of the disaster management cycle (Figure 2.4). Indeed, several fire danger rating systems exist worldwide to support the decision-making process. The three most important operational models are:

- *McArthur Forest Fire Danger Index (FFDI)*: developed by a CSIRO (Commonwealth Scientific and Industrial Research Organisation) scientist called McArthur, provides a measure of forest fire danger that combines a measure of vegetation dryness with air temperature, wind speed and humidity [28]. It is used in Australia and it is based upon the purely empirical model built by McArthur [14];

- *National Fire Danger Rating System (NFDRS)*: used in the United States, it is a collection of fuel condition and fire behaviour indices computed from weather station measurements. It is based upon the semi-empirical model of Rothermel [16].

- *Fire Weather Index (FWI)*: is a collection of different components that account for the effects of fuel moisture and wind on fire behaviour and spread. This system was born in Canada, but it has been proved to be robust and it is used in Europe by EFFIS [29] (European Forest Fire Information System).

At the core of all three fire danger rating systems above mentioned there is the fuel moisture content (FMC). It reflects the ratio of the water contained in the sample to its dry mass and is determined by the formula

$$FMC = \frac{w_t - w_d}{w_d}$$

where w_t is the biomass before drying and w_d is the dried biomass. The moisture content is central as it has a major influence on the properties of the fuel: it will be more difficult to ignite forest fuel with high moisture content, since a large amount of heat is required for moisture evaporation.

Nowadays, remote sensing technology is widely used to determine FMC and other useful parameters to provide a fire danger mapping.

2.2 Remote sensing and Earth observation

Remote sensing is the discipline based on the measurement of some property of an object by an acquisition platform that is not in contact with the object. So the goal is to obtain the most accurate measurement using the most appropriate sensor on the post practical platform.

2.2.1 Historical overview

The first remote sensing technique is photography with camera being the first sensor. It dates back to 1858, when the first aerial photo was taken from a baloon by the French photographer Gaspard-Félix Tournachon, known as "Nadar" over Paris. However, hisphotographs no longer exist and therefore the earliest surviving aerial photograph is from 1860 and it is a picture of Boston taken from 630m (Figure 2.5).

Aerial photography became a recognized as a valuable tool during the First World War. Aerial views of enormous surface regions were made possible by cameras that were mounted aboard aircraft, and these views were crucial for military reconnaissance. Aerial cameras had proven to be a really useful tool to monitor enemy positions, movements and defenses so governments funding were made to further improve this promising technology.

During World War II, the main idea was to expand the acquirable spectrum of the camera. Instead of acquiring only color images, they started to capture other types of images at different wavelengths of non-visible portions of electromagnetic spectrum, thanks to the improvements of radar (radio detection and ranging), thermal infra-red detection, and sonar (sound navigation ranging) systems. Thanks to these technologies, multispectral remote sensing is born.

Figure 2.5: First aerial photo. "Boston, as the Eagle and the Wild Goose See It" taken by James Wallace Black

The development of satellites during the Cold War allowed remote sensing to progress to a global scale. The Soviet Union launched the first artificial satellite, Sputnik 1, into orbit in 1957, with cameras. The Cold War decade brought about rapid developments in satellites and imaging technology. In 1972 Landsat 1, the first satellite designed specifically to study and to monitor the Earth's surface, was launched by the US. The original goal of the Landsat program was to collect data from the Earth through remote sensing techniques. It captured over 300.000 multispectral images, thanks to its Multispectral Scanner (MSS), before its termination in January 1978.

There are a great number satellites in operation today, many of which are used for remote sensing applications: there are currently over 3,600 satellite orbiting the Earth, but only approximately 1400 are operational. Among these satellites, earth observation satellites consist in over 100 and each of them is equipped with a range of sensors that can measure and record information about the Earth. Governments frequently launch these satellites to keep an eye on Earth's resources, but private business organisations are also taking a greater interest in launching earth observation satellites. [30]

Prior to 2008, the costs for a Landsat MSS image varied from 20 dollars (1972–1978) to 200 dollars (1979–1982), but then a free and open data policy was adopted. After that all major governative space agencies applied a free data policy together with ESA, the European Space Agency.

Another important earth observation program is the Copernicus Programme. It builds up on three different components:

- *Space component* with observation satellites and associated ground segment with the mission of observing land, atmospheric and oceanographic parameters). This consists of two categories of satellite missions: Contributing Missions from other space agencies and the five Sentinel (space mission) families of the European Space Agency;

- *In-situ measurements* with ground-based and airborne data-gathering networks providing information on oceans, continental surface and atmosphere);

- *Services* developed by Copernicus and are mostly open-source, that cover six main interacting themes: atmosphere, marine, land, climate, emergency and security [31].

So satellites are specifically designed to monitor and measure specific information about land, ocean, and weather, the table 2.2.1 show some active and inactive governative satellites from different space agencies with a brief description of their use.

Mission	Agency	Activity period	Description
Landsat (1-9)	NASA	1973-	MSI high resolution images of land surfaces and coastal areas
Envisat	ESA	2002-2012	Different sensors to monitor oceans and natural hazards and to obtain a digital elevation model
ACRIMSAT	NASA	1999-2013	Sun's UV to infrared energy output
Jason (1-2)	NASA and CNES	2011-2013	Radar altimeters used to monitor ocean surface height
Sentinel-1(A,B)	ESA	2014-	Copernicus Program's C-SAR sensors
Sentinel-2(A,B,C)	ESA	2015-	Copernicus Program's MSI sensors
Aqua	NASA	2002-	Interactions among oceans, land, atmosphere, and biosphere
COSMO-SkyMed 1 to 4	ASI	2007 -	Seismic hazard analysis, environmental disaster monitoring, defence and security
PRISMA	ASI	2019-	Development and delivery of hyperspectral products
GOSAT	JAXA	2009 -	Greenhouse Gases Observing Satellite

Table 2.2: Some governative earth observation satellites. [32]

2.2.2 Spectroscopy

Spectroscopy is the study of matter and its properties via an analysis of the radiant energy that is absorbed, emitted, or scattered by the target object. Although the study of the interaction between visible light and materials was the original purpose of spectroscopy, it has since been expanded to include the entire electromagnetic spectrum, from short-wavelength X rays to long-wavelength microwaves.

The electromagnetic spectrum, shown in Figure 2.6, is important as the first requirement for remote sensing is to have an energy source to illuminate the target and this energy is indeed in the form of *electromagnetic radiation*. There are several regions of the electromagnetic spectrum which are useful for remote sensing:

- Ultraviolet radiation helps rocks and minerals detection as it fluoresces them and they emit visible light when irradiated;
- Visible radiation provides the visible usual colors. It helps discriminating between objects, for example green reveals green vegetation and trees while bluish-green color describes lakes;
- Infrared radiation can be thermal and it discloses information about the temperature of the Earth's surface;
- Microwave radiation can penetrate cloud cover, haze, dust, and even heavy rainfalls to provide information even in bad weather conditions.

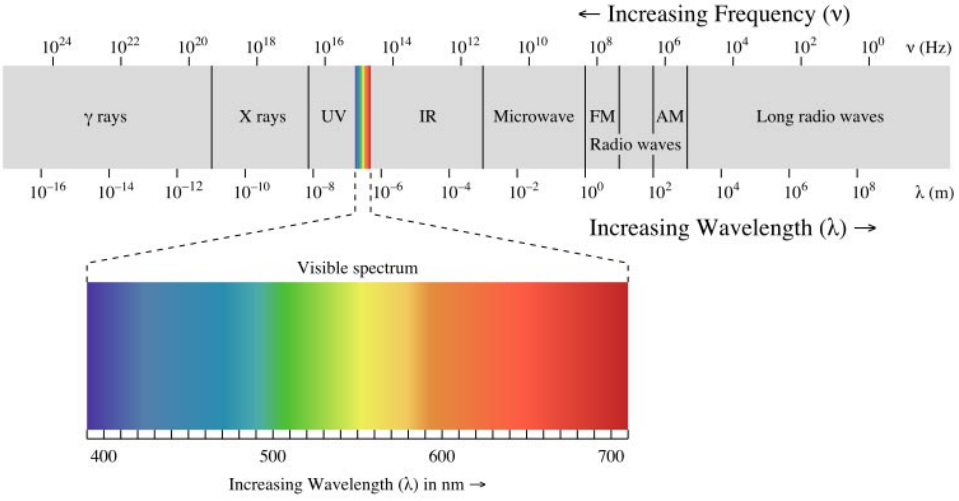

Figure 2.6: Electromagnetic Spectrum.

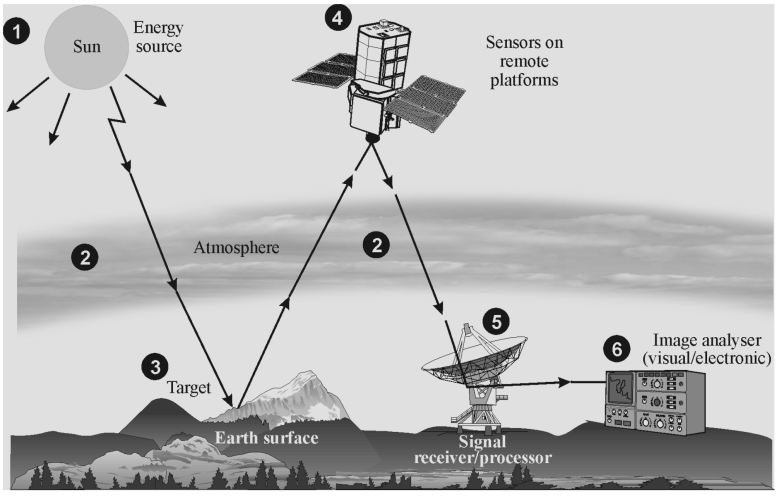

Figure 2.7: Remote sensing acquisition process. (1) The *sun* provides the radiant energy that falls on the Earth's surface; (2) the *atmosphere* is crossed from the source to the target and back again to the sensor platform; (3)the *Earth's surface* illuminated by the radiation reflects and/or reemits the incident energy; (4) a *platform with sensors* receiving the energy reflected or emitted by the Earth's surface; (5) *ground-based receiver* that processes the information sent to it from the sensors on the observation platform; (6) A *final station* dedicated to interpreting the information processed by the ground-based receiver and presenting it in visual, digital, or electronic form.

Through the remote sensing acquisition process decribed in Figure 2.7, incoming light and radiation may be affected by atmospheric particles and gases, and energy passing through the atmosphere will experience absorption and scattering:

- *Scattering* it occurs when a radiation path is deflected from its intended course and it spreads the energy of the incident in all directions. So, the wavelength of the radiation, the quantity of particles or gases, and the distance the radiation travels through the atmosphere all affect how much scattering occurs;

- *Absorption*: it occurs when radiation energy is transformed into the excitation energy of the molecules. The three main components of the atmosphere that absorb radiation are ozone, carbon dioxide, and water vapour. While water vapour absorbs a large part of incoming longwave infrared and shortwave microwave radiation, carbon dioxide absorbs radiation in the far infrared region of the spectrum, trapping heat inside the atmosphere. Ozone also absorbs ultraviolet radiation.

2.2.3 Types of remote sensing

So, in general, it is possible to create a *spectral response* for a target on the Earth's surface by analysing the energy that is reflected (or emitted) by that target over a range of various wavelengths using different remote sensing technique. They can be divided into two main categories based on the signal source that is used to examine the item: passive remote sensing devices rely on reflected light to examine the itme, while active ones need their own source of emission of light to work.

Active sensors sends their signal in the direction of the object and then check the response. They can operate at any time of day as they don't need sunlight. The difference between *active remote sensing* techniques lies in what they send (light or waves) and what they measure (e.g., distance, height, atmospheric conditions, etc.), here are some examples:

- *Radar sensors* uses radio frequencies through an antenna that emits impulses, but energy flow encounters an obstruction and, to some extent, scatters back to the sensor. It is possible to calculate how far away the target is based on the amount and distance travelled;
- *Lidar sensors* use light to measure distance by sending light pulses and measuring the amount received;
- *Laser altimeters* are used to measure elevation.

Unlike active ones, *passive remote sensing* sensors do not direct their own energy toward the surface, but sunlight reflected by the target is the source of natural energy used. Because of this, it can only be used when there is adequate sunshine, as there won't be anything to reflect. Various band combinations are used to measure the obtained quantity through multispectral or hyperspectral sensors. The number of channels used in these combinations varies, but generally the range of bands include both spectra visible and invisible (visible, IR, NIR, TIR, microwave). The most often used passive remote sensing equipment includes several spectrometers and radiometers, for example:

- *Spectrometer* can identify and examine spectral bands;
- *Radiometer* measures the power of radiation emitted by an object in specific band ranges;
- *Hyperspectral radiometer* is able to distinguish among hundreds of spectral bands thanks to its exceptionally high resolution;
- *Imaging radiometer* creates a surface picture by scanning it.

Regardless the type of remote sensing technique used, sensors data quality is evaluated considering four types of resolution:

- *Radiometric resolution* represents the amount of information given by each pixel, so the number of bits that indicates the energy captured.Because there are more values available to store information, there is better ability to distinguish between even tiny variations in energy with increased radiometric resolution;

- *Spatial resolution* is determined by the size of each pixel recorded in a raster image and the area of the Earth's surface that each pixel represents. More detail can be seen when the resolution is finer (lower number);

- *Spectral resolution* is the capacity of a sensor to distinguish finer wavelengths, or having more and smaller bands. Sensors categorised as multispectral typically have from 3 to 10 bands, while hyperspectral sensors have hundreds or thousands of bands. The spectral resolution is finer when the wavelength range for a given band is smaller;

- *Temporal resolution* is The amount of time it takes a satellite to complete an orbit and return to the same observation region. This resolution is influenced by the orbit, the features of the sensor, and the swath size. The temporal resolution is substantially higher for geostationary satellites since they rotate at the same speed as the planet. The temporal resolution of polar orbiting satellites can range from one day to sixteen days.

It is challenging to incorporate all the ideal qualities into a single remote sensor, so trade-offs are necessary. This is why it is crucial to understand the kind of data required for a particular field of study.

Spectral indexes

Spectral bands carry different types of information as every object has its own chemical composition and each composition has its own spectral signature. As shown in Figure 2.8, different Earth features have different spectral signatures: turbid water is different from clear water, while dry soil is different from wet soil. The intuition is that spectral bands can be used in many application, starting from water monitoring, to land cover classification.

To obtain even more information, these bands can be combined into spectral indexes. They enhance the contribution of some particular chemicals or features to provide a more detailed of the object of interest. A wide range of features, such as vegetation photosynthetic activity or vegetation moisture content can be extracted using these indices, which can be computed without any bias or assumption as they are essentially a straightforward modification of spectral bands.

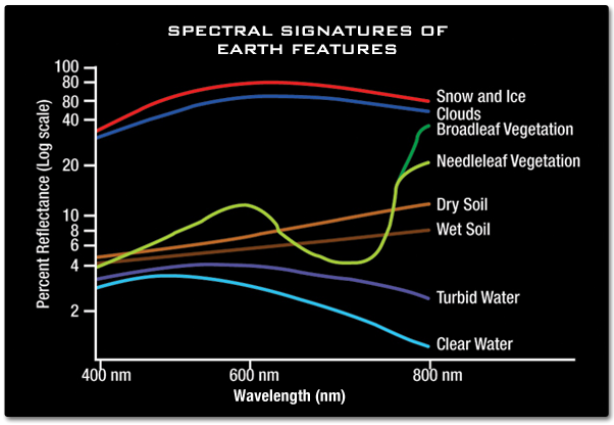

Figure 2.8: Spectral signatures of different Earth features within the visible light spectrum. (Source [33], Credit: Jeannie Allen.)

Typical algebraic formula are normalized differences and ratio:

$$Index = \frac{B_x}{B_y}$$

$$Index = \frac{B_x - B_y}{B_x + B_y}$$

where B_x and B_y are two different spectral bands.

These kind of formulas help to enhance spectral features and to reduce the impacts of illumination and, more significantly, shadows. There isn't a single mathematical formula from which to calculate all spectral indices because equipment, platforms, and resolutions are different between different sensors. Spectral indexes have a lot of potential, they also have some significant limitations: for example, a given surface may have different signatures depending of many different factors, such as approximations of atmospheric corrections or different latitudes, indeed what is actually a normal value for a location at one time may not have the same typical value for that area. So attention needs to be paid when building the index pool for a specific task.

2.3 Remote sensing and wildfires: state-of-the-art

Pre-fire, active fire, and post-fire are the three fire stages that can be used to categorise wildfire monitoring and management tasks, and remote sensing can contribute in each of them.

- *Pre-fire stage*: the monitoring of this stage focuses on fuel types and their conditions with a close relationship to tree species. They are influenced by factors such as topography, land use, vegetation zones, climate, and meteorological variables including wind speed;
- *Active stage*: phase in which the fire starts and spreads. Fire output and intensity can both be measured and the fuel type, topography, and weather all have an impact on this stage;
- *Post-fire stage*: after the fire has been extinguished, it is important to describe what remains of the vegetation. Three possible measurements can be made during this stage: mapping the burned area, determining the intensity of the fire, tracking the vegetation's recovery, and taking restoration measures.

Fuels' type and state are mapped during the pre-fire phase. A combination of factors, including vegetation species, form, and size arrangement, determine fuel. Moisture content and the status of the fuel—live or dead—determine its condition. The majority of remote sensing fuel-type mapping is carried out by classifying plant functional kinds and using vegetation index methodologies ([34], [35]). LiDAR structural data can supplement these analyses by potentially revealing details about the structure of vegetation canopy [36].

Remote sensing methods for the active fire phase include temperature retrieval and fire detection. For instance, MODIS and VIIRS satellites, two NASA satellites with thermal bands, are used to retrieve data regarding active fires for EFFIS Active Fire Monitoring tool [37]. The information is based on the detection of thermal differences between a possible fire and the terrain around it. The presence of a potential fire is identified as a "hot spot" if the temperature differential meets a predetermined threshold.

The assessment of fire intensity and the mapping of burned areas are the main post-fire procedures. Burned area maps for rapid damage assessments can also be obtained from MODIS data. Unsupervised fire mapping processes are visually checked and rectified using visual interpretation of the MODIS images. Because of the 250 m image resolution, tiny flames cannot be precisely mapped. In addition to thermal analysis, these objectives can addressed with specific spectral indices such as the Normalized Burnt Ration (NBR) [38]. Additionally, in order to quantify recovery outcomes, [39] integrated forest spatial structural information with spectral information obtained from synthetic aperture radar and optical remote sensing.

To sum up, the majority of studies rely on optical vision, with thermal imagery dominating for active fire detection. Because they have a higher spatial resolution than MODIS-type sensors, Landsat and Sentinel-2 data are crucial components of many systems. In addition, Sentinel-3 maybe used to supplement MODIS data to identify current forest fires [40].

Chapter 3

Dataset building

Wildfire behavior is driven by the interplay of three components: topography, weather, and fuels, the so-called fire environment triangle presented in Section 2.1.2. These components need to be quantified precisely to identify locations with elevated fire hazard potential: topography and weather data are available and easily accessible, while fuel data strongly vary in space and time, so their detailed characterization often demands resource-intensive field observation. The underlying assumption of this work is that spectral indexes computed from multispectral data can accurately represent continuous fuel properties while preserving their seasonal behaviour.

After a presentation of the area of interest and its fire regime (Section 3.1), all the data used to characterise fuels 3.2, weather 3.3 and topography 3.4 features are described. Finally, the feature engineering work done to build the final dataset is described in the last section 3.5.

3.1 Area of interest and its fire regime

Sicily is the largest island in the Mediterranean Sea with a total size of 25,711 km2. It has a typical Mediterranean climate which is typically characterized by mild and wet winters and hot dry summers. Annual rainfall is highly variable, in general precipitations are concentrated in fall and winter, while summer are characterized by drought. The average annual rainfall varies from less than 50 cm in the southeast coast to over 100 cm in the northeastern highlands. In addition to that, especially in summer, it is common to have the Saharan wind called *Sirocco*.

The majority of Sicily's interior is hilly, and when it is possible, it is heavily farmed and there is a lot of seismic and volcanic activity, indeed Mount Etna is the tallest active volcano in Europe (3,220 metres).

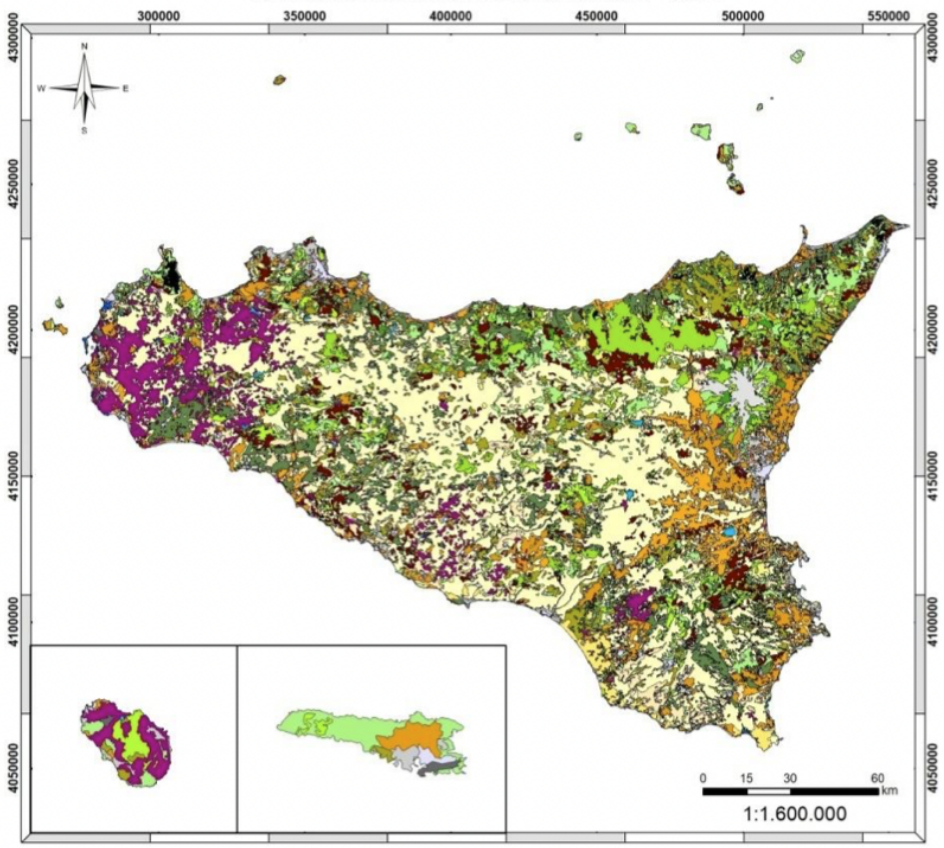

Figure 3.1: Sicily Land Cover Map. Yellow describes agricultural areas; Orange describes permanent crops, like fruit trees; Purple describes vineyards; Green describes forest and seminatural areas; Light grey describes bare rocks. (Source: Arpa Sicilia processing of Copernicus Corine Land Cover data 2018 [41]).

Thanks to Arpa's processing of Copernicus Corine Land Cover data from 2018, the Figure 3.1 gives a more comprehensive representation of Sicily's land cover. Agriculture, woods, and semi-natural regions make up the majority of the land cover. Arable fields are the most common type of agricultural land, followed by fruit trees(orange) and vineyards (purple).

Being covered mostly in vegetation, Sicily is the region most frequently crossed by wildfires, indeed from 2016 to 2021 there has been 6860 fires. Figure 3.2 shows that 2016 and 2017 were the years with the major number of fires, while 2018 were the year with the less number of fires. Typically the number of wildfires slightly increases between May and June, reaching a peak in August. In February and August 2021 there has been the two biggest wildfires during this period of time.

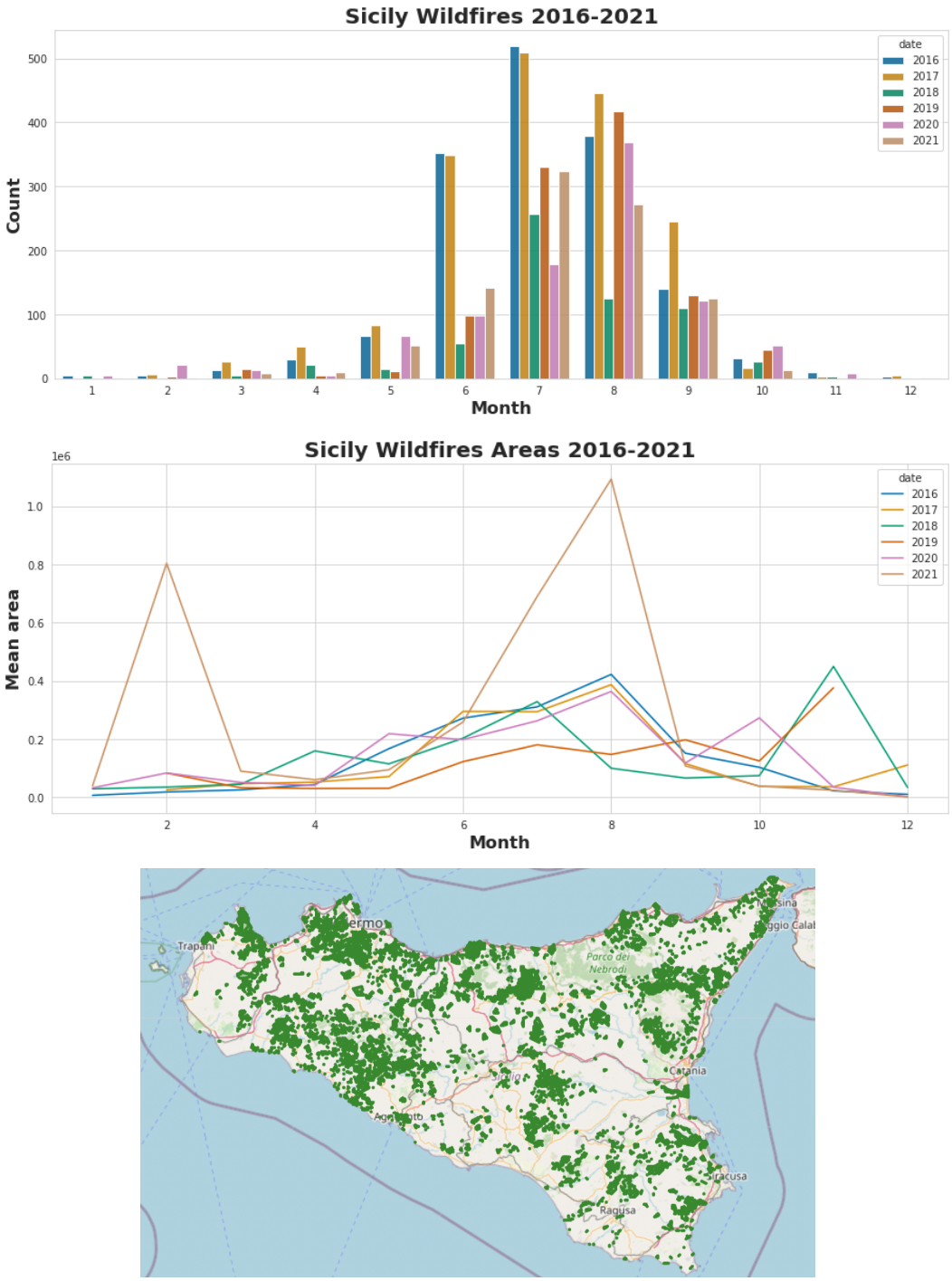

Figure 3.2: Sicily wildfires from 2016 to 2021. The first shows the yearly trend of the number of fires; the second plot shows the mean area yearly trend; in the third plot the green areas are the burnt polygons to describe fires spatial distribution (Data Source Comando del Corpo Forestale della Regione Siciliana)

3.2 Fuel data

Fuel data are essential to make a meaningful monitoring but frequently incomplete or outdated as they should be usually taken in-situ and it is a costly procedure. This work makes use of two different sources: LUCAS dataset and Sentinel-2 multispectral images. LUCAS is a European dataset with in-situ collected data that provides land cover information (Section 3.2.1); while Sentinel-2 images will be used to compute spectral indexes (Section 3.2.2).

3.2.1 LUCAS

LUCAS is a Land Use / Cover Area Frame Survey in which data are actually gathered through direct observations on the entirety of the EU's territory. The aim is to create a uniform framework for coherent sampling plans, classifications, and data collection processes to provide unbiased statistics on land use for agriculture and environmental and landscapes monitoring in the European Union.

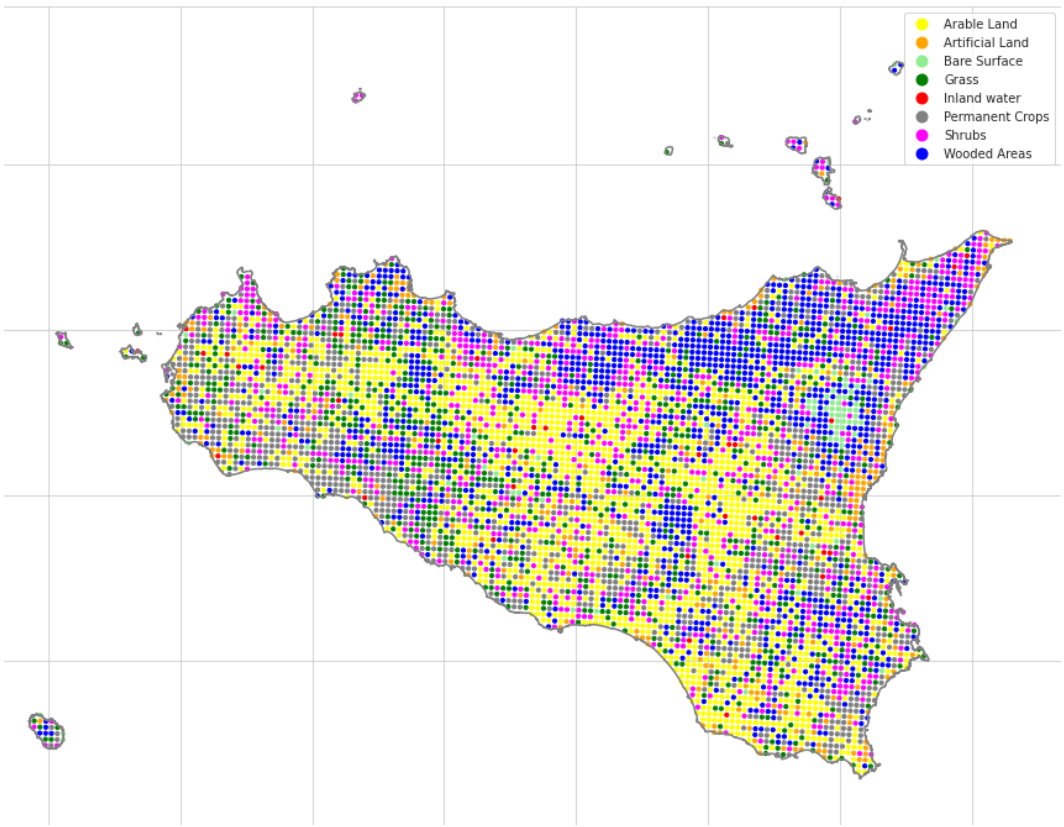

Figure 3.3: Sicily distribution of LUCAS points. (Data Source: LUCAS 2018 [42])

LUCAS points sampling starts from a 1 Km2 grid of 4.400.000 points across the full territory of the European Union. A subset of 1.1 million points that make up a 2km2 grid are divided in land cover classes through remote sensed images. Among these points, a stratified sub-sample of points is selected by an iterative algorithm to run an in-situ survey. To reduce the expense of data gathering effort, sites over 1500 metres high or far from the road network are excluded from the second phase subset. However, these locations are analyzed with image interpretation and classified using regression models that also incorporate data from prior surveys.

In Sicily there are 6423 points of LUCAS grid, and Figure 3.3 shows their distribution: the most common classes are arable land with 30%, wooded areas with 20% and permanent crops with 17%. The major difference between LUCAS and CLC (Figure 3.1) is how the labelling is accomplished: the majority of LUCAS points are in-situ data, whereas CLC is a processing carried out using LUCAS points and multispectral images, without any additional in-ground information. Regarding this work, the LUCAS dataset is preferred due of its detailed information. Additionally, EUROSTAT will unveil a fresh, updated version of LUCAS before the end of 2022 [43].

3.2.2 Multispectral data: Sentinel-2

The European Space Agency (ESA) developed the Sentinel-2 mission to acquire high spatial resolution optical imagery as part of the Copernicus programme of the European Union (EU).

Band	Num	Wavelenght (nm)	Bandwidth (nm)	Res (m)
Coastal	1	442.7	21	60
Blue	2	492.4	66	10
Green	3	559.8	36	10
Red	4	664.6	31	10
Red Edge 1	5	704.1	15	20
Red Edge 2	6	740.5	15	20
Red Edge 3	8	832.8	20	10
NIR 1	8	832.8	106	10
NIR 2	8A	864.7	21	20
Water Vapour	9	945.1	20	60
SWIR 1	10	1373.5	31	60
SWIR 2	11	1613.7	91	20
SWIR 3	12	2202.4	175	20

Table 3.1: Sentinel-2 spectral bands.

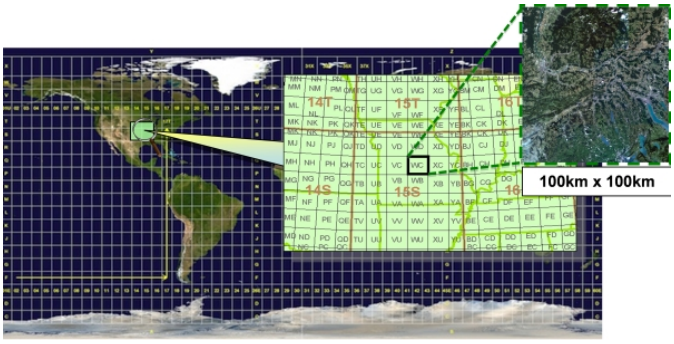

Figure 3.4: Sentinel-2 UTM Tiling. (Data Source: ESA [44])

The Sentinel-2 mission improves the continuation of services that monitor the earth's surface using two satellites Sentinel-2A (launched on 23 June 2015) and Sentinel-SB (launched on 7 March 2017). The satellites are polar-orbiting phased at 180 degrees to each other in the same Sun-synchronous orbit at a mean altitude of 786 km, allowing them to achieve a high revisit time (10 days at the equator with one satellite and 5 days with two satellites under cloud-free conditions). The satellites also carry a MultiSpectral Instrument (MSI), made by Astrium SAS (France), that acquires images composed of 13 different bands: four bands at 10 m (meaning each pixel of the sensed image covers an area of 10 m x 10 m), six bands at 20 m and three bands at 60 m spatial resolution as listed in Table 3.2.2.

Every image the MSI instrument collects is carefully processed at multiple processing levels:

- **Level-0** (L0) products: raw data packaged for long-term storing and upcoming reprocessing processes;
- **Level-1A** (L1A) products: Instrument data that has not been compressed with rough pixel alignment between images from different spectral bands and detector modules. Neither radiometric adjustments nor resampling have been used. These goods are utilised for calibrating;
- **Level-1B** (L1B) products: Top-of-atmosphere (TOA) radiances with complete radiometric adjustments. These items are used for quality assurance, calibration, and validation;
- **Level-1C** (L1C) products: TOA reflectances in the geometry of cartography;
- **Level-2A** (L2A) products: Bottom-Of-Atmosphere (BOA) reflectances in cartographic geometry.

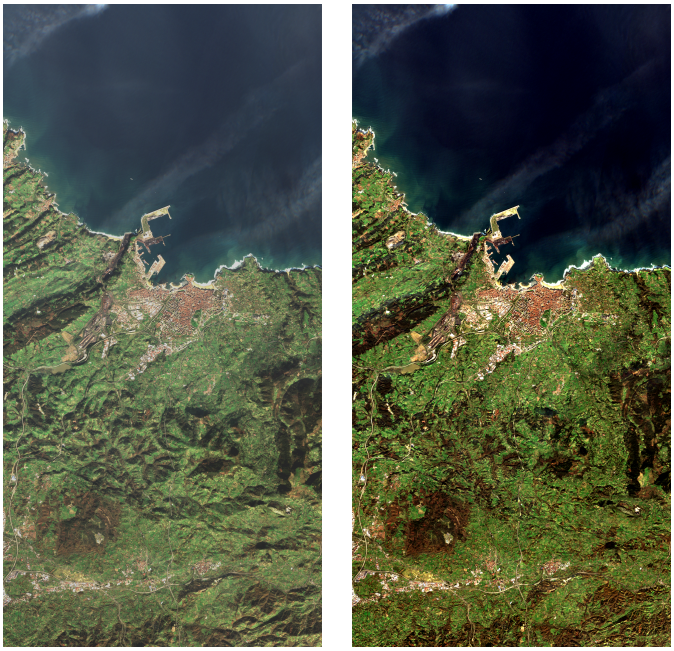

Figure 3.5: L1C and L2A data comparison. Top-of-atmosphere (TOA) Level-1C image data on the left and associated Level-2A Bottom-of-atmosphere (BOA) image data (right) (Image Source: ESA [45])

The granules, also known as **tiles**, for Level-1C and Level-2A are 100x100km2 ortho-images in the UTM/WGS84 projection. The 60 zones that make up the Earth's surface are determined by the UTM (Universal Transverse Mercator) system as shown in Figure 3.4. Images can completely or partially cover tiles depending on their orbit. Both L1C and L2A product are publicly accessible: while L1C from data are available from 2016, while already atmospherically corrected L2A images are from March 2018. Figure 3.5 shows that BOA images (L2A) are clearer and shaper than TOA ones(L1C) and it is for this reason that in sperimental works atmosferically corrected images area used. As wildfire data used in this work dates back to 2016, all images of 2016 and 2017 needs to be atmosferically corrected to obtain L2A products, and so L1C Sicily products will be processed using Sen2Cor ESA processor [46], briefly described in the next section.

Sen2Cor processing tool

The main Sen2Cor processing steps are shown in Figure 3.6. The processing is split into two parts: Scene Classification (SC) that provides a classification map of the image, and Atmospheric Correction which aims at transforming TOA reflectance into BOA reflectance. SC includes four different classes for clouds and six different classifications for shadows, cloud shadows, vegetation, soils/deserts, water and

snow as shown in Figure 3.7. Thanks to this mapping the recovery of Aerosol Optical Thickness (AOT) and Water Vapour (WV) content and Cloud Detection are possible, WV and AOT lead to the final conversion of TOA to Bottom-Of-Atmosphere (BOA). So, Scene Classification Layer (SCL), Quality Indicators for cloud and snow probability, AOT and WV maps, and surface (or BOA) reflectance images, all of which are available at various spatial resolutions, are the Level-2A outputs (60 m, 20 m and 10 m).

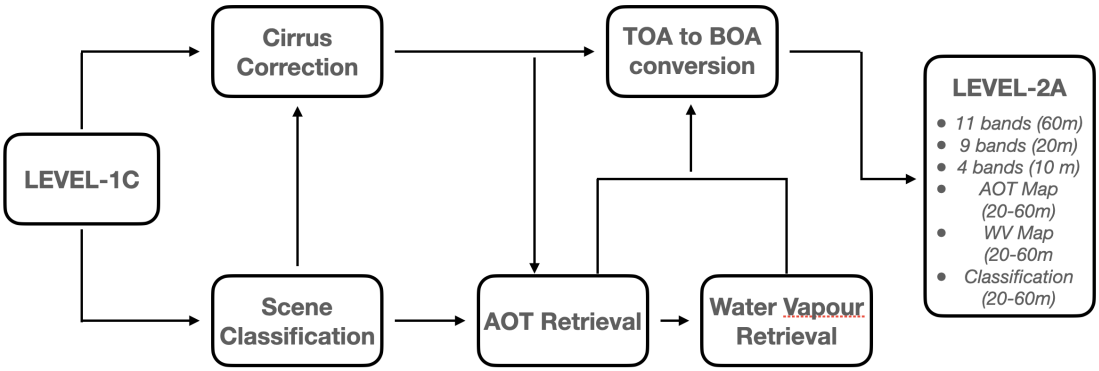

Figure 3.6: Sen2Cor processing steps. Scene classification is made as a first step to perform Cirrus correction, Aerosol Optical Thickness (AOT) and Water Vapour content retrieval. After that, the Top-of-Atmosphere (TOA) to Bottom-of-Atmosphere (BOA) correction is done, to finally produce a L2A image.

Label	Classification
0	NO_DATA
1	SATURATED_OR_DEFECTIVE
2	DARK_AREA_PIXELS
3	CLOUD_SHADOWS
4	VEGETATION
5	NOT_VEGETATED
6	WATER
7	UNCLASSIFIED
8	CLOUD_MEDIUM_PROBABILITY
9	CLOUD_HIGH_PROBABILITY
10	THIN_CIRRUS
11	SNOW

Figure 3.7: Scene Classification Layer Labels. Scene classification classes produces as the first step from Sen2Cor processing tool.

Spectral indexes

Vegetation indices are considered straightforward methods for obtaining specific information from remote sensing products. They are computed mathematically from spectral bands without any assumption on the study region, such as climatic conditions, vegetation that is there, etc.

The selection of the presented pool of indexes has been carefully done using literature focusing on the main task of this thesis. Here's a list of all the chosen families with the description and the formula for every index. For major clarity, formulas are written using both general bands names and Sentinel-2 names.

Broadband greenness family. The broadband greenness indexes are used to evaluate the overall density and vitality of greenery. They are combinations of reflectance metrics sensitive to the combined effects of canopy chlorophyll concentration, foliage chlorophyll concentration, canopy leaf area, foliage clumping, and canopy architecture. They give an indication of the quantity of photosynthetic material present in vegetation overall, which is crucial to understand the condition of vegetation for the prediction of a wildfire event. Here's a list of the two broadband greenness indexes used in this work:

- **NDVI** (i.e Normalized Difference Vegetation Index) [47], [48], [49]: it serves as a measure for abundant, healthy vegetation; it is sensitive to the effects of foliage chlorophyll concentration, canopy leaf area, foliage clumping and canopy architecture. This index's value lies between -1 and 1 and the typical range for greenery is between 0.2 and 0.8:

$$NDVI = \frac{NIR_1 - RED}{NIR_1 + RED} = \frac{B8 - B4}{B8 + B4}$$

- **MSAVI2** (i.e Modified Soil Adjusted Vegetation Index 2) [48]: it is a variation of the MSAVI. The first version of MSAVI suppresses the effects of soil pixels using a canopy background adjustment factor (L), which is a function of vegetation density and frequently necessitates prior knowledge of vegetation amounts. MSAVI2 is based on an inductive method that does not employ a constant L value to emphasise healthy vegetation, and it also decreases soil noise and enhances the dynamic range of the vegetation signal:

$$\begin{aligned} MSAVI2 &= \frac{2 \cdot NIR_1 + 1 - \sqrt{(2 \cdot NIR_1 + 1)^2 - 8 \cdot (NIR_1 - RED)}}{2} \\ &= \frac{2 \cdot B8 + 1 - \sqrt{(2 \cdot B8 + 1)^2 - 8 \cdot (B8 - B4)}}{2} \end{aligned}$$

Narrowband Greenness family. Narrowband greenness indexes are a combination of reflectance measures sensitive to the different effects of canopy leaf area, foliage clumping, and narrowband chlorophyll concentration. The purpose of these indexes is still to offer a measure of the total quantity and quality of photosynthetic material in vegetation, but they are made to be more sensitive to tiny changes in vegetation health than the broadband greenness indexes. There is not a direct connection with a fire event, these indexes can be used to detect and monitor forests which are one of the main fuels for a fire.

- **RENDVI** (i.e Red Edge Normalized Difference Vegetation Index) [47]: This index is an adjustment to the standard broadband NDVI. Instead of employing the primary absorption and reflectance peaks, this index uses bands along the red edge. It takes advantage of the vegetation's red edge's sensitivity to small variations in canopy leaf content, gap fraction, and senescence. This index's value lies between -1 and 1. The typical range for greenery is between 0.2 and 0.9:

$$RENDVI = \frac{E.EDGE_2 - E.EDGE_1}{E.EDGE_2 + E.EDGE_1} = \frac{B6 - B5}{B6 + B5}$$

Burnt indexes family. Burnt indexes use NIR and SWIR bands to highlights burned areas, in this way it is also possible to monitor the recovery phase of an already burnt area. These indexes may not be able to distinguish between water bodies and burned areas because of its low-reflectance characteristic, so the water content family are important also to quantify areas unrelated to flames in order to get over this problem. In general, burned areas reflect more strongly in the shortwave infrared band than they do in the near infrared, on average, so another index is introduced which considers only SWIR bands.

- **NBR** (i.e Normalized Burn Ratio) [49], [38] : This indicator draws attention to burned regions in large fire zones with a normalised difference using NIR and SWIR wavelengths. In general, healthy vegetation is indicated by a high NBR value, whereas bare ground and recently burned areas are indicated by a low value:

$$NBR = \frac{NIR_1 - SWIR_3}{NIR_1 + SWIR_3} = \frac{B08 - B12}{B08 + B12}$$

- **NBR2** (i.e Normalized Burn Ratio2) [49]: it modifies NBR to highlight water sensitivity in vegetation and may be useful in post-fire recovery studies:

$$NBR2 = \frac{SWIR_2 - SWIR_3}{SWIR_2 + SWIR_3} = \frac{B11 - B12}{B11 + B12}$$

Leaf Pigments. These indexes do not measure the chlorophyll content as greenness indexed, but the concentration of stress-related pigments as they are are present in higher concentrations in weakened vegetation. These two pigments are usually anthocyanins and carotenoids, but only the first ones are usually used to detect forest.

- **ARI** [47]: Higher plants often include anthocyanins, so pigments responsible for their red, blue, and purple colouring. They are thought of as indicators of many kinds of plant stressors, they offer useful information on the physiological status of plants. In general, increases in ARI are a sign of new growth or dying vegetation in the canopy. This index takes advantage of the absorption characteristics of stress-related pigments by using reflectance measurements in the visible spectrum:

$$ARI = \frac{1}{GREEN} - \frac{1}{R.EDGE_1} = \frac{1}{B3} - \frac{1}{B5}$$

Water Content family. The canopy water content indexes are an indication of how much water is present in the canopy of leaves. A plant's water content is crucial since a plant with more water tends to be healthier, develop more quickly, and be more resistant to fire. These indexes employ reflectance measurements to obtain measurements of the total column water content and so are really useful in a fire prediction scenario as they can provide a picture of the fuel moisture content.

- **NDWI** (i.e Normalized Difference Water Index) [34]: it is sensitive to changes in water content of vegetative and it is able to detect subtle changes in water content of the water bodies. It ranges from -1 to 1

$$NDWI = \frac{GREEN - NIR_1}{NIR_1 + GREEN} = \frac{B3 - B8}{B3 + B8}$$

- **NMDI** (i.e Normalized Multi-band Drought Index) [34], [49]: it measures the water content of the plant canopy by taking into account a soil moisture background to monitor potential drought conditions. The of NIR and SWIR bands eliminates water changes due to the internal structure and dry matter content of the leaf, increasing the precision of the calculation of the vegetation's total column water content [50]. In general, this index values range from 0.7 to 1 for dry soil, 0.6 to 0.7 for soil with intermediate moisture, and less than 0.6 for wet soil

$$NMDI = \frac{NIR_2 - SWIR_2}{NIR_2 + SWIR_2} = \frac{B8A - B11}{B8A + B11}$$

3.3 Weather data

Weather is another element of the fire environment and it will be included in this analysis with the use of **ERA5** dataset. ERA5 is a reanalysis dataset produced by ECMWF(European Centre for Medium-Range Weather Forecasts). It combines vast observations from satellites, aircraft, land and sea based weather sensors with atmospheric model data into a globally complete and consistent dataset using the laws of physics. This method is uses numerical weather prediction models to forecast weather parameters which are later combined with in-situ/satellite observations in a physical optimal way. ERA5 provides hourly data on many atmospheric, land-surface and sea-state parameters, on regular latitude-longitude grids at $0.25° \times 0.25°$ (approx 28km) lat/lon resolution.

The family of ERA5 datasets comprised the **ERA5-Land** dataset [51], which is a land surface dataset produced at higher resolution ($0.1° \times 0.1°$, approx 9km) with no additional data assimilation [52]. This enhanced spatial resolution makes this dataset very useful for all kind of land surface applications [51].

Relying on the variables used in all physical wildfire models, in the table 3.3 there is the list of the weather parameters used and how they were aggregated to have a weekly value. Following, there are a more detailed description for some less intuitive parameter.

Parameter	**Unit**
Temperature ($2t$)	K
Eastward wind speed ($u10$)	ms^{-1}
Northward wind speed ($v10$)	ms^{-1}
Relative humidity (u)	%
Surface solar radiation downwards ($ssrd$)	$J\ m^{-2}$
Total Precipitation (tp)	m

Table 3.2: Weather variables and statistics used in this work. Temperature, wind, precipitation and solar radiation are native ERA5Land paramenters, while relative humidity has been computed using vapour pressure and dewpoint temperature.

Relative Humidity. It is a parameter which describes the amount of water vapour present in air expressed as a percentage of the amount needed for saturation at the same temperature. It is not archived directly in ERA5 datasets, but using temperature (T), and dewpoint temperature (T_d), is possible to compute relative humidity as the ratio of vapor pressure ($e(T_d)$) to saturation vapor pressures ($e_s(T_d)$) [53].

So the relative humidity u is given by

$$rh = \frac{e(T_d)}{e_s(T)}$$

where $R_{dry} = 287.0597$, $R_{vap} = 461.5250$, $a_1 = 611.21$ hPa, $a_3 = 17.502$, $a_4 = 32.19$ K and $T_0 = 273.16$ K are constant parameters taken form Chapter 7 of [54]. Saturation vapour pressure $e_s(T)$ is computed with the following formula

$$e_s(T) = a_1 e^{a_3\left(\frac{T_d - T_0}{T_d - a_4}\right)}$$

Low humidity takes moisture from the fuels, and fuels in turn, take moisture from the air when the humidity is high. Light fuels gain and lose moisture quickly with changes in relative humidity, so when the the value of relative humidity drops, fire behavior increases as they become drier. While, heavy fuels, respond to humidity changes more slowly.

Surface Solar Radiation Downwards. It is the amount of solar radiation reaching the surface of the Earth, then scattered, absorbed or transmitted by the atmosphere and reflected or absorbed by the surface[51]. This variable comprises both direct and diffuse solar radiation as follows

$$S_{surf}^{dn} = S_{surf}^{dn,direct} + S_{surf}^{dn,diffuse}$$

where $S_{surf}^{dn,direct}$ is the surface solar radiation direct and $S_{surf}^{dn,diffuse}$ surface solar radiation diffuse, as shown in Figure 3.8.

This variable is a good approximation of the total amount of energy provided by solar radiation. Indeed, thermal radiation is one of the main heat transfer processes and, as previously underlined, heat is necessary to reach the ignition temperature.

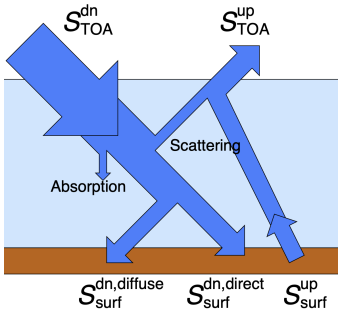

Figure 3.8: Schematic of the short-wave radiative energy flows in the atmosphere. Radiation from the Sun is partly reflected back to space by clouds and particles in the atmosphere (aerosols) and some of it is absorbed. The rest is incident on the Earth's surface (Source ECMWFF [55]).

3.4 Topography data

The topographical features will be obtained from European Digital Elevation Model developed by Copernicus Programme [56]. A digital elevation model (DEM) is a representation of the topographic surface of the Earth's naked ground (bare earth), without considering trees, structures, or other surface items. It is a raster file with an elevation value for each 25x25 m2 area.

Elevation is not the only interesting topographical characteristic, indeed starting from EU-DEM other variables maybe computed, such as slope and aspect.

- *Elevation* is required for temperature and humidity adjustment and it's given directly by EU-DEM, so no further computations are required;
- *Slope* is useful to understand direct effects on fire spread, determining the angle of incident solar radiation to alter fuel moisture, and converting spread rates and directions from the surface to horizontal coordinates. It was computed with algorithm implemented in RichDEM [57] based on [58]. (Figure 3.9);
- *Aspect* identifies the directions of the slopes faces. For example, a south-facing hill is significantly hotter, dryer, while north-facing slopes receive less direct sunlight due to their orientation, so fuels may have very different characteristics. It was computed with algorithm implemented in RichDEM [57] based on [58]. (Figure 3.9)

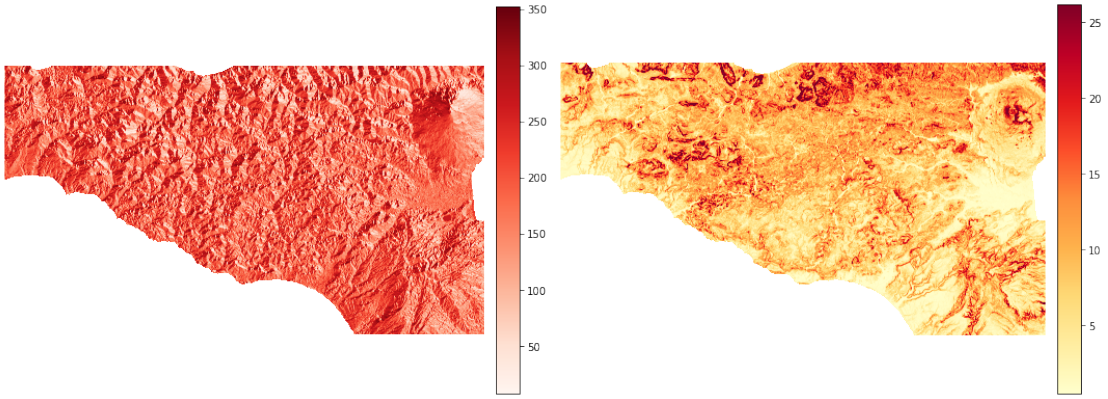

Figure 3.9: Sicily topographical features. On the left, the aspect map in degrees describes the direction of the maximum slope of the cell, while the map on the right describes values in degree of the slope on the area. (Data Source: DEM Copernicus [56])

3.5 Feature engineering sub-framework

This section outlines every step that was taken to construct the training and test sets. The train dataset will span the years 2016 through 2020, while the test will include all of the data from the year 2021. Older wildfire data were available, but as Sentinel-2 hadn't been launched until 2016, all spectral indexes data would have been missing.

There are four sequential step performed: subarea selection, dataset core building, dataset features computing and dataset filling. Every step will be explained in the following sections

Subarea selection

Sentinel-2 tiles were used to select the subarea since it is the most effective technique to obtain the wider burnt region while limiting the amount of MS image storage. Figure 3.10 depicts 12 distinct tiles, each covering an area of 100x100km2 and it is clear how many tiles are sea for the greatest part. In addition to that, given that each MSI image is approximately 1 Gb, that there are between 70 and 100 images accessible each year, and this thesis spans 6 years, the necessary storage would have been the order of 50 Tb. So two tiles have been chosen considering the statistics shown in Figure 3.11: amount of land in the single tile, amount of Sicily land in the tile and number of fires. The tiles **33SVB** and **33SUB** contains 4745 fires over the 6860 provided by Comando del Corpo Forestale della Regione Siciliana. They are highlighted in green in Figure 3.10.

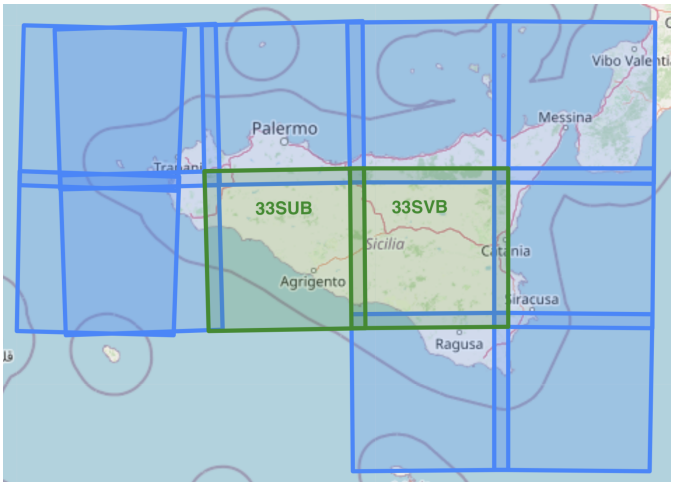

Figure 3.10: Sicily Sentinel-2 Tiles. Sentinel-2 tiles are 100x100 km^2 areas in UTM projection. Squares in the picture represent tiles that intersects Sicily: the green ones are the two tiles used in this thesis.

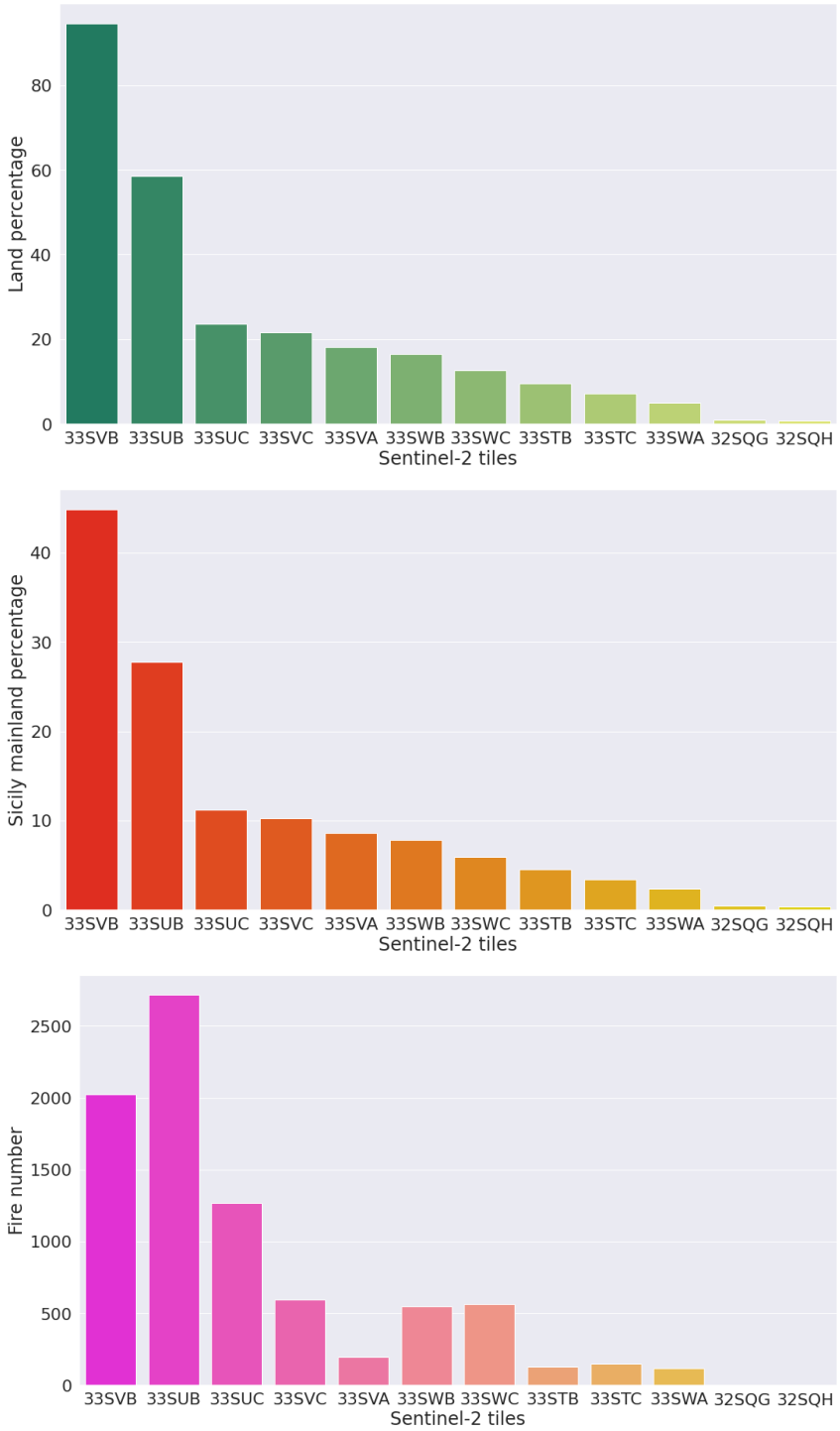

Figure 3.11: Sentinel2 tiles over Sicily statistics. The green distribution shows the percentage of land for each tile; the red distribution describes the percentage of land for the entire region; the pink distribution is the number of fires per tiles.

Dataset core building

Since feature extraction from raw data is a finite operation, the entire subarea is covered by the grid. A point grid would be denser and have superior spatial resolution, but it will not be able to consider the area in its entirety, whereas a square grid would provide statistics over the area and not punctual information, but can cover the full area without missing anything. Considering the task and the spatial resolution of the features used, the choice has been to use a square grid, more precisely a grid of 200x200 m^2 squares, shown in Figure 3.12.

The gridded area contains 43963 distinct areas 10.01% of which have experienced at least one burn. There are many unburnt areas which are undoubtedly really similar cases, so it is useless to consider one instance for one area as it does not help the model to distinguish between burned and unburnt areas. Instead, the goal is to feed the model varying temporal information for a particular area so that it can be used to rebuild a historical perspective. With an average of 3/4 multispectral images per month and a 20% cloud coverage, it is impossible to perform it as a daily task, so the temporal resolution of this work is being set at one week.

According to the theory, If there has been a fire in a certain area on a specific week, the fuel's moisture content has probably dropped in the weeks before, or the current week's weather forecast does not call for rain. Additionally, only 67.58% of burnt areas experienced only one fire, so there are regions where there have been several fires: 20.55% had two fires, 7.14% three fires and so on. Probably, some fires have overlapped and some of are just very close and fall in the same square. But for this reason, it's also crucial that the model recognises that just because a region has previously burned, it doesn't indicate that it can't do so again or even more than once. There are areas that have a burnt 11 time from 2016 to 2021, for example. Due to this, a binary *already burned* feature and a *previously fires count* are two additional features that are taken into account.

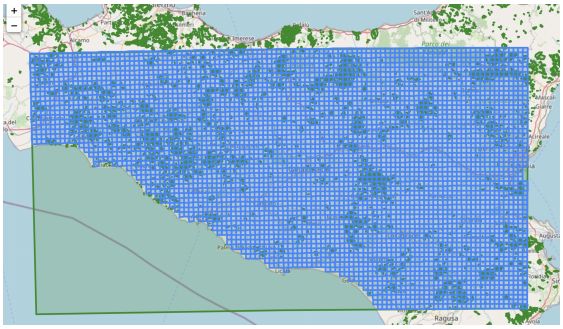

Figure 3.12: Gridded subarea example. The grid covers only the land portion of the area and the green polygons under blue gridded squares are areas burnt between 2016 an 2021.

Following this idea, each area will have multiple instances in the dataset, describing a sort of time series for its features. A single fire event area is filled in the dataset as follows:

- Present instance: the week-year of the fire event;
- Past instances: there are two different instances, the older one is exactly one year before the fire event, in order to have seasonality pattern that didn't lead to fire while the more recent one is taken only 4 weeks before the event to have trend behaviour;
- Future instance: 6 weeks after the fire event. In these instances the already burnt feature is set to 1 and the count increases.

In multi-fire instances, the same theory would produce inaccurate information. Consider a location that experienced a fire in week 30 of 2017 and another fire in week 25 of 2018. When considering the second event, the older past instance on week 25 of 2017 would not have been already burnt, but on week 21 of 2018 the fire already happened, so the flag should be set to False in the first past instance and to True in the more recent instance. Since the complexity of these cases increases when there are more fires in a region, each multiple case has been handled independently to avoid having this illogical information.

Both the train set and the test set are designed with this concept in mind as even the test set has multi-fires areas, indeed 6% of burnt areas had two different fire events in 2021.

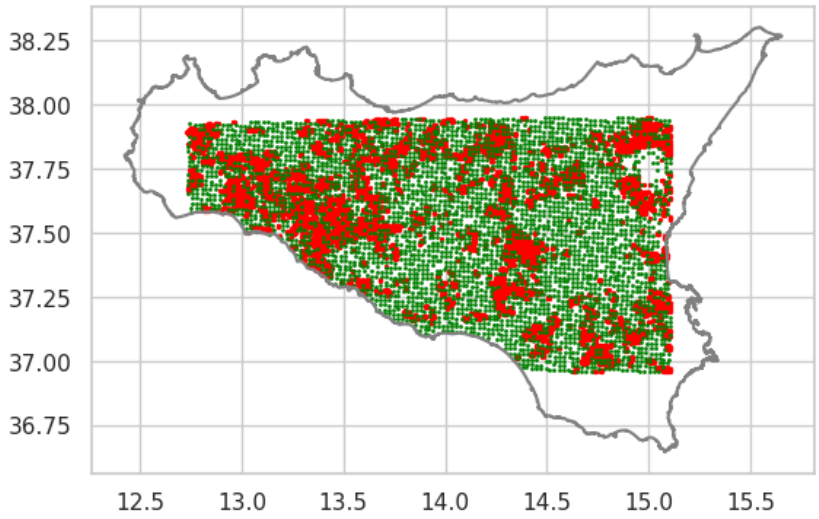

Figure 3.13: Areas selected for the dataset. The plot shows the centroids of the selected areas: the red dots are burnt regions while green ones are vegetation areas never burnt.

Without any further addition, the dataset would at this point consist entirely of locations that have burned at least once. This idea is really far from reality, thus it is necessary to select some areas that have never burned. The never-burnt areas sampling cannot be performed randomly to have a uniform spatial distribution so LUCAS dataset's points have been considered: all never-burnt squares which contained a LUCAS points labelled with a vegetation class have been added to the dataset. Figure 3.13 shows how this idea lead to a spatial uniform dataset.

It is crucial to emphasise that all these criteria are applied to both the training and test sets, which should present a picture as accurate to reality as possible. As a result, there should be areas that have already burned, areas that have never burned where a fire occurred, and areas that have never burned. In addition to that, the spatial area information have been omitted in order to make the model as general as possible.

At the end of this phase it is possible to check the label distribution. Figure 3.14 displays the target distribution in both training and testing: A fire event is described by 20% of instances in both datasets. This imbalance is obviously out of proportion with reality, which finds a maximum rate of 7 percent for burned areas in 2021. Even though it is not a perfect representation of reality, this 80/20 imbalance is a significant imbalance for a machine learning model and will require special attention.

The creation of a dataset that could accurately depict the seasonal fire trend was the other key goal. In Figure 3.15 the dataset's train and test distributions are compared to actual yearly distributions and it is clear that they all have a very similar trend.

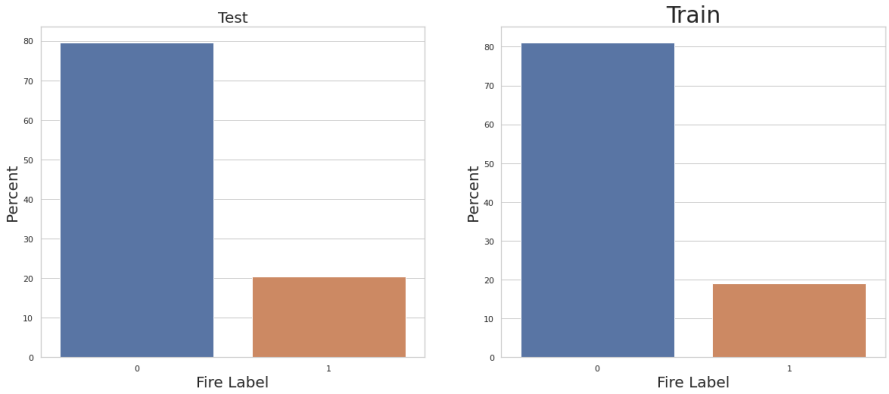

Figure 3.14: Target distribution in the dataset. These pltos shows target distribution in the test (on the left) and in the train (on the right). The positive class (orange) are the instances of a fire event occurence, while the negative class (blue) are non-fire cases.

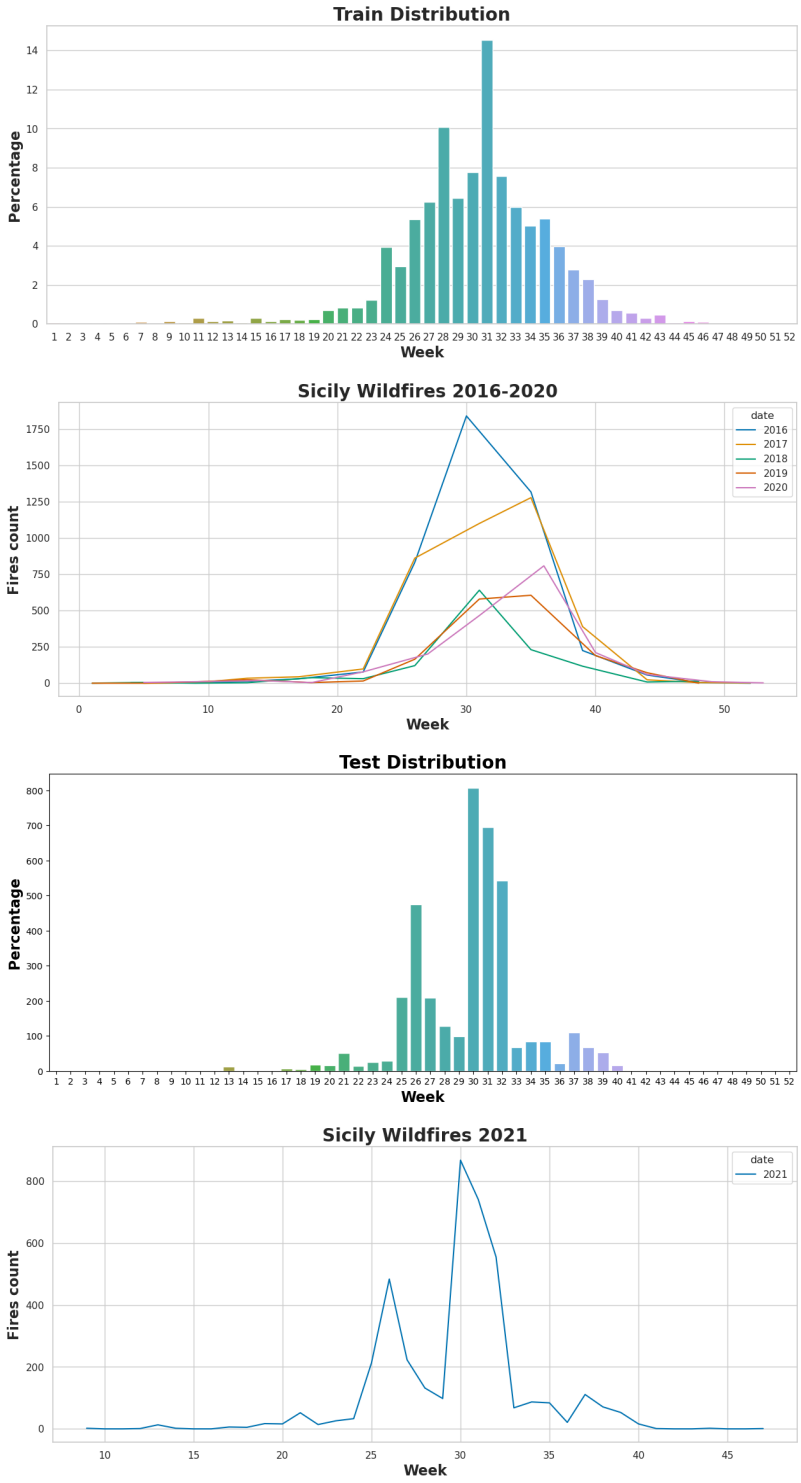

Figure 3.15: Train and test distributions. The histograms show positive fire instances of the dataset, while the lineplots presents is the true fire trend.

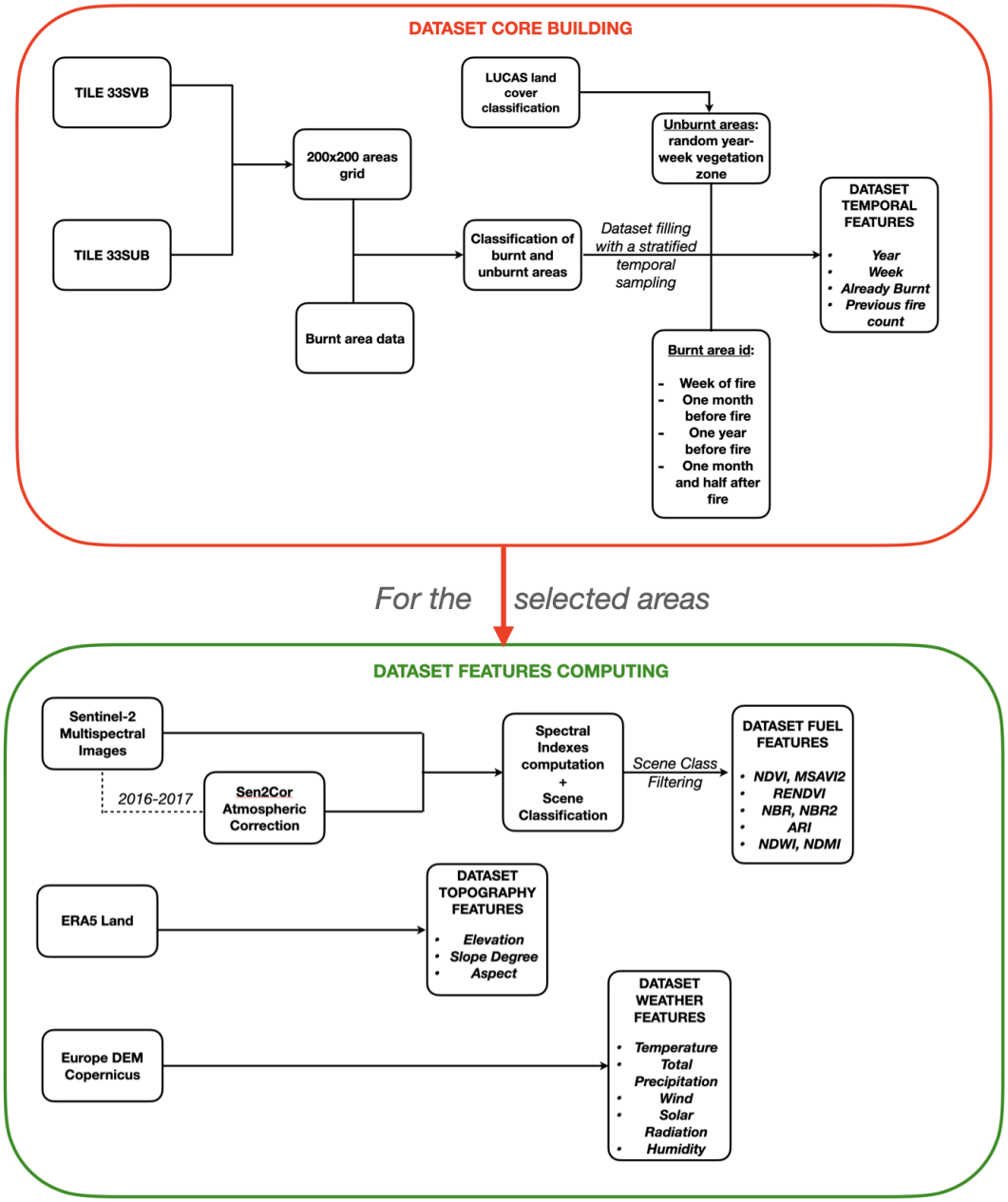

Figure 3.16: The second and the third step of the dataset building. The result of the subarea selection was the union of two Sentinel-2 tiles: 33SUB and 33SVB. In the dataset core building phase (red), the squares of the subarea grid were divided into burnt and unburnt in order to fill the dataset with different instances that respect the seasonal fire trend and to add never-burnt instances. For this subset of areas, all the features have been computed in the third step (green). These features have been statistical aggregated both spatially (over the area) and temporally(for a given week).

Dataset features computations

A preliminar step for this phase is the acquisition of Sentinel-2 images. It is significant to mention that clouds are a significant problem for multispectral images, since the presence of a cloud in the image can alter the value of the bands, and as a results, of the derived spectral indexes. Due to this, an option called "cloud coverage" that designates the area of the sky that is typically veiled by clouds can be adjusted while images are being retrieved. In this work, a 20% cloud coverage was chosen, yielding a total of around 500 photos for 650 Gb. Because L2A products weren't accessible until 2018, L1C products for 2016 and 2017 were downloaded, and they were atmospherically corrected using the Sen2Cor (v5.5) tool to have uniform images (L2A) for the full time period taken into consideration.

It is now possible to compute the feature required to fill the dataset using the areas chosen in the dataset core building stage (Figure 3.16).

- Topographical features have been computed for every area id using RichDEM package [57];
- Hourly weather features for each area were extracted from ERA5Land and relative Humidity was computed using MetPy package [53]. Temperature, Wind, Solar Radiation and Relative Humidity were aggregated by week with mean and standard deviation. Total Precipitations were instead aggregated as the sum of weekly precipitations;
- Spectral indexes for the selected areas where computed from multi-spectral images using zonal statistics of mean and standard deviation. In addition, a scene classification label, selected with majority voting, was matched to every single area id.

Multispectral images could have some anomalies and the scene classification layer is used to filter these cases: for example, even if the cloud coverage is low, there is still a possibility of finding a cloud, which will alter the value of computed spectral indexes. So, all instances with scene classification labels not in range 3-8 are dropped (see Figure 3.7 for scene classification labels).

Three new datasets, one for each feature, will be produced as a result of these calculations: the topography dataset will simply contain variables for all locations that were chosen, while the weather and spectral index datasets both have a time series structure. Weekly weather time series don't have any missing values, but spectral indices do. There are two types of missing values: those related to the absence of data during the specified week, maybe as a result of high cloud cover, and those related to scene classification filtering, therefore the number of missing values may vary by region.

Dataset filling

The efforts done in the dataset core and the features computations steps are combined in this final stage. Keep in mind that a unique dataset core instance is given by area, year, and week, to fill each instance, area id is used to pick the topographical features, while year and week attributes are used to extract values from time series datasets of weather and spectral indexes corresponding to specific weeks (Figure 3.17). Since the only information that would be available in a real-world scenario to estimate a fire risk label would come from weather forecasting, weather data are taken for the current week. Instead, fuel data are instead taken for the two weeks prior to the event as an information about the fuel status trend. When the last two weeks were missing for the spectral indexes time series, the last two available information were taken.

So, an instance of the final dataset would have five different information:

- *Past information* with fuel data of the last two weeks;
- *Future information* with weather forecasting;
- *Area fire regime information* with the count of previous fires and the already burnt flag;
- *Seasonality information* thanks to year and week attributes;
- *Topographical information* for the considered area.

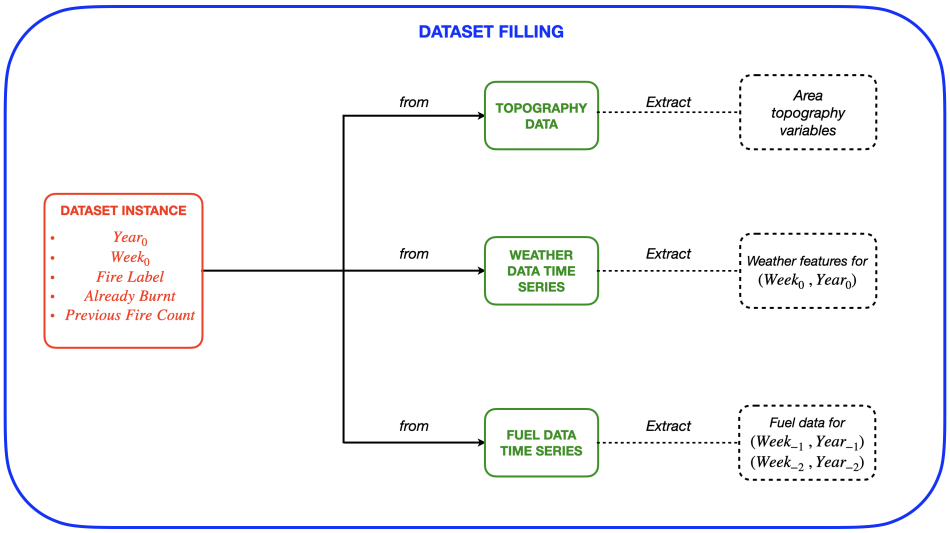

Figure 3.17: Dataset filling step considering one area. For each instance of the output of dataset core step(red), features are extracted from the three outputs of the dataset features computation step(green).

Chapter 4

Methodology

This chapter presents all the methods of pre-processing and classification used on the dataset created applying the sub-framework presented in the previous section.

The dataset contains features of many formats and scales, ranging from Kelvin to dimensionless indices, so a feature scaling (Section 4.2) is executed along with a feature selection to reduce the tasks's dimensionality (Section 4.1).

The classification task will be performed both using classical machine learning models (Section 4.3), such as Support Vector Machines and K-Nearest Neighbour and using ensemble methods (Section 4.4), such as Gradient Boosting or Bagging. The validation methods and the metrics used are also explained in Section 4.5 and Section4.6 respectively.

From the exploratory analysis carried out in the previous section (Figure 3.14), it emerged that the dataset is highly unbalanced because of the nature of the target. As most of the machine learning methods are built on the assumption of an equal number of examples for each class, imbalanced classification is a challenge. In the final section (Section 4.7), some specific techniques to deal with this problem are presented.

4.1 Feature Selection

The basic idea of feature selection is to lower the number of input variables in order to enhance the efficiency of the model while also lowering the computational cost of modelling.

One popular technique used for features selection is **Recursive Feature Elimination** (RFE). Using all of the features in the training dataset as a starting point, RFE attempts identify a subset of features by sequentially removing one at a time until the desired number of features is left. This is accomplished by first fitting a machine learning algorithm (Random Forest in this case), ranking the features

according to relevance, eliminating the least important features, and then re-fitting the model. This process is repeated until a fixed amount of attributes is present. In Section 5.1 the list of removed features is shown.

4.2 Feature Scaling

Feature scaling is a method for standardizing the independent features in the data over a predetermined range, typically between 0 and 1. It is done to deal with extremely variable magnitudes, values, or units. In this study case there are many different scales, from Kelvin to Pascal, from slope degrees to dimensionless indexes, so feature scaling is essential to avoid classification algorithms prioritising larger values over smaller ones.

The scaling method used in this thesis is **standardization**, which centres the numbers around the mean and uses a unit standard deviation. As a result, the attribute's mean becomes zero, and the distribution that results has a unit standard deviation. The formula for standardization is the following

$$X' = \frac{X - \mu}{\sigma}$$

where μ is the mean of the feature values and σ the standard deviation.

4.3 Classification models

In this thesis the machine learning task performed is supervised classification, so in this section some of the most popular models are presented, with an analysis of their advantages and disadvantages.

4.3.1 Random Forest

The Random Forest is a supervised learning technique that consists in growing multiple decision trees which are later combined to produce a prediction based on majority voting.

Each decision tree starts with a root node and it expands into many branches, forming a structure similar to the one of a tree. It simply asks a question and based on Yes/No answer and it expands into subtrees, each trained on a separate subset of the training data in order to avoid having identical trees.

Classifying a test record is straightforward once a decision tree has been constructed: starting from the root node, the test condition is applied to the record and the appropriate branch based on the outcome of the test is followed. To grow

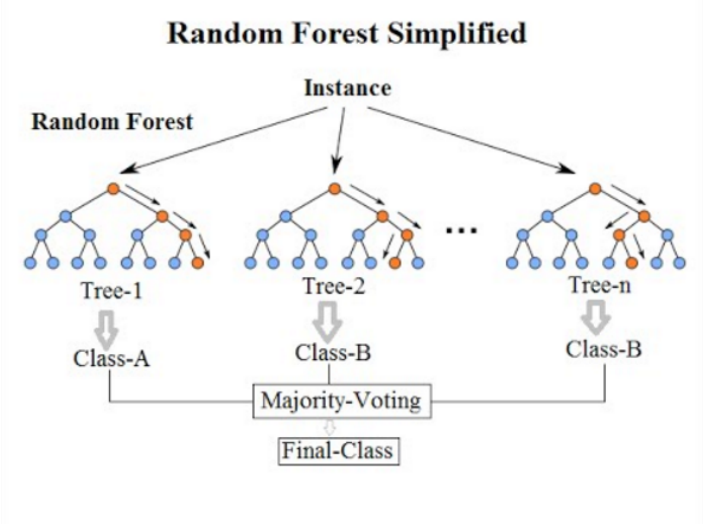

Figure 4.1: Random Forest algorithm.

a tree, however, requires making choices on the features to use, the conditions to use for splitting, as well as understanding when to stop.

Not all features are taken into considerations. Each decision tree only considers a random subset of all the data, often as large as the square root of the entire number of features, to subdivide the nodes. By doing this, even if each tree may have a large variance relative to a specific set of training data, the forest as a whole will have a reduced variance, leading to better predictions thanks to a simple majority voting system among all trees.

Another problem is to choose the optimal decision tree without computing all the possible trees as it is computationally unfeasible. Indeed, in training phase we take a *top-down* (splitting the predictor space from the top of the tree), *greedy approach* (at each step the local best split is performed, rather than the one that would lead to the best global result) that is known as *recursive binary splitting*. The "best split" is the split that minimizes a function that measures the node purity considering all the features. The most common functions used to choose the best split are the *Gini index* and *Cross Entropy*.

The procedure goes on until a stopping requirement is met for example, we may continue until no region contains more than a give number of observations. A common stopping criterion is the maximum depth a single tree can reach, it is the the length of the longest path from a root to a leaf.

In prediction phase each node in every tree acts as a test case for some feature, and each edge descending from the node corresponds to the possible answer to the test case. This operation is done for each subtree rooted at the new node until a terminal node, associated with a given target variable, is reached.

To summarize the characteristics of the random forest:

- **Advantages**:
  - Less prone to overfitting than decision trees;
  - Handles non-linearity, missing values and outliers;
  - Can handle every type of data, both numerical and categorical;
- **Disadvantages**:
  - Complex, not easy to understand;
  - Long training period.

4.3.2 SVM

Support Vector Machines (SVMs) are supervised learning models whose aim is to find the hyperplane that best separates data. It represents each data as a point in a space mapped in a way that objects from different categories are divided by an empty space as wide as possible. The new points will be classified according to which side of the gap they belong to, see Figure 4.3.2.

This gap is called *margin* and it is marked by two *support vectors* that, in a two-dimensional space, are simply lines. These support vector are decided using the two nearest point to the hyperplane, if they are removed or modified they alter the position of the hyperplane.

To find the best hyperplane, the following steps need to be done:

- Find a linearly separable hyperplane that divides values from two classes, if there are more hyperplanes that respect this condition, choose the one with the highest margin;
- If the hyperplane does not exist, SVM uses a non-linear mapping to transform the training data into higher dimension as in this way they can always be divided by an hyperplane.

SVM is based on the strong assumption that an hyperplane can linearly separate data, but this is far from real life. This idea is referred to as "hard margin" classifier. However it is possible to soften this idea by choosing an hyperplane that separate data almost linearly, so the model should be more tolerant towards errors, leading to a greater generalization ability. This is defined as "soft-margin" classifier. To obtain this flexible behaviour a regularization parameter is used. It defines the tolerance of the model towards errors. Smaller values of this parameter $\lambda = 1/C$ leads to bigger margins, with greater tolerance, while bigger values of λ lead to the hard-margin paradigm.

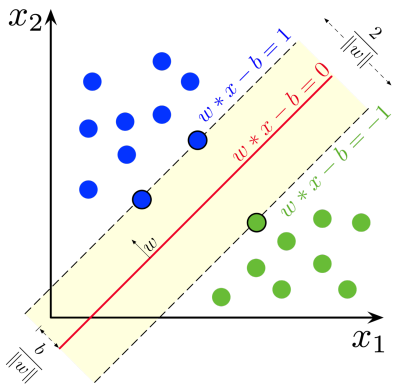

Figure 4.2: SVM.

By mapping non-linearly separable data onto a higher dimensional space, where the data could become linearly separable, the so-called *kernel trick* can be used to expand the SVM method to non-linear instances. A linear SVM model may be then trained to classify the data in the new feature space once the data transposed. This approach is highly expensive to compute, though, so the idea is to use a a kernel function to avoid the need of explicitly translating input data.

So, to summarize:

- **Advantages**:
  - Good when there is no clue on the data;
  - Efficient in high dimensional spaces;
  - Handles non linear cases thanks to kernel trick;
  - Good generalization, so less overfitting
- **Disadvantages**:
  - Sensitive to noisy data;
  - Long training time for large dataset;
  - Difficult interpretation.

4.3.3 KNN

It is a non-parametric classification technique that finds the k closest data points for a given unknown data point and predicts the output class based on the most frequent class among those k neighbours. It based the idea of similarity of distance since it assumes that comparable items occur within close proximity. In fact, it

employs distance metrics like Manhattan distance or Euclidean distance to identify the neighbours.

Since it simply memorises every point without doing anything else during the training phase, it is referred to as a lazy-learning algorithm. This approach led to a relatively quick training period. The intesitve work is done during the validation stage: each time the model has to categorise data, it must compute every distance between that point and all the others, store these distances in memory, and then classify the data by examining labels of nearest k neighbours.

For KNN, the selection of k is crucial: smaller k can result in a lot of noise while larger k can cause the model to classify based on the more prevalent class. Due to the difficulty in classifying the minority class, this approach is not particularly useful in situations when class distribution is skewed.

To summarize:

- **Advantages**:
  - No training period, much faster than other algorithms;
  - New data can be added seamlessly and they will not impact the accuracy of the algorithm;
- **Disadvantages**:
  - High cost for computing the distance between the new point and each existing points in a large dataset;
  - Difficulties with high dimensions because it becomes difficult for the algorithm to calculate the distance in each dimension;
  - Sensitive to noise in the dataset.

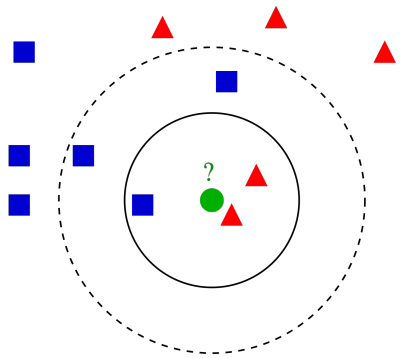

Figure 4.3: KNN. The green dot is classified basing on how many neighbours the model considers: if $k = 2$ it will be classified as a red triangle e, while if $k = 3$ it will be a blue square.

4.4 Ensemble methods

By mixing the predictions from various models, **ensemble learning** is a machine learning approach that tries to improve predictive performance. There are two different types of methods: *sequential ensemble methods* and *parallel ensemble methods*.

In sequential ensemble methods, base learners are created consecutively. The main motivation to use this kind of methods is to use the dependence between the base learners by weighing previously mislabeled examples with higher weight. While parallel ensemble methods are applied wherever the base learners can be generated in parallel. The main idea of parallel methods it to use independence between base learners and reducing the errors by averaging.

Even though there are an apparent infinite amount of ensembles you can create, there are three techniques that rule the world of ensemble learning: bagging, stacking and boosting.

Stacking. The general process of stacking involves training a learner to combine the different learners. The combiner is usually referred to as second-level learner and the individual learners are referred to as first-level learners. Although there might be more layers of models, the most popular hierarchy has two levels. For instance, we might have 3 or 5 level-1 models instead of just one, and a single level-2 model that combines the level-1 models' projections to produce a prediction.

Bagging. The idea behind bagging is to combine the results of multiple weak learners to get a generalized result from a final single model. It is a combination of bootstrapping and aggregation which is possible to reduce variance of an estimate by taking the mean of multiple estimates.

Three steps are needed to perform bagging, as shown on the left in Figure 4.4:

- Create randomly sampled datasets of the original training data (bootstrapping);
- Build and fit several classifiers to each of these diverse copies;
- Take the average of all the predictions and to make a final overall prediction (aggregation).

Random Forest can be compared to Bagging with a small modification. Bagged Decision Trees have access to all available features when selecting where to divide and how to make decisions, so, even if the bootstrapped samples may differ slightly, the data will typically split off with the same characteristics for each model. Instead, Random Forest models select features at random to determine where to split.

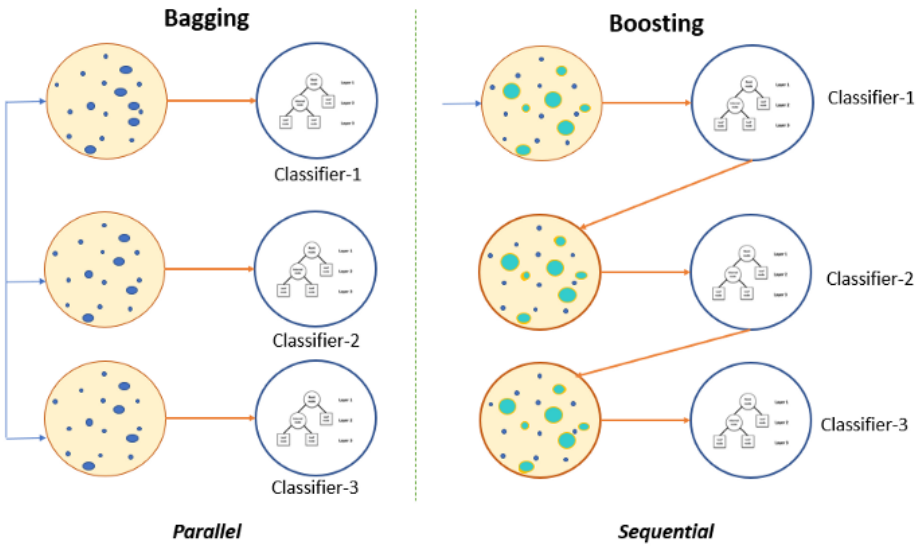

Figure 4.4: Bagging and Boosting. Bagging is described on the left: bootstrapped subsamples are drawn from a dataset, for each a decision tree is created and the prediction is done by averaging the outcomes of each tree. Boosting is shown on the right: every new weak learner tries to correct the errors of the previous learner by weighting data, until a strong learner is obtained as a weighted mean of all the weak learnes. (Image source [59])

Boosting. It is a sequential process where the errors of the previous model is corrected by each subsequent model. Unlike the bagging, boosting relies on the dependence of weak learners. When weak learners update the weights of the data points based on the outcomes of the preceding weak learners, the weak learners become strong learners. By attempting to raise the weight attached to an observation that was mistakenly categorised, boosting modifies that observation's weight. Boosting often reduces bias error, however it can occasionally cause the training dataset to become overfit. In general, the algorithm is able to identify the parameters it needs to concentrate on to perform better thanks to this redistribution of weights. The sequential process of creating and aggregating weak learners varies between boosting techniques.

The steps in the boosting process are as follows:

- The training dataset is split into a subset with identical weights for each data point and for the initial dataset, a based model is developed, and predictions are made using this model for the complete dataset;
- The predicted and actual values are used to calculate errors. The observation that was mistakenly predicted is given more weight;

- Boosting attempts to fix the errors in the preceding model by creating a new weak learner;
- Multiple models are created using the same procedure, each one fixing the errors in the previous model;
- The weighted mean of all the models (weak learners) is the final model, which is a strong learner.

Indeed, the main difference between bagging and boosting (Figure 4.4) is how they are trained: in bagging, each model is built independently of the others and with identical weight, while in boosting the results of the earlier weak learners influence the newer learners which are weighted according to their performance. Bagging typically reduces variance rather than bias. While Boosting seeks to lessen the issue of bias, Bagging attempts to address the issue of overfitting training data.

There are three common categories of boosting techniques:

- **Adaptive boosting or AdaBoost**: In order to reduce the training error, this approach iteratively locates misclassified data points and modifies their weights. The model keeps improving in a sequential manner until the strongest predictor is produced. In general weak learners are decision trees with a single split, called decision stumps which in a complex scenario are usually not enough;
- **Gradient boosting**: it is a technique that gradually adds predictors to an ensemble, with each one repairing the mistakes made by the one before it, but opposed to AdaBoost, gradient boosting trains on the remaining errors from the prior predictor. It treats boosting as a numerical optimization problem where the objective is to minimize the loss function of the model by adding weak learners using a gradient-descent like procedure;
- **Extreme gradient boosting or XGBoost**: it is a gradient boosting system created for speed and scale, indeed through the use of the CPU's numerous cores, XGBoost enables parallel learning during training. It doesn't run multiple trees in parallel, the parallelisation happens during the construction of each trees, then each independent branches of the tree are trained separately.

4.5 Model validation

A training phase and a validation phase are often included in the creation of a classification model. The train/validation split and the k-fold cross-validation technique are the two most used methods for model assessment. When used to classification issues with a strong class imbalance, both techniques have the

potential to fail and can, nonetheless, produce findings that are deceptive. For this reason, in this work is used **stratified 2-fold** cross validation which is is able to stratify the sample by the class label. However, in a real world scenario where fire data are available yearly, if we are testing for 2020 the data for the current year would no be available yet, so it is reasonable to say that a random cross validation in this case it is like cheating. So, another validation approach is tested, where the samples are stratified by year and not by distribution, the folds will be divided as follows:

- First fold
  - Training: 2016-2017-2018
  - Validation: 2019
- Second fold
  - Training: 2016-2017-2018-2019
  - Validation: 2020

4.6 Performance measures

A metric is a function that compares the expected class label to the predicted class label and it gives information about how well the model performs. Generally, problems can be solved using standard metrics, such as accuracy, but they are misleading when classes are imbalanced. Indeed they treat both minority and majority classes equally, without differentiating between the number of correctly classified examples of various classes. In general, in imbalanced classification tasks, errors that affect the minority class are often more significant than those that affect the majority class, and as a result, performance criteria that emphasise the minority class may be necessary.

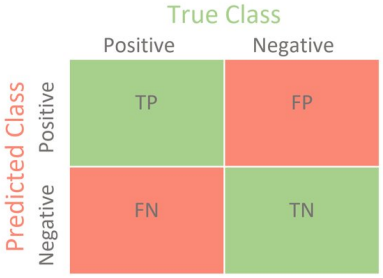

Figure 4.5: Confusion matrix for a binary target.

Each measure that will be presented will be based on a confusion matrix (Figure 4.6, which is a highly helpful tool when dealing with binary classification issues. Using one of the classes as a reference, the values inside the matrix show:

- True Positive (TP): the number of samples that were correctly identified as not belonging to the reference class;
- False Positive (FP): the number of samples that were incorrectly identified as belonging to the target class;
- False Negative (FN): the number of samples that were correctly identified as not belonging to the reference class;
- True Negative (TN): the number of samples correctly predicted as not belonging to the reference class.

Precision is the fraction of True Positive elements divided by the total number of positively predicted units (column sum of the predicted positives). Precision expresses the proportion of units our model says are Positive and they actually Positive. In other words, Precision tells us how much we can trust the model when it predicts an individual as Positive;

$$Precision = \frac{TP}{TP + FP}$$

Recall: it is the fraction of True Positive elements divided by the total number of positively classified units (row sum of the actual positives). It measures the ability of the model to find all the Positive units in the dataset:

$$Recall = \frac{TP}{TP + FN}$$

F-Score. It aggregates Precision and Recall measures under the concept of harmonic mean, where balance of precision and recall in the calculation of the harmonic mean is controlled by a coefficient called β, which is chosen such that recall is considered β times as important as precision, as can be seen from the following formula

$$F_\beta = (1 + \beta^2) \cdot \frac{Precision \cdot Recall}{(\beta^2\ Precision) + Recall}$$

There are three typical values for the beta parameter:

- $F_{0.5}$ ($\beta = 0.5$): Precision is valued more highly than recall;
- F_1 ($\beta = 1$): Equal the importance of recall and precision, most common;

- F_2 ($\beta = 2$): Precision is less important than recall

In this work, F_2 score will be used since it focuses more on minimising false negatives than minimising false positives.

ROC Curves. Receiver Operating Characteristic, or ROC, is an abbreviation that stands for a branch of research that evaluates binary classifiers based on their capacity to distinguish between classes. The behaviour of a model is summarised by a ROC curve (Figure 4.6, which calculates the false positive rate and true positive rate given a series of predictions made by the model under various thresholds.

- True positive rate (TPR), is defined as

$$TPR = \frac{TP}{TP + FN}$$

  it measures the percentage of positive data points that, when compared to all positive data points, are correctly regarded as positive. In other words, we will miss fewer positive data points if TPR is higher;

- False positive rate (FPR) is defined as

$$FPR = \frac{FP}{FP + TN}$$

  This fraction is intuitively related to the percentage of negative data points that are wrongly interpreted as positive. In other words, the more negative data points that are incorrectly identified, the higher the FPR.

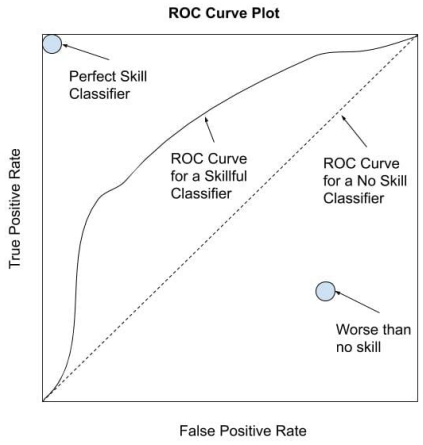

Figure 4.6: ROC curve with false positive rate over true positive rate.

Computing the area under the ROC curve yields a single score that may be used to compare models: a classifier with no skills will receive a score of 0.5, whereas a classifier with perfect skills will receive a score of 1.0. This measure can be overly optimistic in situations of extreme class inequality, particularly when there are few examples in the minority class, despite being usually effective.

An alternative to the ROC Curve in an imbalanced scenario, is the **precision-recall curve** (Figure 4.6), which works similarly to the ROC Curve but concentrates on the performance of the classifier on the minority class. To combine the precision and recall into one single metric the model computes the two metrics with various thresholds and then they are all plot together to make up a single curve. So, classifiers will be scored higher if they outperform others under a variety of different thresholds.

A horizontal line with a precision proportionate to the number of positive cases in the dataset will represent a no-skill classifier, for example in a balanced dataset it will have a value of 0.5. When the classes are severely unbalanced, this statistic is a helpful indicator of prediction success. Indeed, high recall but low precision produces a large number of results, the majority of the predicted labels are different from the training labels. In contrast, a system with high precision but low recall produces very few results, yet most of its projected labels match the training labels. A perfect system with great precision and recall will provide a large number of outcomes, all of which will be correctly categorised.

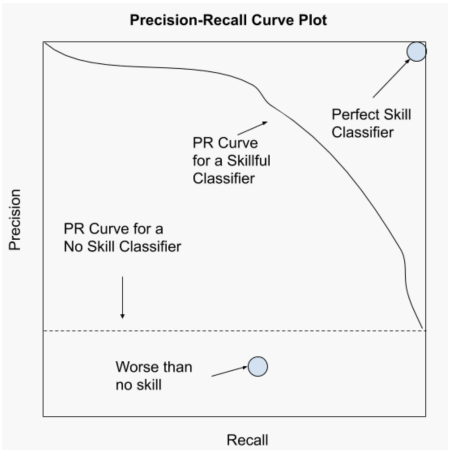

Figure 4.7: ROC curve with recall over precision.

4.7 Dealing with imbalanced data

In the previous section, Figure 3.14 shows that the target distribution is imbalanced. Generally, models are not able to deal with class imbalance without any additional adjustments, so in this sections some sampling solution are presented [60].

Sampling techniques include modifying an unbalanced data set using certain procedures to produce a balanced distribution. The sampling can be done in two directions: eliminating samples from the majority class (under-sampling) or adding data to the minority class (over-sampling). A combination of both technique can be also implemented.

It is important to underline that the training dataset is the only one to which the class distribution has been altered as the goal is to alter how well the models fit. The test dataset used to evaluate a model's effectiveness does not undergo resampling.

4.7.1 Over-sampling methods

There are several sampling methods that increases the number of samples in the dataset:

- **Random.** The minority class instances are randomly duplicated and added to the training set until a more balanced distribution is achieved. In some circumstances, trying to balance out an unbalanced dataset might lead to algorithms being overfit to the minority class, as the model sees the same samples over and over, so the sampling strategy is usually tuned;

- **Synthetic sampling.** The Synthetic Minority over-sampling Techniques (SMOTE) provides the minority class some artificially generated data points so that it can be compared to the majority class. In fact, it selects the k closest minority class neighbours after randomly choosing an instance x_i in the minority class. A synthetic instance is then produced by joining x_i and $\hat{x_i}$ to form a line segment in the feature space, where $\hat{x_i}$ is a k nearest neighbours randomly selected (Figure 4.7.1). One of the major drawbacks of SMOTE algorithm is over generalization, indeed it produces the same number of synthetic data samples for each original minority case without taking into account nearby examples, which enhances the likelihood of class overlap;

- **Adaptive Synthetic Sampling.** To overcome SMOTE overlapping problem, adaptive sampling methods. **Borderline-SMOTE** starts by classifying minority class observations. If all of the neighbours are members of the majority class, it labels any minority observation as a noise point and ignores it while producing synthetic data. So it resamples with a SMOTE technique entirely

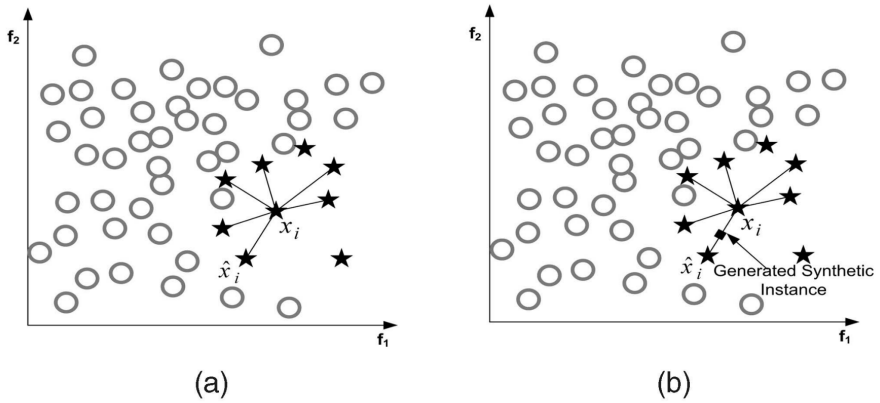

Figure 4.8: SMOTE. On the left, an example of the K-nearest neighbors for the x_i and on the right the data creation based on euclidian distance is shown. (Image Source [60])

from a small subset of points which have the same number of majority and minority class neighbours. One major problem is that it eventually pay more credence to these observations. Another adaptive technique is **ADASYN** which creates synthetic data consistent with their distributions. It determines the impurity of the neighbourhood for each minority observation by taking the ratio of the majority observations in the neighbourhood.

4.7.2 Under-sampling methods

To reduce the number of samples of from the majority class, two main techniques are considered:

- **Random.** The majority class instances are randomly discarded until a more balanced distribution is achieved. The main drawback of removing samples is that it impossible to save more informative instances from the majority class. In addition, random undersampling is used also in methods like random forest and bagging when there is a sampling phase. In those cases, each boostrap sample is random under-sampled to balance it;

- **Informed Under-sampling**: These methods address the information loss problem that the standard random under-sampling method have. **EasyEnsemble** creates an ensemble learning system by separately picking a number of subsets from the majority class and creating several classifiers based on the combination of each subset with the data from the minority class. By exploring

the majority class data using independent random sampling with replacement, EasyEnsemble can be viewed as an unsupervised learning algorithm. Instead, in **NearMiss** methods under-sampling is accomplished using the K-nearest neighbour (KNN) classifier. There are five different proposed methods, but in this work only NearMiss-1 and 2 are considered. NearMiss-1 eliminates the majority class examples with the smallest average distance to the minority class examples that are furthest away from them, while NearMiss-2 removes the majority class examples with the smallest average distance to the three minority class examples that are closest to them.

4.7.3 Sampling combination methods

It is also possible to combine over-sampling and under-sampling methods to reach a more balanced distribution. It can be done in any order and multiple times.

- **Random.** Random over-sampling and random under-sampling can be mixed at any extent: the sampling distribution of both tecniques can be tuned.
- **Sampling with Data Cleaning. SMOTE-ENN** combines the strengths of SMOTE over-sampling, which can produce synthetic examples for minority classes, with ENN under-sampling which can eliminate some observations from both classes if they have a different class from its K-nearest neighbour majority class. The K-nearest neighbour of each observation is identified first, and then the ENN technique determines whether or not the majority class from the observation's k-nearest neighbour matches the observation's class. The observation and its K-nearest neighbour are removed from the dataset if the majority class of the observation's K-nearest neighbour and the observation's class vary.

Chapter 5

Results

In the chapter that follows all the experimental results are shown. Initially, the section 5.1 provides an explanation of how the models will be tuned on a stratified sample of the entire dataset. All these experiments, listed in section 5.2, differ for how they tackle the dataset imbalance. Finally, from each tuning setup, the best model will be extracted and will be trained on the entire dataset. In the final section 5.3 all these models are compared to determine the most effective strategy to address the imbalance in the dataset.

5.1 Implementation details

Python has been used throughout to do this task on an Amazon web services virtual machine. Along with the most well-known data science tools, such `pandas` or `sklearn`, libraries that can handle raster data and geographic locations, like `geopandas`, `rasterio`, `xarray`, and `shapely`, have been extensively employed. Additionally, `imblearn` has been used for sampling methods.

Before starting with the model tuning and training, feature scaling and feature selection have been performed using respectively a StandarScaler and RFECV. The features discarded are:

- Standard deviations of three spectral indexes: MSAVI2, NDVI, RENDVI. They are all indexes from greenness families, both broadband and narrowband, which evaluates combined effects of chlorophyll concentration;
- Three weather features: standard deviation for eastward wind speed and both nean and standard deviation for relative humidity.

A stratified selection of 30%, with fixed random state, samples from both the training and test were used for the tuning phase. Therefore, it would be possible to

compare two alternative validation techniques: stratified and annual cross validation. It might be viewed as a technique to validate that the classifier's knowledge is updated annually.

By optimizing F2-score, which is the harmonic mean of recall and precision with a stronger emphasis on recall, `GridSearch` is used to fine-tune the parameters. Precision can be thought of as a measurement of true alarms; in fact, low precision indicates that only a small percentage of positive predictions are accurate, indicating that if the model predicted a fire event, it was probably a false alarm. High precision, in contrast, suggests that the alarm was most likely accurate. Also recall has an important interpretation in this context: it measures how well a classifier can identify a fire incident. Generally speaking, it is more crucial to accurately detect a fire event than to miss it, but precision cannot be neglected because, if we try to maximise recall, the problem might translate in a model that predicts largely fire occurrences with a high number of false alarms. The ideal compromise is indeed F2-score, which takes into account both indicators but places a greater emphasis on accurately forecasting wildfire incidents.

The experiments were done on four different model and sampling approach families. Testing was done on SVM, KNN, Random Forest, Gradient Bosting, Bagging, and XGBoost as

- baseline, so without dealing with the imbalance (Section 5.2.1);
- over-sampling, so the creation of new minority class samples (Section 5.2.2);
- undes-sampling, so removing majority class samples (Section 5.2.3);
- over- and under-sampling combination methods (Section 5.2.4).

The best model for each sampling family was taken, trained on the entire dataset using parameters obtained from yearly cross validation and tested on the full dataset. In the Section 5.3 all the best four models are compared.

5.2 Hyperparameters tuning

In general, determinining the appropriate hyperparameter values for a learning algorithm is challenging. Additionally, in this set of experiments, sampling methods hyperparameter are also tuned to determine the best sampling strategy.

5.2.1 Baseline results

The first set of experiments is done without any technique that deals with dataset unbalance and the results are showed in Table 5.2.1. There is no classifier able to

Model	Precision		Recall		F2-Score		AUROC		AUPR	
	KF	*YF*	*KF*	*YF*	*KF*	*YF*	*KF*	*YF*	*KF*	*YF*
SVM	0.52	0.52	**0.47**	**0.47**	**0.48**	**0.48**	0.68	0.68	0.34	0.34
KNN	0.54	0.54	0.43	0.43	0.44	0.44	0.67	0.67	0.34	0.34
RF	**0.86**	**0.85**	0.23	0.23	0.26	0.26	0.61	0.61	0.34	0.34
BG	0.77	0.77	0.38	0.38	0.42	0.42	0.67	0.67	0.41	0.41
GB	0.82	0.81	0.44	0.42	**0.48**	0.47	**0.71**	**0.70**	**0.46**	**0.45**
XGB	0.75	0.75	0.37	0.37	0.41	0.41	0.67	0.67	0.39	0.39

Table 5.1: Results without using any technique to manage imbalanced dataset. KF and YF are the two different validation methods used, KF = StratifiedKFold, YF = Yearly folds. In the models column is highlighted the best method for this section: Gradient Boosting.

go above 0.5 recall, which means they are no able to correctly predict even 1 of out 2 fires.

The methods with more consistent precision-recall values are SVM and KNN as they generally don't suffer from dataset imbalance, because the algorithms are not affected by the class size. Instead, Bagging and Random Forest both rely on Decision Trees, and each tree is constructed using a fraction of the data, so each tree will be biased in the same direction of class imbalance if the dataset and these samples are unbalanced. In fact, the results had low recall, making it impossible to accurately predict a fire event, but great precision, indicating that the few correct predictions are likely true. If the dataset imbalance, which affects the subsets, is handled, these results might be improved. Probably, Bagging performs slightly better than Random Forest because it has access to all the available feature when selecting where to divide and be, so even if the dataset samples maybe different, the split will consider all the characteristics of the model.

In general, boosting techniques benefit from the fact that they sequentially update mistaken classified samples, so Gradient Boosting is the best method as it has high value precision and a recall of 0.44/0.42 and it has the best AUPR value. Indeed, high area under the curve represents both high recall and high precision.

5.2.2 Over-sampling results

In these section, two different oversampling approaches are used: random over-sampling and synthetic sampling generation. Every method is an improvement of the previous one: random over-sampling tends to overfit data as it creates duplicates of existing points, so SMOTE tries to mitigate this problem by generating synthetic samples rather than replicating instances. However, replicating without paying attention to overlapping data could be uninformative, so the last two adaptive

Model	Precision		Recall		F2-Score		AUROC		AUPR	
	KF	*YF*	*KF*	*YF*	*KF*	*YF*	*KF*	*YF*	*KF*	*YF*
Random Oversampling										
SVM	0.47	0.50	0.55	0.59	0.53	0.57	0.70	0.72	0.34	0.37
KNN	0.44	0.44	0.58	0.62	0.55	0.58	0.70	0.72	0.33	0.35
RF	0.67	0.63	0.46	0.52	0.49	0.54	0.70	0.73	0.41	0.42
BG	0.74	**0.74**	0.43	0.44	0.47	0.47	0.70	0.70	0.43	**0.43**
GB	**0.79**	0.60	0.50	0.50	0.54	0.51	0.71	0.71	**0.49**	0.39
XGB	0.78	0.49	0.46	0.45	0.50	0.46	0.71	0.67	0.46	0.32
SMOTE										
SVM	0.47	0.52	0.44	0.48	0.45	0.49	0.66	0.69	0.32	0.35
KNN	0.45	0.44	0.62	0.65	0.58	0.59	**0.72**	0.73	0.35	0.35
RF	0.69	0.65	0.43	0.55	0.47	0.50	0.69	0.71	0.41	0.41
BG	0.61	0.58	0.35	0.36	0.38	0.39	0.65	0.65	0.33	0.33
GB	0.72	0.62	0.48	0.39	0.52	0.42	**0.72**	0.68	0.45	0.36
XGB	0.72	0.60	0.46	0.44	0.50	0.46	0.71	0.68	0.43	0.37
SMOTE Borderline										
SVM	0.47	0.50	0.47	0.50	0.47	0.50	0.37	0.69	0.32	0.34
KNN	0.42	0.40	0.66	0.71	0.58	0.61	**0.72**	0.73	0.34	0.34
RF	0.66	0.60	0.46	0.52	0.48	0.53	0.70	0.72	0.40	0.40
BG	0.56	0.57	0.37	0.37	0.40	0.40	0.62	0.65	0.33	0.33
GB	0.71	0.58	0.49	0.40	0.53	0.43	**0.72**	0.67	0.45	0.35
XGB	0.69	0.39	0.45	0.43	0.48	0.42	0.70	0.63	0.41	0.28
ADASYN										
SVM	0.50	0.50	0.48	0.48	0.48	0.48	0.63	0.68	0.34	0.34
KNN	0.41	0.40	**0.67**	**0.80**	**0.59**	**0.62**	**0.72**	**0.74**	0.33	0.34
RF	0.63	0.61	0.47	0.50	0.50	0.52	0.70	0.71	0.40	0.40
BG	0.57	0.62	0.43	0.41	0.44	0.43	0.67	0.67	0.33	0.36
GB	0.72	0.59	0.49	0.41	0.52	0.43	**0.72**	0.67	0.45	0.35
XGB	0.70	0.54	0.47	0.45	0.50	0.46	0.71	0.68	0.43	0.34

Table 5.2: Results using oversampling techniques. KF and YF are the two different validation methods used, KF = StratifiedKFold, YF = Yearly folds. In the models column is highlighted the best method for this section: Gradient Boosting

methods are considered: SMOTE-Borderline which classify minority classes samples to understand which points should be sampled and ADASYN which creates data consistent with their distributions. The results are shown in Table 5.2.2

All over-sampling methods have a parameter which describes how many new sample should be generated and it needs tuning too. Sampling strategy parameter,

for over-sampling methods, corresponds to the ratio

$$\frac{N_{rm}}{N_M}$$

where N_{rm} and N_M are the number of samples in the minority class after re-sampling and the number of samples in the majority class, respectively. The majority of sampling methods reached values of 0.6/0.7 for most models, meaning that the samples in the minority class needs to be almost doubled.

With random over-sampling KNN and SVM still performs quite similarly to the baseline methods. KNN with synthetic sampling methods improves its recall measures as, the sampling generated leads to have more simila neighbours around the minority class, so it is more easy for KNN to classify. Indeed, the best recall measure of this slot of experiments is reached by KNN with ADASYN technique. This model, however has a low precision which means that increasing the number of minority samples leads KNN to predict the positive class more than necessary and for this reason, it is not considered the best classifier for this section.

Recall measures for SVM instead worsen from baselines probably because it has more difficulties in finding the right support vectors in a noisier dataset. Instead, Random Forest increases its recall but decreases its precision because probably it tends to overfit data that are generated and it is not able to built general tree. Even in this case, boosting methods are the ones with the best balance between precision and recall, fact that is confirmed by the AUPR score.

Considering the four different sampling strategies values from SMOTE to SMOTE Borderline and ADASYN generally increases, which means that the overlapping problem cited above lead to wrong predictions. While the basic implementation of SMOTE does not distinguish between easy and hard samples to be categorised using the nearest neighbours rule, ADASYN focuses on producing samples next to the original samples that are incorrectly identified using a k-Nearest Neighbors classifier, which probably provides more informative synthetic samples.

ADASYN models have the highest F2-Score because they have high recall, but the precision values are really low compared with the over-sampling ones. Because of their better trade-off between recall and precision, the best model is still Gradient Boosting along with a random over-sampling techniques.

5.2.3 Under-sampling results

Another way of adjusting class imbalance is by under-sampling majority class. Generally, under-sampling methods can also help improve run time and storage problems by reducing the number of training data samples when the training data set is huge, but it can discard potentially useful information which could be important for building rule based classifiers.

Model	Precision		Recall		F2-Score		AUROC		AUPR	
	KF	*YF*	*KF*	*YF*	*KF*	*YF*	*KF*	*YF*	*KF*	*YF*
Random Undersampling										
SVM	0.46	0.48	0.61	0.36	0.57	0.38	0.72	0.63	0.36	0.29
KNN	0.44	0.47	0.61	0.62	0.57	0.58	0.71	0.73	0.34	0.36
RF	0.67	0.63	0.48	0.55	0.51	0.56	0.71	0.73	0.42	0.43
BG	0.62	0.60	0.58	0.63	0.59	0.62	0.75	0.76	0.44	0.44
GB	0.65	0.62	0.64	0.52	0.64	0.53	0.78	0.72	0.48	0.41
XGB	0.62	0.59	0.61	0.64	0.61	0.63	0.76	0.76	0.45	0.45
NearMiss										
SVM	0.35	0.27	0.57	0.75	0.51	0.55	0.66	0.64	0.28	0.25
KNN	0.32	0.31	0.64	0.58	0.49	0.50	0.64	0.64	0.26	0.26
RF	0.35	0.32	0.62	0.69	0.54	0.56	0.67	0.67	0.29	0.28
BG	0.34	0.33	**0.73**	**0.79**	0.60	0.61	0.70	0.70	0.30	0.30
GB	0.48	0.33	0.61	0.67	0.58	0.56	0.73	0.68	0.37	0.29
XGB	0.48	0.32	0.59	0.78	0.57	0.61	0.72	0.70	0.36	0.29
EasyEnsembleClassifier										
EEC	0.58	0.53	0.29	0.29	0.32	0.32	0.61	0.61	0.29	0.29
Balanced under-sampling										
RF	**0.75**	**0.75**	0.68	0.68	**0.70**	**0.70**	**0.81**	**0.86**	**0.57**	**0.56**
BG	0.73	0.72	0.50	0.50	0.53	0.53	0.73	0.73	0.46	0.46

Table 5.3: Results using undersampling techniques.KF and YF are the two different validation methods used, KF = StratifiedKFold, YF = Yearly folds. In the models column is highlighted the best method for this section: Balanced Random Forest

Four different under-sampling approaches have been investigated and, even in this case, the sampling strategy parameter has been tuned. Sampling strategy parameter, for under-sampling methods, corresponds to the ratio

$$\frac{N_m}{N_{rM}}$$

where N_m and N_{rM} are the number of samples in the minority class and the number of samples in the majority class after re-sampling, respectively. All models chooses a sampling strategy of 0.4/0.5. In general, the results from the random under-sampling approach are better than the ones for the random over-sampling in the previous section, meaning that the majority class features make the classification more difficult for the model.

The Near-Miss approach incorporates various sorts of criteria that can be chosen

using the parameter version, which have been tuned along with the sampling strategy. While NearMiss-2 selects the majority samples whose average distance to the N furthest minority class samples is the smallest, NearMiss-1 chooses the majority samples whose average difference to the N nearest minority class samples is the smallest. Version 1 has always been chosen over version 2 in the tuning phase because possibly because it performs a better job in cleaning the overlapping positive and negative instances. This approach with the Bagging model reached the highest value of recall, but with really low values of precision, the lowest precision values among all experiments. Indeed, models start predicting many true labels as they encounter more difficulties in understanding the characteristics linked to the majority class.

Another approach analyzed is the Easy Ensemble Classifier, which in an ensemble method based on AdaBoost learners trained on different samples balanced using random under-sampling. In AdaBoost, data that are hard to categorise are given increasingly bigger weights until the algorithm finds a model that classify these samples correctly. As a result, each iteration of the algorithm must learn a different element of the data, concentrating on regions that include samples that are challenging to categorise. Probably this technique is not able to apply correct weights on data as the the random under-sampling may remove data with useful information.

On Random Forest and Bagging Classifier, the random under-sampling can be applied in two ways: on the initial dataset, so before the creation of the bootstrapped sub-datasets or on these subsets, after their creation usinf the entire dataset. These final methods are called Balanced Bagging and Balanced Random Forest. The Balanced Random Forest is the best model in this category since it has the best values in all metrics except recall. It has higher precision than the plain Random Forest as it still can potentially see all possible samples from the original dataset.

5.2.4 Combination of sampling techniques results

Over-sampling and under-sampling methods could be combined in order to reduce the negative effects of both methods: so a combined approach could lead to less over-fitting and less loss of informative data.

For the random approach two different sampling parameters have been tuned: the under-sampling sampling strategy was always higher than the over-sampling one, with the first value of 0.6 and the second ov 0.4 on average. This results confirm the tendency of the model to prefer under-sampling rather than augmenting the dataset with new minority class samples.

However the best values for recall and F2-score were reached by the SMOTE-ENN approach. It generates new minority data using SMOTE and delete majority

Model	Precision		Recall		F2-Score		AUROC		AUPR	
	KF	*YF*	*KF*	*YF*	*KF*	*YF*	*KF*	*YF*	*KF*	*YF*
Random Oversampling/Undersampling										
SVM	0.45	0.49	0.60	0.62	0.56	0.59	0.71	0.73	0.35	0.37
KNN	0.43	0.47	0.59	0.59	0.55	0.56	0.70	0.72	0.33	0.42
RF	0.67	0.68	0.47	0.47	0.50	0.50	0.71	0.71	0.42	0.42
BG	0.67	0.70	0.49	0.52	0.52	0.55	0.71	0.73	0.43	0.45
GB	**0.73**	0.64	0.53	0.44	0.56	0.47	**0.74**	0.69	**0.48**	0.39
XGB	0.71	**0.72**	0.52	0.53	0.55	0.56	0.73	0.74	0.46	**0.47**
SMOTE-ENN										
SVM	0.49	0.51	0.57	0.57	0.55	0.56	0.71	0.72	0.36	0.37
KNN	0.42	0.43	**0.70**	**0.70**	**0.62**	**0.62**	**0.74**	**0.74**	0.36	0.36
RF	0.57	0.60	0.55	0.55	0.56	0.56	0.73	0.73	0.41	0.41
BG	0.49	0.49	0.47	0.47	0.48	0.48	0.68	0.68	0.33	0.33
GB	0.60	0.60	0.56	0.55	0.57	0.56	**0.74**	0.73	0.42	0.42
XGB	0.60	0.59	0.54	0.53	0.55	0.54	0.72	0.72	0.41	0.40

Table 5.4: Results using a combination of oversampling and undersampling techniques. KF and YF are the two different validation methods used, KF = StratifiedKFold, YF = Yearly folds. In the models column is highlighted the best method for this section: XGBoost.

data. When the the observed sampled and the majority class of its K-nearest neighbours are dissimilar, ENN removes both the observation and its K-nearest neighbour rather than simply the observation and its 1-nearest neighbour. It performs an in-depth data cleaning which achieved high recall score with KNN as it has more minority class neighbours, but the precision score is too low to be considered a good model, probably the data cleaning is too strong. So, because its better trade-off between precision and recall, XGBoost is chosen as the best model of this section.

5.3 Best models comparison

In this section all the best models selected in the previous sections are analyzed. Every model is re-trained over the entire training set from 2016 to 2020 and then tested on 2021.

In Table 5.3 the results achieved are shown. The baseline Gradient Boosting has better performances than the over-sampling one, indeed, even if they have similar F2 Score and recall, the value of precision is different. With oversampling the Gradient Boosting lower its precision from 0.81 to 0.60, meaning that it has a higher

number of false alarms. It can be also seen by their respective precision-recall curves in Figure 5.1 and 5.2: the over-sampling curve (yellow) has a faster decreases than the baseline one, meaning that the precision reduces faster as the recall increases.

Considering the over/undersampling combination, the XGBoost had a better recall but lower precision compared to the baseline. Given that under-sampling strategies consistently performed better than over-sampling ones, this increased recall value could be connected to the under-sampling strategy. Indeed, with an F2-Score of 0.85 and an AUPR of 0.77, the best model is the Balanced Random Forest which uses an under-sampling strategy on the sampled subsets instead of applying it over the entire dataset. This model is able to detect more than eight fires ten and with a precision of around 90%. Additionally, the Figure 5.3 shows this optimal balance between recall and precision is robust.

The excellent interpretability of tree-based models makes easy to analyse the relative importance of each attribute in the model classification. In fact, feature importance may be studied for all of these top tree-base models and all plots can be found in the next pages: red plot in Figure 5.1 for baseline Gradient Boosting, yellow plot in Figure 5.2 for over-sampling Gradient Boosting, green plot in Figure 5.3 for Balanced Random Forest and pink plot in Figure 5.4 for XGBoost.

The only element that all of these models have in common is that the two-weeks prior NBR2 mean value is the most important spectral index feature. The reason of this could be its ability to recognise plant water sensitivity. In the first ten features there are always weather-related attributes with total precipitation and solar radiation typically being the most informative. In several circumstances, solar radiation has been shown to be more significant than temperature. Year and week-specific seasonal factors always have a significant impact on the prediction and even the already-burned feature have shown to be really helpful, particularly for XGBoost, where it had an importance value of around 35%. It is important to emphasize that Balanced Random Forest is able to distribute weights uniformly, and that many of the most crucial features are spectral indexes. For instance, NDMI (water content index) is more significant than solar radiation, and ARI index (antocyanins content) is more significant than wind and temperature.

Approach	**Model**	**Precision**	**Recall**	**F2**	**AUROC**	**AUPR**
Baseline	GB	0.81	0.50	0.54	0.73	0.50
Over-sampling	GB	0.60	0.52	0.53	0.72	0.40
Under-sampling	**BRF**	**0.87**	**0.85**	**0.85**	**0.91**	**0.77**
Comb sampling	XGB	0.73	0.60	0.62	0.77	0.52

Table 5.5: Best models metrics comparison. The best models from the four different setup are trained on the entire dataset and are tested on 2021 samples. The Balanced Random Forest achieved the best results in every metric considered.

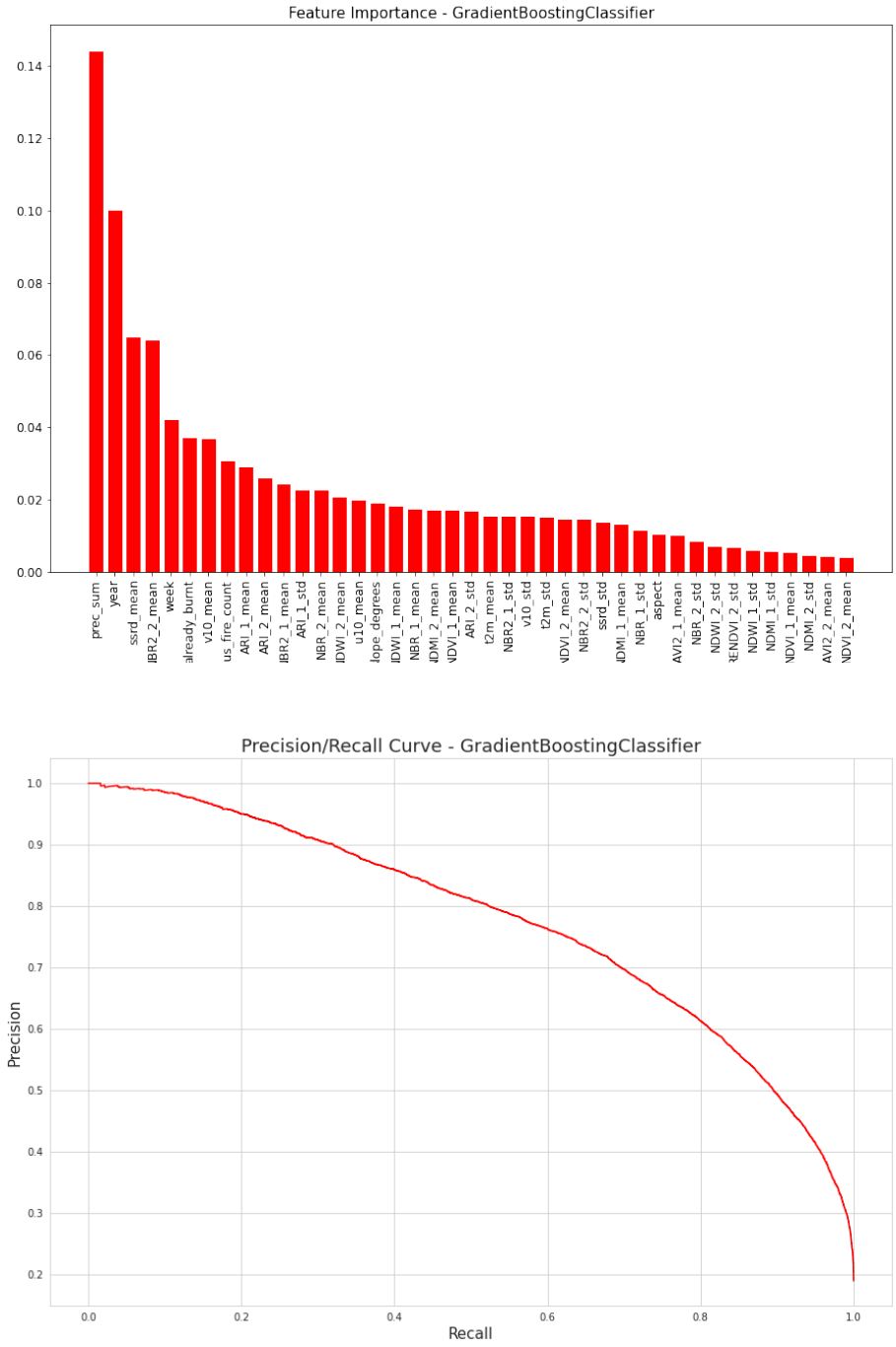

Figure 5.1: Best baseline results. These results are achieved using a Gradient Boosting Classifier with 0.1 learning rate, 350 estimators, each with max depth of 30 and max leaf nodes of 170.

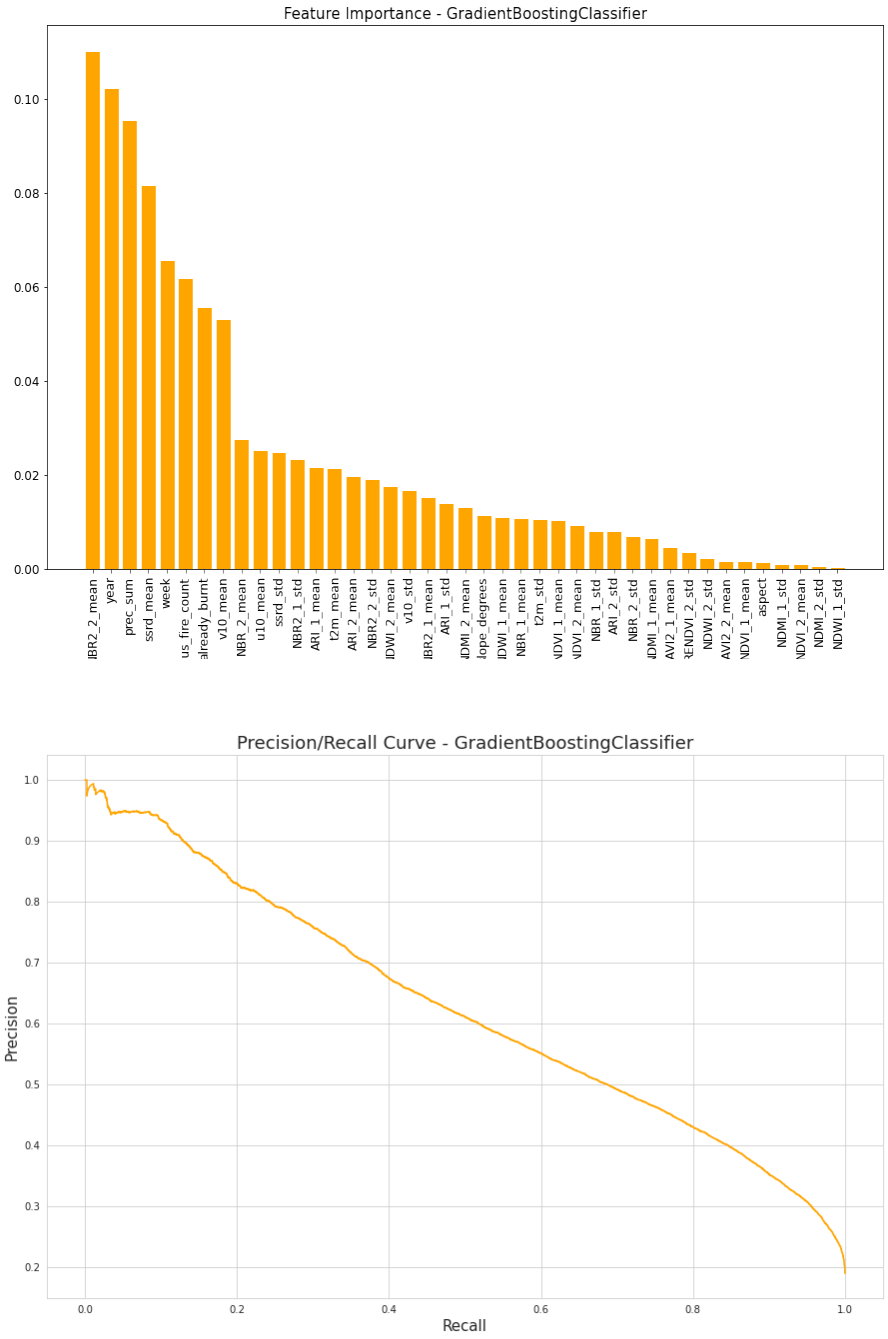

Figure 5.2: Best over-sampling results. These results are achieved using a Gradient Boosting Classifier with 0.01 learning rate, 350 estimators, each with max depth of 30 and max leaf nodes of 120. The over-sampling strategy was 0.7.

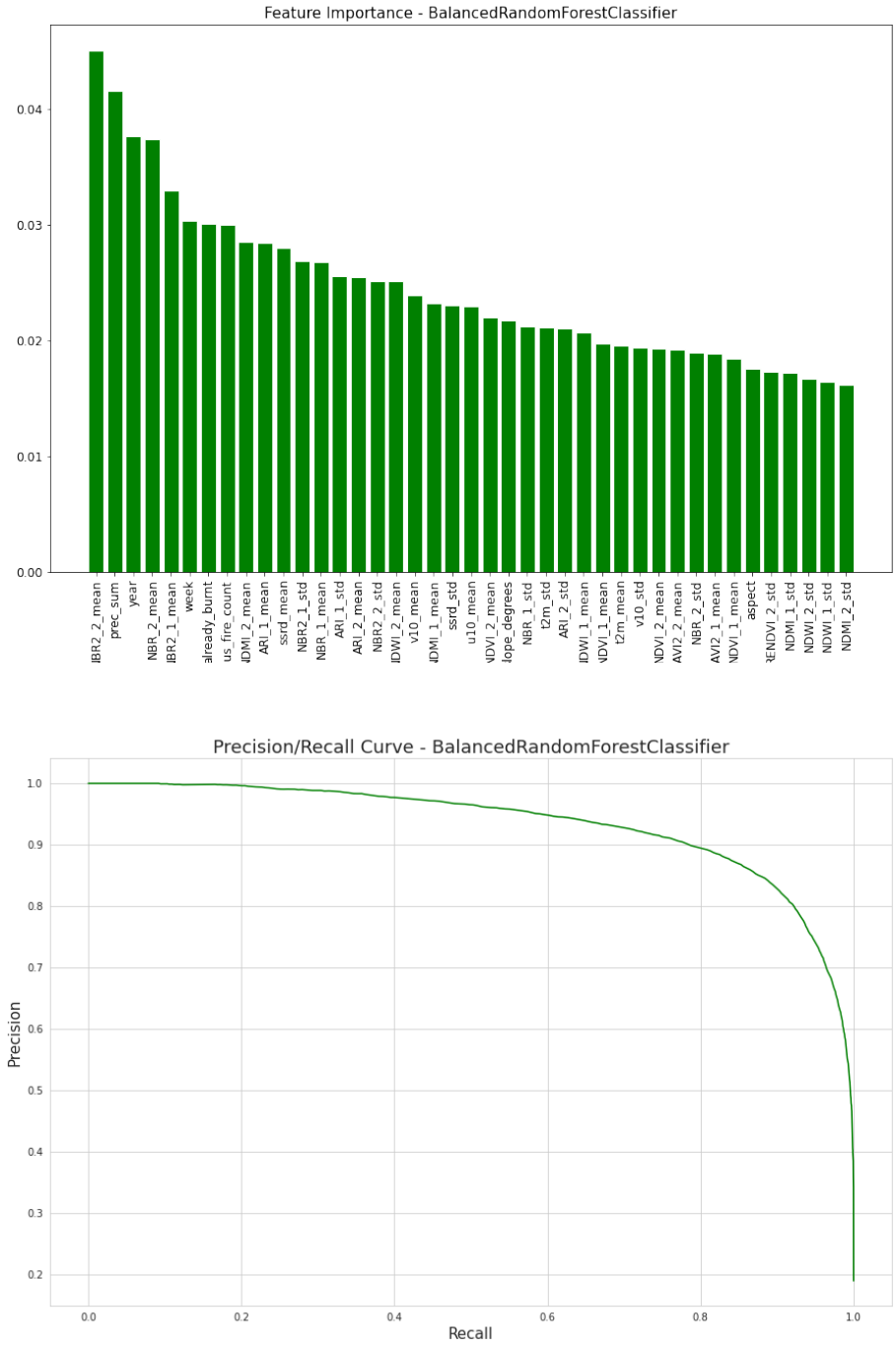

Figure 5.3: Best under-sampling results. These results are achieved using a Gradient Boosting Classifier with gini as split criterion and 450 estimators. The under-sampling strategy was 0.6.

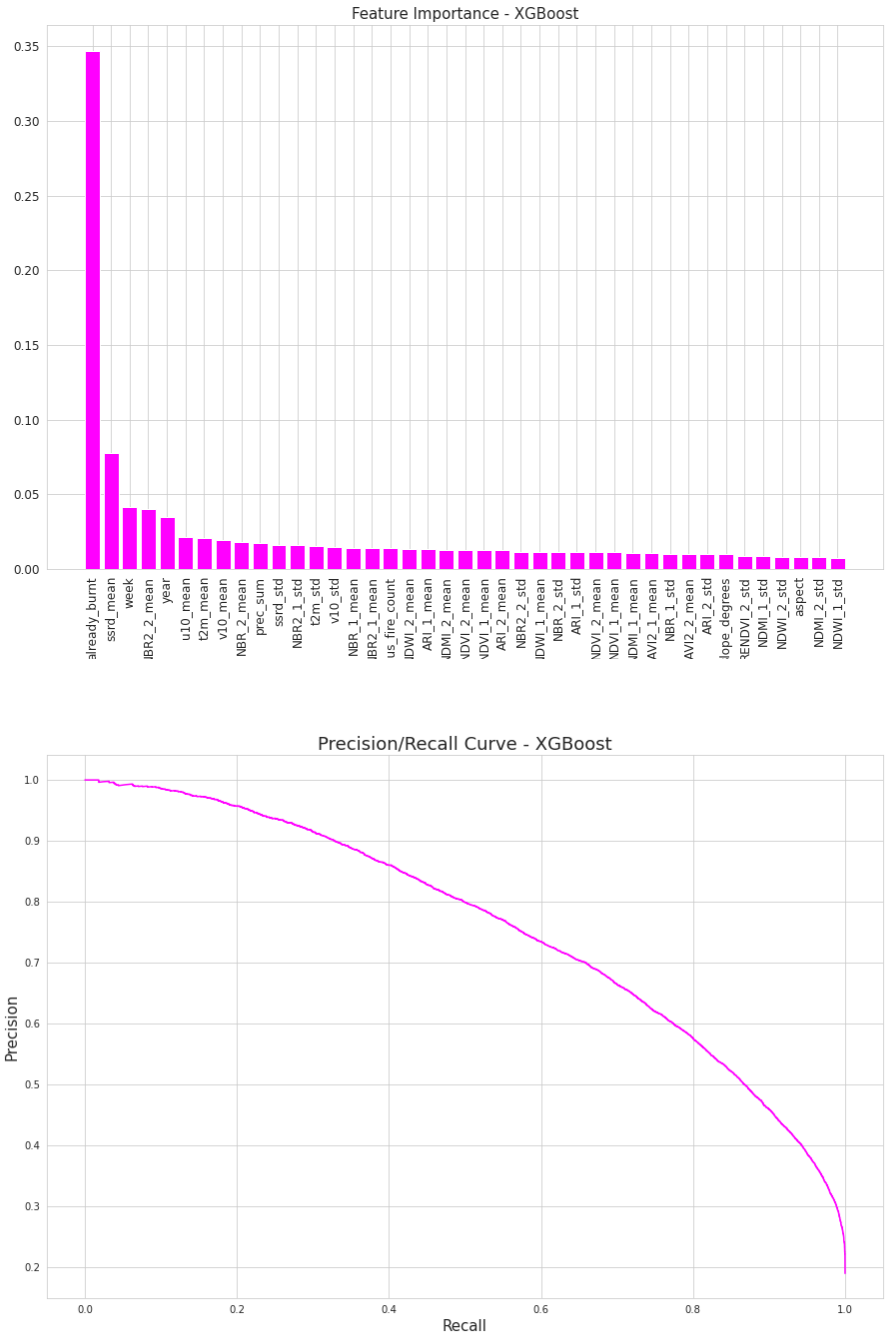

Figure 5.4: Best sampling combination results. These results are achieved using a XGBoost with 0.1 learning rate, 350 estimators, each with max depth of 15 and max leaf nodes of 170. The over-sampling strategy was 0.4 while the under-sampling strategy was 0.6.

Chapter 6

Conclusions

By creating a novel dataset using fire data provided by Comando del Corpo Forestale della Regione Siciliana, this thesis addressed the problem of forecasting a weekly wildfire hazard in Sicily. In fact, Sicily is the Italian region most frequently affected by fires, thus it is necessary to have an automated system for monitoring to have extensive control over areas with a higher fire risk.

The dataset is based on a 200x200 m2 square grid that has been built over a southern region that was selected using the Sentinel-2 tiling system. A single instance of the dataset contains five different types of data: past data using fuel data from the previous two weeks, future data using weather predictions, area fire regime data using the count of prior fires, seasonality data using year and week attributes, and topographical data for the area under consideration.

Due to the nature of the event under consideration, it was important to apply measurements and sampling strategies that could help the training of the models dealing with the imbalanced target distribution. Both ensemble approaches and common classification models were used, along with strategies that addressed class imbalance. Precision, recall, F2-score, AUROC, and AUPR were employed as the five measures to evaluate the performance of the models. The model that performed the best across all five metrics was Balanced Random Forest. Additionally it was noted how it was the model which made the most intensive use of spectral indexes.

6.1 Limitations

Even if this work has achieved encouraging results, there are two main limitations:

- Areas of 200x200 m^2 are too large for a real-world scenario as it would be extremely challenging to monitor the entire area if the fire danger was really high. Indeed, a reasonable data area, for instance, should be at least 60x60

m^2, similar to the highest spatial resolution of Sentinel-2 bands, although more processing power is required to perform a similar task;

- Spectral indexes times series have missing values which have not been filled so some instances may have outdated fuel trends. Two practical ways to overcome this problem is by examining the seasonality of each index or augmenting the dataset using Landasat-8 multi-spectral images processed with ESA's Sen2Like tool. Another more experimental solution would be try computing spectral indexes from Sentinel-3 bands that are similar to Sentinel-2 ones. This idea it needs further testing to understand if it is a practicable.

6.2 Next steps and further works

The following steps can be taken to enable real-world applications for this work:

- The first step is to use actual forecasting meteorological data and the the model against the intrinsic amount of uncertainty and understand how it can affect the performances;
- Another test should be done with respect to the robustness of the framework in a scenario with different land cover characteristics and understand at what extent it can be scaled over the entire country;
- Finally an automated process for the data acquisition and processing should be built in order to update the dataset real-time.

Along with these more concrete steps, some more theoretical investigation could be done on the methodologies used in this work. For example, other approaches to deal with the target imbalance could be tested or maybe, the problem could be treated as an anomaly detection task.

Further improvements could be made to spectral images data in two ways: either by increasing the number of bands, so switching from multi-spectral to hyper-spectral or by enhancing the spatial resolution of the bands. Additionally, this dataset might benefit from SAR data from the Sentinel-1 mission. Since it is an active remote sensing technique, radar data can be obtained at any time of day and they have proven to be quite helpful in mapping drought [61].

In conclusion, even with its limitations, this thesis proved that spectral indexes are able to provide a good representation of the fuel characteristics and that good wildfire risk predictions could be achieved through a dataset that provides information about fire regime, seasonality, weather forecasting, topographical characteristics and fuel trends.

List of Figures

List of Tables

Bibliography

[1] San-Miguel-Ayanz J et al. «Advance report on wildfires in Europe, Middle East and North Africa 2021». In: KJ-NA-31028-EN-N (online) (2022). ISSN: 1831-9424 (online). DOI: 10.2760/039729(online) (cit. on p. 1).

[2] Antonio Nicoletti Enrico Fontana Antonino Morabito. «Italia in fumo: Gli incendi del patrimonio naturale, i fattori di rischio e le proposte di Legambiente». In: (2022). URL: https://www.legambiente.it/wp-content/uploads/2021/11/report-incendi-2022.pdf (cit. on pp. 2, 3).

[3] Grazia Pellizzaro et al. «Impact of Four Large Fires on Air Quality in Sardinia (Italy)». In: *Environmental Sciences Proceedings* 17.1 (2022). ISSN: 2673-4931. DOI: 10.3390/environsciproc2022017093. URL: https://www.mdpi.com/2673-4931/17/1/93 (cit. on p. 3).

[4] Bachisio Arca et al. «Evaluating Wildfire Simulators Based on the 2021 Large Fires Occurring in Sardinia». In: *Environmental Sciences Proceedings* 17.1 (2022). ISSN: 2673-4931. DOI: 10.3390/environsciproc2022017074. URL: https://www.mdpi.com/2673-4931/17/1/74 (cit. on p. 3).

[5] Gargiulo Dell'Aglio Gambardella. «Fires Monitoring by Means of Spectral Indices from Sentinel-2 Data: the Case Study of the Vesuvius and Lattari Mountains, Campania (Italy)». In: (2020). DOI: 10.20944/preprints202009.0069.v1 (cit. on p. 3).

[6] Antonio Pepe, Matteo Sali, Mirco Boschetti, and Daniela Stroppiana. «Mapping Burned Areas from Sentinel-1 and Sentinel-2 Data». In: *Environmental Sciences Proceedings* 17.1 (2022). ISSN: 2673-4931. DOI: 10.3390/environsciproc2022017062. URL: https://www.mdpi.com/2673-4931/17/1/62 (cit. on p. 3).

[7] Jen Ciarochi. *The history of wildfire modelling*. URL: https://triplebyte.com/blog/the-history-of-wildfire-modeling (cit. on p. 6).

[8] Andrew L. Sullivan. «Wildland surface fire spread modelling, 1990 - 2007. 2: Empirical and quasi-empirical models». In: *International Journal of Wildland Fire* 18.4 (2009), p. 369. DOI: 10.1071/wf06142. URL: https://doi.org/10.48550/arXiv.0706.4128 (cit. on pp. 6, 7).

[9] Terry Clark, Mary Jenkins, Janice Coen, and David Packham. «A Coupled AtmosphereFire Model: Convective Feedback on Fire-Line Dynamics». In: *Journal of Applied Meteorology - J APPL METEOROL* 35 (June 1996), pp. 875–901. DOI: 10.1175/1520-0450(1996)035<0875:ACAMCF>2.0.CO;2 (cit. on p. 6).

[10] E Pastor, L Zárate, E Planas, and J Arnaldos. «Mathematical models and calculation systems for the study of wildland fire behaviour». In: *Progress in Energy and Combustion Science* 29.2 (2003), pp. 139–153. ISSN: 0360-1285. DOI: https://doi.org/10.1016/S0360-1285(03)00017-0. URL: https://www.sciencedirect.com/science/article/pii/S0360128503000170 (cit. on p. 7).

[11] Andrew L. Sullivan. «Wildland surface fire spread modelling, 1990 - 2007. 1: Physical and quasi-physical models». In: *International Journal of Wildland Fire* 18.4 (2009), p. 349. DOI: 10.1071/wf06143. URL: https://doi.org/10.48550/arXiv.0706.3074 (cit. on pp. 7, 10).

[12] Andrew L. Sullivan. «Wildland surface fire spread modelling, 1990 - 2007. 3: Simulation and mathematical analogue models». In: *International Journal of Wildland Fire* 18.4 (2009), p. 387. DOI: 10.1071/wf06144. URL: https://doi.org/10.48550/arXiv.0706.4130 (cit. on p. 7).

[13] W. L. Fons. «Analysis of Fire Spread in Light Forest Fuels». In: *Journal of Agricultural Research* 72.3 (1946), pp. 92–121. URL: https://www.fs.usda.gov/treesearch/pubs/58122 (cit. on p. 7).

[14] McArthur AG. «Weather and grassland fire behaviour». In: *Forest Research Institute, Forest and Timber Bureau of Australia* (1966). URL: https://nla.gov.au/nla.cat-vn752731 (cit. on pp. 7, 15).

[15] C. E. Van Wagner. «Conditions for the start and spread of crown fire». In: *Canadian Journal of Forest Research* (1967). DOI: https://doi.org/10.1139/x77-004 (cit. on p. 7).

[16] Richard C. Rothermel. «A mathematical model for predicting fire spread in wildland fuels». In: *USDA Forest Service* (1972). URL: https://www.fs.usda.gov/treesearch/pubs/32533 (cit. on pp. 7, 15).

[17] Richard C. Rothermel. «Predicting behaviour and size of crown fires in the northern Rocky Mountains.» In: *USDA Forest Service* (1991). URL: https://www.fs.usda.gov/treesearch/pubs/26696 (cit. on p. 7).

[18] Grishin AM. «A mathematical model of forest fires and new methods of fighting them.» In: *Publishing House of the Tomsk State University* (1997) (cit. on p. 7).

[19] R R Linn. «A transport model for prediction of wildfire behavior». In: (1997). DOI: 10.2172/505313. URL: https://www.osti.gov/biblio/505313 (cit. on p. 7).

[20] Moreno FG. Tarifa CS Del Notario PP. «On the flight paths and lifetimes of burning particles of wood.» In: (1965) (cit. on p. 7).

[21] Albini FA. «Spot fire distance from burning trees: a predictive model.» In: *USDA Forest Service* (1979) (cit. on p. 7).

[22] William H. Frandsen. «The influence of moisture and mineral soil on the combustion limits of smoldering forest duff.» In: *Canadian Journal of Forest Research* (1987) (cit. on p. 7).

[23] *Fire behaviour*. URL: https://www.nrcan.gc.ca/our-natural-resources/forests/wildland-fires-insects-disturbances/forest-fires/fire-behaviour/13145 (cit. on p. 8).

[24] Clive M. Countryman. «Heat: Heat conduction and wildland fire». In: (1976) (cit. on p. 10).

[25] Nikolay Viktorovich Baranovskiy and Viktoriya Andreevna Kirienko. «Forest Fuel Drying, Pyrolysis and Ignition Processes during Forest Fire: A Review». In: *Processes* 10.1 (2022). ISSN: 2227-9717. DOI: 10.3390/pr10010089. URL: https://www.mdpi.com/2227-9717/10/1/89 (cit. on p. 11).

[26] Carmine Maffei. «Remote sensing-based prediction of forest fire characteristics». In: (2022). URL: https://doi.org/10.4233/uuid:8938bc7b-27e7-4b72-b744-1d8a1b0928a5 (cit. on pp. 13, 14).

[27] *United Nations: Risks and Disasters*. URL: https://www.un-spider.org/risks-and-disasters (cit. on p. 14).

[28] *Forest Fire Danger Index*. URL: https://research.csiro.au/bushfire/assessing-bushfire-hazards/hazard-identification/fire-danger-index/ (cit. on p. 15).

[29] *Fire Danger Forecast*. URL: https://effis.jrc.ec.europa.eu/about-effis/technical-background/fire-danger-forecast (cit. on p. 15).

[30] *History of Remote Sensing*. URL: http://gsp.humboldt.edu/olm/Courses/GSP_216/online/lesson1/history.html (cit. on p. 17).

[31] *Copernicus Programme*. URL: https://en.wikipedia.org/wiki/Copernicus_Programme (cit. on p. 17).

[32] *List of Earth observation satellites*. URL: https://en.wikipedia.org/wiki/List_of_Earth_observation_satellites (cit. on p. 18).

[33] *What is Remote Sensing?* URL: https://www.earthdata.nasa.gov/learn/backgrounders/remote-sensing (cit. on p. 23).

[34] Marina D'Este, Mario Elia, Vincenzo Giannico, Giuseppina Spano, Raffaele Lafortezza, and Giovanni Sanesi. «Machine Learning Techniques for Fine Dead Fuel Load Estimation Using Multi-Source Remote Sensing Data». In: *Remote Sensing* 13.9 (2021). ISSN: 2072-4292. DOI: 10.3390/rs13091658. URL: https://www.mdpi.com/2072-4292/13/9/1658 (cit. on pp. 24, 37).

[35] Elena Aragoneses and Emilio Chuvieco. «Generation and Mapping of Fuel Types for Fire Risk Assessment». In: *Fire* 4.3 (2021). ISSN: 2571-6255. DOI: 10.3390/fire4030059. URL: https://www.mdpi.com/2571-6255/4/3/59 (cit. on p. 24).

[36] Johannes Heisig, Edward Olson, and Edzer Pebesma. «Predicting Wildfire Fuels and Hazard in a Central European Temperate Forest Using Active and Passive Remote Sensing». In: *Fire* 5.1 (2022). ISSN: 2571-6255. DOI: 10.3390/fire5010029. URL: https://www.mdpi.com/2571-6255/5/1/29 (cit. on p. 24).

[37] *Active Fire Detection*. URL: https://effis.jrc.ec.europa.eu/about-effis/technical-background/active-fire-detection (cit. on p. 24).

[38] Daniela Smiraglia, Federico Filipponi, Stefania Mandrone, Tornato Antonella, and Andrea Taramelli. «Agreement Index for Burned Area Mapping: Integration of Multiple Spectral Indices Using Sentinel-2 Satellite Images». In: *Remote Sensing* 12 (June 2020), p. 1862. DOI: 10.3390/rs12111862 (cit. on pp. 24, 36).

[39] Hui Zhou, Fu Xu, Jinwei Dong, Zhiqi Yang, Guosong Zhao, Jun Zhai, Yuanwei Qin, and Xiangming Xiao. «Tracking Reforestation in the Loess Plateau, China after the "Grain for Green" Project through Integrating PALSAR and Landsat Imagery». In: *Remote Sensing* 11.22 (2019). ISSN: 2072-4292. DOI: 10.3390/rs11222685. URL: https://www.mdpi.com/2072-4292/11/22/2685 (cit. on p. 24).

[40] Alekseenko N.A. Gizatullin A.T. «Prediction of Wildfires Based on the Spatio-Temporal Variability of Fire Danger Factors». In: (). DOI: https://doi.org/10.24057/2071-9388-2021-139 (cit. on p. 24).

[41] ARPA Sicilia. *Corine Land Cover (CLC) del territorio siciliano al 2012 e al 2018*. URL: https://www.snpambiente.it/wp-content/uploads/2019/02/15-1-19_pdf_Relazione_ARPA_CLC.pdf (cit. on p. 28).

[42] *LUCAS - Land use and land cover survey*. URL: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=LUCAS_-_Land_use_and_land_cover_survey#Defining_land_use.2C_land_cover_and_landscape (cit. on p. 30).

[43] *Information on LUCAS 2022*. URL: https://ec.europa.eu/eurostat/web/lucas/data/primary-data/2022 (cit. on p. 31).

[44] *Product Types*. URL: https://sentinels.copernicus.eu/web/sentinel/user-guides/sentinel-2-msi/product-types (cit. on p. 32).

[45] *Level 2A*. URL: https://sentinels.copernicus.eu/web/sentinel/user-guides/sentinel-2-msi/product-types/level-2a (cit. on p. 33).

[46] *Sen2Cor*. URL: https://step.esa.int/main/snap-supported-plugins/sen2cor/ (cit. on p. 33).

[47] Cristiano Castaldi Nicola Puletti1 Francesco Chianucci. «Use of Sentinel-2 for forest classification in Mediterranean environments». In: *Annals of Silvicultural Research* (2017). DOI: http://dx.doi.org/10.12899/asr-1463 (cit. on pp. 35–37).

[48] Bang Nguyen Tran, Mihai A. Tanase, Lauren T. Bennett, and Cristina Aponte. «Evaluation of Spectral Indices for Assessing Fire Severity in Australian Temperate Forests». In: *Remote Sensing* 10.11 (2018). ISSN: 2072-4292. DOI: 10.3390/rs10111680. URL: https://www.mdpi.com/2072-4292/10/11/1680 (cit. on p. 35).

[49] Samuel Hislop, Simon Jones, Mariela Soto-Berelov, Andrew Skidmore, Andrew Haywood, and Trung H. Nguyen. «Using Landsat Spectral Indices in Time-Series to Assess Wildfire Disturbance and Recovery». In: *Remote Sensing* 10.3 (2018). ISSN: 2072-4292. DOI: 10.3390/rs10030460. URL: https://www.mdpi.com/2072-4292/10/3/460 (cit. on pp. 35–37).

[50] Klemas Hardisky and Smart. «The Influence of Soil Salinity, Growth Form, and Leaf Moisture on the Spectral Radiance of Spartina alterniflora Canopies.» In: *Photogrammetric Engineering and Remote Sensing* (1983). URL: https://www.asprs.org/wp-content/uploads/pers/1983journal/jan/1983_jan_77-83.pdf (cit. on p. 37).

[51] J. Muñoz Sabater. «ERA5-Land hourly data from 1950 to 1980». In: *Copernicus Climate Change Service (C3S) Climate Data Store (CDS). (Accessed on < 04-09-2022 >)* (2021). DOI: 10.24381/cds.e2161bac (cit. on pp. 38, 39).

[52] *The family of ERA5 datasets*. URL: https://confluence.ecmwf.int/pages/viewpage.action?pageId=181127817 (cit. on p. 38).

[53] Ryan M. May, Sean C. Arms, Patrick Marsh, Eric Bruning, John R. Leeman, Kevin Goebbert, Jonathan E. Thielen, Zachary S Bruick, and M. Drew. Camron. *MetPy: A Python Package for Meteorological Data*. Version 1.3.1. Unidata, 2022. DOI: 10.5065/D6WW7G29. URL: https://github.com/Unidata/MetPy (cit. on pp. 38, 48).

[54] ECMWF. «IFS Documentation CY41R2 - Part IV: Physical Processes». In: *IFS Documentation CY41R2*. IFS Documentation 4. ECMWF, 2016. DOI: 10.21957/tr5rv27xu. URL: https://www.ecmwf.int/node/16648 (cit. on p. 39).

[55] Robin Hogan. «Radiation Quantities in the ECMWF model and MARS». In: (). URL: https://www.ecmwf.int/sites/default/files/elibrary/2015/18490-radiation-quantities-ecmwf-model-and-mars.pdf (cit. on p. 39).

[56] *EU-DEM*. URL: https://land.copernicus.eu/imagery-in-situ/eu-dem (cit. on p. 40).

[57] Richard Barnes. *RichDEM: Terrain Analysis Software*. 2016. URL: http://github.com/r-barnes/richdem (cit. on pp. 40, 48).

[58] B.K.P. Horn. «Hill shading and the reflectance map». In: *Proceedings of the IEEE* 69.1 (1981), pp. 14–47. DOI: 10.1109/PROC.1981.11918 (cit. on p. 40).

[59] *Bagging and Boosting*. URL: https://pluralsight2.imgix.net/guides/81232a78-2e99-4ccc-ba8e-8cd873625fdf_2.jpg (cit. on p. 58).

[60] Haibo He and E.A. Garcia. «Learning from Imbalanced Data». In: *Knowledge and Data Engineering, IEEE Transactions on* 21 (Oct. 2009), pp. 1263–1284. DOI: 10.1109/TKDE.2008.239 (cit. on pp. 64, 65).

[61] Maurice Shorachi, Vineet Kumar, and Susan C. Steele-Dunne. «Sentinel-1 SAR Backscatter Response to Agricultural Drought in The Netherlands». In: *Remote Sensing* 14.10 (2022). ISSN: 2072-4292. DOI: 10.3390/rs14102435. URL: https://www.mdpi.com/2072-4292/14/10/2435 (cit. on p. 84).