UNIVERSITY OF WEST ATTICA

FACULTY OF ENGINEERING

DEPARTMENT OF ELECTRICAL & ELECTRONICS ENGINEERING

POSTGRADUATE THESIS

With title

WILDFIRE PREDICTION USING MACHINE LEARNING

Stafylas Demetrios (Register Number: 0016)

Supervisor: Leligou Aikaterini -Eleni

Athens, 16 June 2022

UNIVERSITY OF WEST ATTICA

FACULTY OF ENGINEERING

DEPARTMENT OF ELECTRICAL & ELECTRONICS ENGINEERING

POSTGRADUATE THESIS

With title

WILDFIRE PREDICTION USING MACHINE LEARNING

Stafylas Demetrios (Register Number: 0016)

Supervisor: Leligou Aikaterini -Eleni

three-member examination committee

Leligou Aikaterini -Eleni — Eleni Aikaterini Leligkou — Digitally signed by Eleni Aikaterini Leligkou Date: 2022.07.27 16:35:27 +03'00'

Papadopoulos Periklis — Periklis Papadopoulos — Digitally signed by Periklis Papadopoulos Date: 2022.07.27 17:16:43 +03'00'

Papoutsidakis Michalis — Michail Papoutsidakis — Digitally signed by Michail Papoutsidakis Date: 2022.07.27 19:48:01 +03'00'

Stafylas Demetrios

Electrical Engineer & Computer Engineer (NTUA)

To my daughter Eva

Acknowledgements

The present postgraduate thesis was prepared within the framework of the postgraduate course, "Master of Science in Artificial Intelligence and Deep Learning" of the University of West Attica under the supervision of Professor Leligou Aikaterini-Eleni. So, I would like to express my gratitude to her for the opportunity she gave me to deal with such an interesting subject-matter, which fully corresponds to my scientific interests as well as for freedom of movement and full support throughout the elaboration. Finally, I would like to thank Mr. Papadopoulos Periklis and Mr. Papoutsidakis Michalis who honored me with their participation in the three-member examination committee.

ABSTRACT

The use of supervised Machine Learning algorithms is widespread in the science of fires. The objective of this postgraduate thesis was to conduct three experiments utilizing only weather variables for the region of the Attica basin. More specifically, the prediction of the probability of fire occurrence (binary classification) for 12, 4 and 2 weather variables respectively, was implemented as first experiment, the prediction of the fire scale (multi-class classification: small fire, medium fire, large fire, wildfire) for 12 weather variables as second experiment and the prediction of the size of the burned area of forest fires for 12 and 4 weather variables as third experiment (regression task). Initially, a new dataset named "wildfire" was synthesized that included the prevailing weather conditions during the forest fires occurrences in the Attica basin. Based on this, an attempt was made to conduct the three experiments with the resulting predictions proving to be particularly impressive. The performance of the formed wildfire dataset was compared with the known prior art Montesinho dataset in order to evaluate which of the two functioned best in the application of supervised Machine Learning algorithms.

The comparative results showed that for all 12 weather variables extracted by the wildfire dataset, a tuned Random Forest model (70%) outperformed other classification models regarding prediction accuracy of fire occurrence. In alternative embodiments for the best 4 and 2 selected weather features correspondingly the Extreme Gradient Boosting (XGBoost) prediction model achieved the best accuracy (67.4%) in terms of fire occurrence prediction and the Neural Networks performed marginally better (63.6%) than the Random Forest (63.3%). As for the problem of multi-class classification of fire scale prediction (small fire, medium fire, large fire, wildfire), it demonstrated that the model of the K- nearest neighbors implemented better (50%) than the other prediction models. The findings for forecasting of size of burned area of forest fires turned out that by using all the weather variables the K-nearest neighbors (r^2 score value 70%) outperformed other regression models while for 4 chosen weather features poor outcomes were provided by regression models with only the Linear Regression algorithm to carry out better than others (r^2 score value 2%).

Finally, a comparison was made with the known prior art Montesinho dataset for 4 and 2 selected weather variables for the first experiment, as well as for 4 weather variables for the third experiment. The results showed that the newly created wildfire dataset functioned much better when applying the supervised Machine Learning algorithms.

Key words: machine learning, wildfire, random forest, support vector machine, logistic regression, linear regression, neural network, decision trees, extreme gradient boosting, k - nearest neighbors, fire occurrence, fire scale, burned area.

Contents

List of Figures

List of tables

CHAPTER 1 -- Introduction

1.1Historical data of fires

Forest fires are today the most well-known common problem faced by our forests and the natural environment destroying important ecosystems and areas of social importance every year. Fighting forest fires is a specialized issue that requires special knowledge and planning. The main purpose of this study is to further contribute to the investigation of this important issue. The problem in our country is acute, with occasional extreme disasters and climatic as well as meteorological conditions being a determining factor for both the onset and the evolution of a forest fire. The collection and analysis of fire data becomes imperative in order to have knowledge which can be integrated into the design of fire prevention and suppression but also the restoration of burned areas. All the catastrophes that have been caused by fires in recent years make us painfully realize that the Mediterranean area is intertwined with fires. In this scope of mind, the specific study was realized and in the hope of contributing, at least in part, to alleviating the intense problem of fires.

By looking at the records of officially recorded fires some very useful information can be retrieved regarding the months, days, hours of fire occurrence and prevailing weather conditions. For instance, the total losses of agricultural and forest areas from fires in the Prefecture of Attica amount to 761,428 acres for the period 1983 -2008 [1] with the most destructive fire being recorded on 11/8/1985 with a total of 78,067 burning acres. In particular, the fire broke out in a forest with fully grassy soil and dense tree vegetation in an area at an elevation of 580 m, with strong soil slopes (60-80%). The cause was a malicious arson, but the perpetrator was not identified. The intervention time of the fire brigade was 30 minutes, and it took 4 days and 8 hours to extinguish it. The fire burned 78,067 acres of forest and agricultural land (49,800 acres of forest and 28,267 acres) and a house. On the day of the event the relative humidity was 31% and the temperature was 33 ° C with strong northly winds of 4.1-7.0 BF. The fire developed into a mixed form and was extinguished only by ground means. In the following map is being illustrated the average annual number of fires in the Prefectures of Greece (period 1983-2008).

Additionally, from the distribution of forest fires based on the month of their event, most appeared from July to September. Within this quarter 62% was recorded of the country's fires, which caused the 85% of its burned areas. The average monthly intensity of fires was also maximized during this period, ranging from 299 in September to 639 in July. August was the most fire prone month of the year, as it accounted for 24% of the incidents and 36% of the burned areas of the country. There was also a clear differentiation of the percentages of burned areas caused by fires, with start day on Sunday but also on Saturday. Specifically, Sunday fires were responsible for 20% of the burned areas of the country and Saturday for 17%, while all on the other days the corresponding percentages ranged from 12% to 14%. Thus, fires were caused on Saturday but mainly on Sunday were characterized by great severity (400 acres and 457 acres burned per incident), versus of the remaining days (values less than 343 acres of burnt area per

incident). In addition to the time of the event, it was observed that 51% of the country's forest fires occurred between 12:00 and 16:00 with a significant spate of fire occurrences at 14:00 (12% of the total), smaller numbers in the morning and afternoon, and few at night.

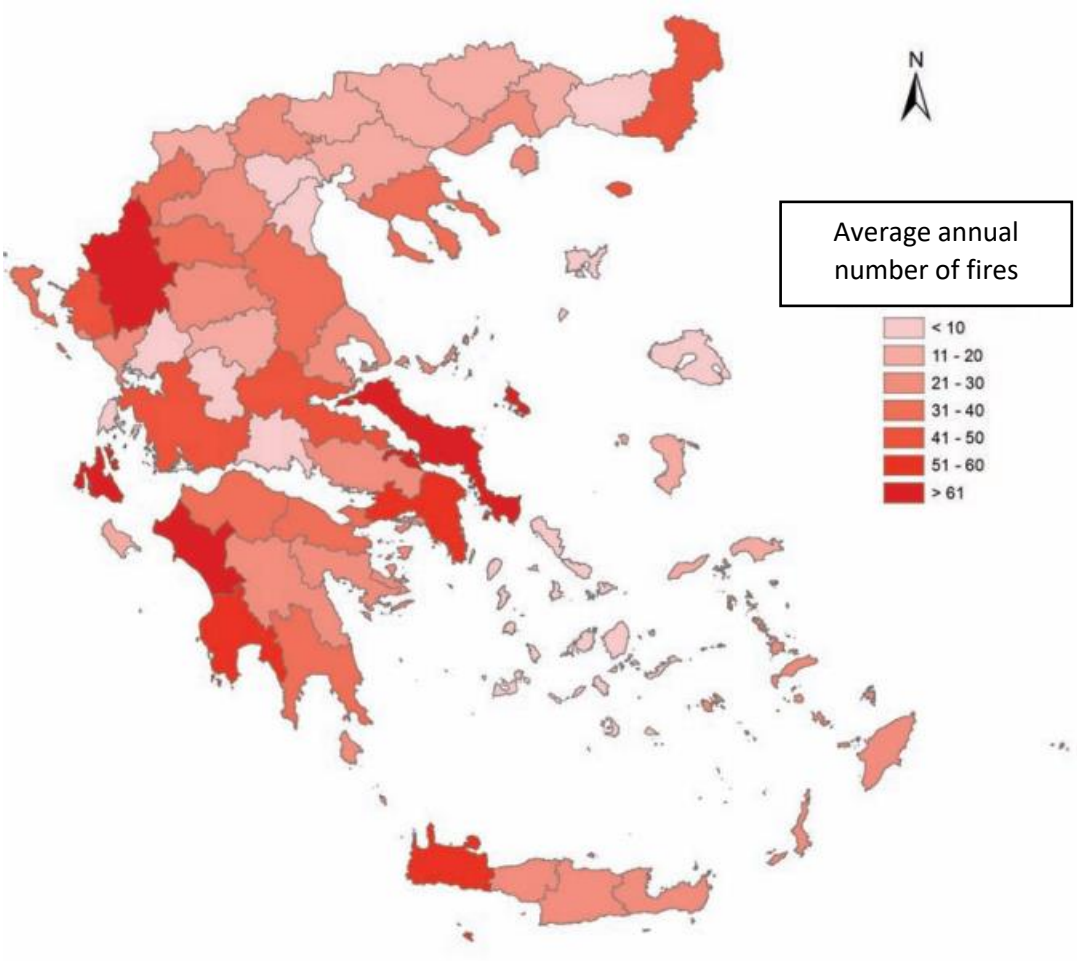


Figure 1 : Map average annual number of fires in the Prefectures of Greece (time period 1983-2008).

Analyzing the causes of forest fires for the period 1983-2008, 47% of the burned areas came from incidents of unknown causes. With the malicious arson following and being responsible for 18% of the burned areas. Although fires by arsonists and psychopaths were recorded as quite rare and accounted for about 1% of the burned areas, they were nevertheless quite severe with an average intensity of 2110 acres of burned area per incident. Less catastrophic were the lightning fires with an average of intensity 876 acres of burned area per incident.

As for the magnitude of the damage caused to the country's forest areas, the losses in human lives but also in animals should not be ignored. The deadly fire of July 23 2018, in Mati

Attica left behind 102 dead, being the second largest tragedy in number of victims of natural disaster related to the weather in our country, after the heat of July 1987. The prevailing meteorological conditions in combination with the topography of the area, made the fire extreme and uncontrollable. On the fateful day, the temperature on the east coast of Attica reached 39 degrees, the humidity dropped to 19% before the fire broke out, while the gusts of wind exceeded 95 kilometers per hour. In essence, the Mati fire broke out under extreme fire-meteorological conditions and exhibited extreme fire behavior, rendering ineffective the effort made to limit and control it.

In recent years, catastrophic fires in Greece continue unabated, with an average of 21,207 hectares of burned forest areas from 2008 to 2020 and with 107,117 hectares for 2021 exceeding the average burned forest areas of the previous 13 years. The following chart shows the total burned area by fires in Mediterranean countries of Southern Europe such as Turkey, Greece, Spain, and Portugal for the period 2008 - 2021.

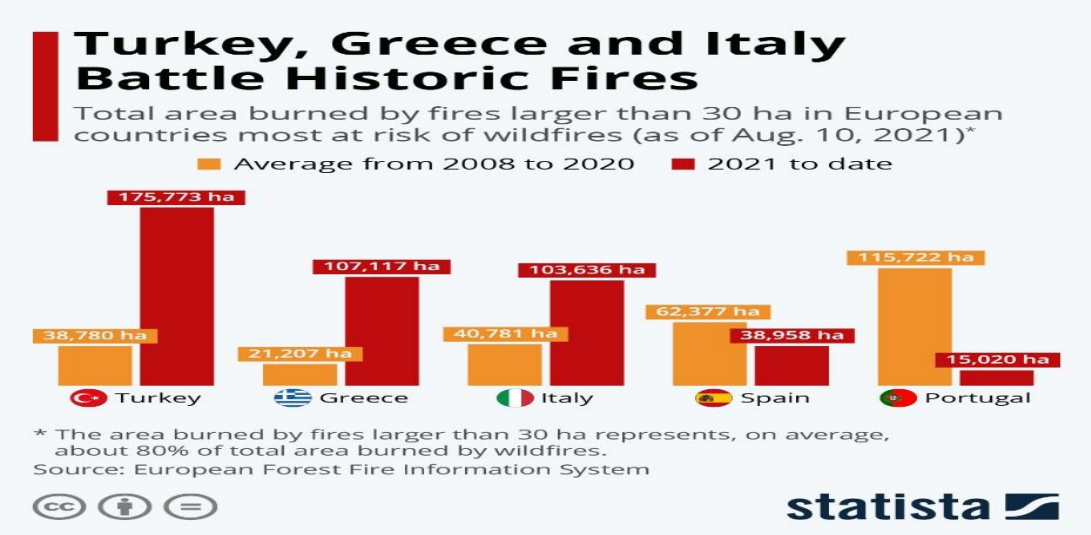


Figure 2: Total burned area in Southern European countries, (time period 2008-2021).
https://www.statista.com/chart/25504/hectares-burned-in-wildfires-in-europe/

1.2The climate of Greece

Greece located in the eastern of Mediterranean basin, belongs to this climatic type, the general characteristics of which is the presence mild and rainy winters, the relatively warm and dry summers and the almost sunshine all year. In the summer months there are few to no rainfall and the dry season can often start as early as April. For the study period 1983-2008 [1], some substantial conclusions were drawn regarding the weather variables (relative humidity, temperature, wind, rainfall) both for the fire occurrences and their intensity.

Relative humidity is a weather factor quite decisive in the occurrence or non-occurrence of fires. In conditions of high relative humidity, the atmosphere becomes wetter with the

consequence that the burnt areas are reduced. When humid conditions prevailed (relative humidity> 80%), a few fires occurred (5.3% of the total) and low intensity (average value 99 acres of burned area per incident), with the result that the total damage corresponded to only 1.5% of the burned area of country, for the period 1983-2008. On the contrary in dryer atmospheric conditions (relative humidity < 40%) fires had increased severity (482 acres burned per incident) for the same period.

Accordingly, the temperature affects the severity of the fires. The most fires (32.1%) occurred at air temperatures from 25 ° C to 30 ° C, with an average intensity of 321 acres of burnt area per incident for the period 1983-2008. At daily temperatures 30-35 ° C the incidents were less (21.8% of the total), but they caused the most disasters (34.3% of all burns areas). However, fires were more severe in extremely hot conditions (> 35 ° C), with average intensity 939 acres of burned area per incident.

There is also a clear correlation between the wind and the severity of forest fires. More specifically, the severity of the fires appears an increasing trend proportionally with the intensity of the wind. The frequency of forest fires in Greece is maximized under moderate wind conditions (1,1-4,0 BF). However, more rarely though, the most catastrophic incidents were associated with stormy winds (> 9.1 BF) and had an average intensity of 2,326 acres area per incident. The most catastrophic fires of the period 1983-2008 were recorded with northerly winds and had an average intensity of 416 acres of burned area per incident. Whereas in conditions of apnea and prevailing east winds, incidents of fires were scarcer.

Another influential factor in the occurrence of fires is the daily rainfall. Annual rainfall ranges from 381 to 1630 mm and is more at higher altitudes in western Greece where these specific areas have more than 100 rainy days per year. The annual rainfall can exceed 2000 mm. On the contrary, the southeastern part of Greece, has frequent annual rainfall less than 400 mm, which is among the lowest in Europe. As a result, this territory usually faces serious drought problems and fire risk [2].

1.3The subject -matter of the study

Forest disasters always cause great human losses. Therefore, the purpose of this work was to develop supervised Machine Learning algorithms for predicting the probability of fire occurrence, the fire scale and the size of the burned forest areas respectively based on daily weather variables. The present study includes the synthesis of a dataset for conducting the above predictions, consisting of the daily prevailing weather variables during the fire occurrences for the period 2010-2019 and for the most fire prone months May to August in Attica basin.

The aims of this study were to compare the results obtained from the applied supervised Machine Learning models in terms of 1) identifying the most influential weather variables caused fire occurrence, the size of burned areas in the study area 2) modelling the probability of fire occurrence (experiment one), fire scale (experiment two) and size of burned area (experiment

three) correspondingly. More particularly, from the 12 total weather variables, the 4 best variables were chosen in the first phase and then the 2 best variables to predict the probability of fire occurrence and the size of the burned forest areas. Then a comparison and selection of the best Classifier / Regressor was performed to determine which provided the best results. Moreover, a comparison with known Montesinho natural park dataset was conducted for 4 and 2 best weather variables respectively in order to draw conclusions which dataset functioned optimally.

Despite the growing needs and interests in fire prevention, there is still much work to be done on this particular problem in the field of Machine Learning. The main contribution of this study is the use of state-of-the-art supervised Machine Learning algorithms for the realization of the before-mentioned predictions using only weather variables. The results proved to be very encouraging as they offered an adequate solution to the problem of fires that has plagued Greece in recent years.

1.4 Study structure

Chapter 2 comprises an exhaustive literature review by defining other works that have been done on the same subject matter. Additionally, it was analyzed the methodology of retrieving the relevant prior art documents. Chapter 3 presents the dataset used to conduct the predictions, as well as the known Montesinho natural park dataset where the comparisons were made. Chapter 4 reveals the models and the supervised Machine Learning algorithms utilized for realizing the mentioned predictions, while the overall experimental results were summarized. Finally, Chapter 5 addresses the conclusions of the work. Possible directions for expansion and optimization are also discussed, as well as future survey.

`

CHAPTER 2 -- Relevant literature

2.1 Literature search

A literature search was performed in ResearchGate and Google Scholars databases by using a combination of words such as "wildfire", "forest fires", "conflagration", "fire weather", “fire occurrence”, “fire prediction/forecast”, “deep learning", "Machine learning", "Artificial Intelligence" for retrieving relevant prior art documents. An extra inquiry was carried out in a cluster of patent databases (European Patent Office QUEry, Derwent Word Patent Index, Non-Patent Literature) to repossess state of the art patents relevant to the fire occurrence prediction. A mix of keywords and technical classification terms was employed for regaining the closest prior art patent documents. An exemplary list of search statements in a cluster of databases (EWN) is being shown as follows.

The classification term is a code system that group the inventions to the technical field, which means that similar inventions are grouped in the same classification. This results in easy search and retrieval of patent documents. Regarding the topic of the study, it is demonstrated the fire occurrence and burned area prediction documents are classified in the following technical fields in Cooperative Patent Classification and Derwent Word Patent Index.

$EWN	SS Status	Results	Query
1		5.360	AND [WILD,]FIRE?, (OR PREDICT+,FORECAST+)
2		107	AND 1,G06N20/00/LOW/C/IC
3		5	AND 2,(OR CLIMATE,WEATHER)
4		5.372	AND (OR [WILD,]FIRE?,CONFLAGRATION,A62C3/0271/C,G08B17/005/C),(OR PREDICT+,FORECAST+)
5		354	AND 4,(OR G06N20/00/LOW/C/IC,G06N3/08/C/IC,G06N3/02/C/IC,G06N3/0454/C,G06N5/003/C)
6		369	AND 4,(OR G06N20/00/LOW/C/IC,G06N3/08/C/IC,G06N3/02/C/IC,G06N3/0454/C,G06N5/003/C,T01-J16C2/MC)
7		19	AND 6,(OR CLIMATE,WEATHER,S03-D05/MC,G01W1/10/C/IC)
8		7	AND 7,(OR HISTORIC+,G08B31/00/C/IC)

Table 1：A list of search statements for retrieving relevant prior art documents.

G06N20/00
Machine learning

- G06N20/10 • using kernel methods, e.g. support vector machines [SVM]
- G06N20/20 • Ensemble learning

G06N3/00
Computing arrangements based on biological models

- **G06N3/02** • using neural network models

_ G06N3/08 • • Learning methods

G06N3/04 • • Architectures, e.g. interconnection topology
G06N3/0454 • • • using a combination of multiple neural nets

A62C3/00
Fire prevention, containment or extinguishing specially adapted for particular objects or places (in oil wells E21B29/08, A62C35/00; in mines or tunnels E21F5/00 ; for nuclear reactors G21C9/04)

_ A62C3/02 • for area conflagrations, e.g. forest fires, subterranean fires
_ **A62C3/0271** • • Detection of area conflagration fires (fire alarms for forest fires G08B17/005)

G08B17/00
Fire alarms; Alarms responsive to explosion

_ G08B17/005 • for forest fires, e.g. detecting fires spread over a large or outdoors area (fire fighting forest fires A62C3/02)

G08B31/00
Predictive alarm systems characterised by extrapolation or other computation using updated historic data

T01-J	Data processing systems
T01-J16	• Artificial intelligence (AI)
T01-J16C	• • Knowledge processing
T01-J16C2	• • • Learning

A total of 23 documents were recovered pertinent about predicting the possibility of fire occurrence, the fire scale, and the size of the burned forest areas.

2.2 Literature review

A detailed scoping review of 300 papers related to the application of Machine Learning algorithms in the science and management of forest wildfires was implemented by Piyush & al [3]. Their attempt based on the identified challenges during wildfire management with the ultimate goal of improving knowledge of Machine Learning models in the specific field. It is widely accepted that both the quality and the quantity of the datasets greatly affect the performance of the Machine Learning algorithms. Therefore, it could not be answered with certainty which was the most appropriate model of Machine Learning as it always depends on the impending problem

of fire management but also mainly on the available datasets. The fire occurrence is due to a combination of factors such as climatic conditions, topography, fuel, ignition source, etc. Focusing our interest on wildfire occurrence and burned area predictions concerning prevailing weather conditions, it was proved through the research that the relative humidity, the accumulated precipitation, the high temperature, the prolonged period of drought, the topography and the meteorological, climatic, and lightning characteristics identified as quite important factors. A remarkable conclusion drawn from the scoping review is that despite the growing number of Machine Learning methods have been applied in various areas of fire science little effort has been devoted to predicting fire occurrence. While they ended up that machine learning algorithms are suitable in fire science and in general problem management only when there is sufficient and high-quality data.

The study of (J. Xiong, J. Wu, Z. Chen) [4] focalized on predicting wildfire size grounded on climatic data using Machine Learning algorithms. The analysis was conducted by processing a Kaggle dataset containing more than 1.8 million fires in the United States. 12 climatic characteristics (different wind speeds at different height, precipitation, temperature measurements, vegetation) and 2 geometric characteristics (longitude and latitude) of the location were retrieved for analysis with the aim of determining the weather conditions at that time of fire occurrence. Various Machine Learning methods were applied such as (Random Forest, Support Vector Machine (Linear), Decision Trees, K nearest neighbors) with the best prediction accuracy being of 32% and achieved with the use Gradient Boosting Trees (GBT) and Deep Neural Networks. The pretty poor accuracy results were attributed to the unpredictable human factor as the main cause of fires as well as the high degree of bias of the data regarding the fire severity and lack of geographical features.

Sakr et al [5] tried to develop a mechanism for predicting the fire occurrence suitable for developing countries that lacked technical infrastructure using only two weather parameters, relative humidity and cumulative precipitation. More specifically, in an effort to reduce costs as well as to eliminate the need for weather forecasting mechanisms and to avoid errors due to inaccurate forecasting, the number of monitored weather features was reduced to two features. A dataset from the territory of Lebanon for the period 2000 - 2008 was utilized and for the season of June to October. Support Vector Machine and Artificial Neural Network were implemented for multiclass fire occurrence prediction with the latter performed marginally better. In case of binary classification (fire/no fire) SVM outperformed over the ANN. In a similar vein, the same group of authors [6] introduced a fire index, having no dependence of any weather prediction mechanism, corresponding to the potential number of fires that could be break out on a specific day. By applying Support Vector Machine (linear) a satisfactory prediction achieved as for the number of fires and their scale whereas for binary classification (fire/ no fire) the accuracy reached up to 96%.

Arias et al [7] focusing their study interest on fire occurrence due to lightning. With a study area the central plateau of the Iberian Peninsula and data sources extracted from the

period 2000-2010 in the specific area for the months of May-September it was turned out that the type of vegetation played a major role in fire prediction. By implementing two Machine Learning algorithms (Logistic Regression, Random Forest) for common five variables (percentage of coniferous forests, percentage of mixed forests and agriculture crops, altitude, slope and mean peak current of negative flashes), Random Forest performed slightly better. What is more the most influential variables regarding the fire occurrence proved to be the percentage of coniferous forests and agricultural crops. Moreover, regarding the topographic variables, the slope of the ground seems to have an effect on the fire occurrence due to lightning, while it was demonstrated that as the altitude increased, the probability of fire decreased. Finally, the mean peak current of negative flashes and the average number of thunderstorms had a significant role in contrast to the polarity of lightning activity.

Researching in the same technical field Blouin et al [8] attempted by combining geographic and temporal variables with weather observations to generate a series of 6-h and 24-hours lightning forecast models for the Alberta province of Canada from April to October of period 1999-2011. In particular, focusing their study on the Boreal and Foothill zones, Balanced Random Forest models were constructed for two time frames (6h, 24h) by processing training data from seven randomly selected years (1999, 2000, 2002, 2004, 2006, 2007, 2009) using as predictors weather data (air temperature, winds, surface pressure, humidity, precipitable water) and geographical and temporal data (latitude, longitude, Julian day, time of day, elevation, convective available potential energy). Predictions were made with the best-fit Random Forest models utilizing validation data form the 6 remaining years (2001,2003,2005,2008,2010,2011). The Showalter Index constituted a crucial predictor for all applicable models. The daily (24h) lightning prediction model for the Foothills zone achieved the best overall performance associated with the five most important variables (Showalter Index, latitude, longitude, elevation, Julian day).

Asley & al [9] evaluated the significance of fire dynamics in the tropical zone in Caribbean and more specifically in Puerto Rico. Climate data (minimum-maximum temperature, precipitation, wind speed), socio-economic data (unemployment rate), historical fire data for the period 2003-2011 were processed and used as inputs in Random Forest Machine Learning algorithm to predict both fire occurrence and the extent of the fire. Daily precipitation was the most important factor in predicting the occurrence and extent of fire. Also crucial were the minimum temperature and the historical fire data regarding the prediction of fire occurrence, while regarding the prediction of the extent of the fire, the maximum temperature and the wind speed contributed significantly. It was also shown that the selection of the unemployment rate did not contribute effectively to the prediction of fire occurrence in contrast to periods of extended drought (especially in winter season) where the probability of fire increased remarkably.

Aldersley et al [10] used a range of ecological, climatic, socio-economic datasets with the aim of the investigating to what extent the relationship between burned areas and climate was

influenced by the human factor both globally and regionally. As data source ten variables were utilized (burned area, tree cover, cropland and pasture cover, population density, Gross Domestic Product, road density, lightning, climatic data) with reference year in 2000 and with the research area to have been delineated to 14 sub-continental regions. Regression Trees and Random Forest were applied for processing multifaceted data. Regarding the global analysis, this model got climatic conditions very seriously. Especially the determination of the average monthly temperature in combination with the percentage of wet days was an important drive in the effect of the average burned area. In addition, the rate of cumulative precipitation and Gross Domestic Product had been shown to suppress fires. The latter was considered as the only measure of human influence. In terms of the regional analysis, it was confirmed that although climate variables played a significant role, temperature was not a significant factor. On the other hand, lightning and wet days had a special effect on the occurrence of fires. Summarizing the present work proved the application of the Random Forest model yielded high predictive power. Moreover, climatic factors were shown to be superior to human factors in the global analysis while in the regional analysis, the high variability in the interaction between environmental and anthropogenic variables was a deterrent to the development of effective predictive models.

On the same wavelength, Guo et al [11] applied Logistic Regression and Random Forest Machine Learning algorithms for determining biophysical and human activity factors as main causes of anthropogenic fire occurrence in boreal forest in China. During processing of predictor variables included climate factors (mean temperature, daily precipitation, relative humidity), forest type, topographic features, human infrastructure (distance to settlement, distance to railway, distance to roadway) and socio-economic factors (unemployment rate, Gross domestic Product, population density) in the applied Logistic Regression and Random Forest models respectively it was demonstrated that Random Forest outperformed. In more detail, quantifying the predictive ability of the LR, RF models by utilizing the Receiver Operating Characteristic and Area Under Curve, it was revealed that the correct prediction rates 60.8% for LR and 70.8% for RF. In addition, in this study, the distance to the railway and the type of forest were identified as the most important factors for the anthropogenic fire occurrence for both models. The distance to settlement and road network were useful information for the RF model. Socio-economic factors, on the other hand, did not seem to have much of an impact on anthropogenic fire prediction models, thus confirming that such fires are more likely to occur near railways, roads and settlements depending on the type of vegetation at a time.

Li-Ming et al [12] built a forest fire prediction modeling method in Japan territory based on artificial intelligence neural networks using population density and weather data (relative humidity, wind speed, daily hours of sunshine). Since most of the influencing factors had non-linear relationships with the risk of forest fire, neural networks emerged as the ideal solution as they have the ability to manage non-linear problems. When first correlating the probability of forest fire with population density and subsequently the probability of forest fire and combinations of population density with weather parameters, it was shown that Back

Propagation Neural Networks better captured this non-linearity of the influencing factors and gave better results in comparison with polynomial regression.

Preeti et al [13] preprocessed a Kaggle data set considering fire meteorological parameters for extracting variables such as temperature, relative humidity and wind speed. By applying regression Machine Learning techniques such as Random Forest (RMSE:0,07), Decision Trees, Support Vector Regression and Artificial neural Network they verified that high temperatures, mild humidity and strong wind speeds are the most essential parameters for predicting the occurrence of fire.

Stojanova et al [14] employed an improved fire forecasting model for the country of Slovenia by processing data from the Geographic Information System, telescopic imagery data and weather data from the ALADIN forecasting model. More specifically, datasets were processed from the continental part, coastal and Kras region of the territory of Slovenia, applying Machine Learning algorithms to predict the fire occurrence. Modeling of the relationships between the threat of fire and the influencing factors (weather conditions, climate data, direction-wind speed) was considered important for the prediction of the possibility of fire occurrence. In the present work, a variety of classifiers were implemented, both single (KNN, Logistic Regression, SVM) and ensemble methods (Boosting, Bagging and Random Forest) in order to evaluate the most appropriate with the best fire prediction performance. The ensemble bagging Decision Trees appeared to outperform other models.

Cortez and Morais [15] proposed a data mining approach using only meteorological data as detected in real time by local meteorological station sensors. A research area was a Montesinho natural park in the Northeast territory of Portugal so as to predict the size of the burned areas of forest fires. The data were collected from two databases, with the first one from the park surveillance area receiving information such as time, date, location, vegetation type and six spatial and temporal components of the Fire Weather Index (fine fuel moisture code, duff moisture code, drought code, initial spread index, build up index, fire weather index) while the second one from the Polytechnic Institute weather station collecting weather observations. For the formed regression dataset, Root Mean Squared Error and Mean Absolute Deviation utilized as global metrics to evaluate the overall performance of the models for four feature selections (combinations of spatial, temporal, Fire Weather Index elements and meteorological variables). Various Machine Learning (Decision Trees, Random Forest, Neural Network, Support Vector Machine) techniques were applied for regression tasks with Support Vector Machine proving to be the best at predicting small fires only for four weather variables (temperature, rain, relative humidity, wind speed). In a similar fashion Xie and Peng [16] explored the ensemble learning methods potential for accurate prediction of both burned area of forest fires and large-scale forest fires for the same study area and dataset. As for the prediction of the burned areas the Random Forest proved to have outperformed other regression models. Regarding the classification models for the prediction of large-scale fires, it was demonstrated that the Extreme

Gradient Boosting and more specifically the Gradient Boosting Regression Tree performed much better than any other model.

Bisquert et al [17] studied for the region of Galicia in Spain the prediction of fire danger using MODIS images (Moderate resolution imaging spectroradiometer) to obtain remote sensing data such as land surface temperature (LST) and enhanced vegetation index (EVI). The input variables that were analyzed to assess the forest fire danger were: 16 day EVI composition, 16 day EVI variations compositions, average and maximum LST compositions of different days and LST variations, difference between punctual values of EVI and LST, a period of year and fire history. Logistic regression was implemented to find the best combination of the above variables (8 day LST, fire history, period of year) for introducing them in an artificial intelligence network in order to predict the fire danger at three levels (low, medium, high danger). It was turned out the artificial intelligence network performed much better than logistic regression with accuracy 76% and precision 66% respectively.

Regarding Greek territory Vasilakos et al [18] presented a fire ignition forecasting system for the island of Lesvos in the northeastern Aegean sea based on meteorological data, vegetation, topographic data, human factor and remote sensing data. A large-scale fire ignition prediction system was developed using neural networks, taking as inputs the Fire Weather Index, Fire Risk Index and Fire Risk Index and giving as an output the Fire Ignition Index.

Concerning the published patents, Tohidi et al [19] disclosed a fire monitoring system and method for estimating the state of fire (fire perimeter, fire intensity, flame height) when taking dynamic characteristics such as satellite imagery, weather variables (wind speed and direction, temperature, humidity, cloud cover) but also static characteristics such as land use, slope, elevation, soil moisture, vegetation, fuel type. The current fire situation could be assessed and modified in real time by receiving constant information while together with a model of predicting fire due to lightning, ember modeling and instability parameters, the rate of fire evolution was predicted. The combination of the fire-related inputs (physical model) together with the Machine Learning algorithms (Logistic Regression, Random Forest, Deep Neural Networks, Support Vector Machine) gave a satisfactory accuracy from 70% to 92% for the prediction of fire occurrence and its evolution. Minglang [20] used a neural network (LSTM) to predict forest fires in the mountains by constructing an automatic feature extraction and combining the collected meteorological data with spatial data. Dan [21] utilized deep learning (Convolutional Neural Network and Recurrent Neural network) for early warning of mountain fires grounded on meteorological and remote sensing data. White et al [22] disclosed a method of a gridded prediction of a wildfire occurrence and extent of fire in a geographical defined area and time period by processing weather, climate, historical and remote sensing data. Li Jinsong et al [23] revealed a forecasting method of a forest fire risk grade for a power transmission line crossing a mountain. A gradient boosting tree was applied by processing historical, meteorological, remote sensing data and vegetation types. Guo

Shuchang et al [24] deployed a lightning forecasting method suitable for forest fire prevention based on convolutional neural network. Watt et al [25] divulged a neural network architecture comprising either a plurality of Recurrent Neural Network elements and at least one Long Short-Term Memory or a transformer element connected in series or a parallel. The purpose of this neural network was to model mainly the climate, weather data and location data.

CHAPTER 3 -- Study area and Datasets

3.1 Attica basin

As a study area was defined the Attica basin with a latitude of 37.98, a longitude of 23.72 . The extent of the capital region is 427 km2 and with an estimated population approximately to 4 million. Geographically the Attica basin is bounded by four large mountains of Aigaleo, Parnitha, Hymettus, Penteli and the Saronic Gulf to the southwest. The Attica region has experienced the last two decades several wildfires resulting in having been burned down important parts of the forest national park of Parnitha, Penteli and Hymettus and causing great side-effects on the fauna and flora. What is more, the air quality of the capital has been affected significantly raising worries about living conditions. Administratively the Attica basin is divided in western, northern, and eastern Attica belonging to Attica prefecture the rest of Piraeus prefecture where they belong by administrative point of view Salamis, Aegina, Hydra, Poros, Spetses, Kythera and Antikythera, and the province Troizinias located in the Peloponnese.

Figure 3: A map of Attica basin.

Athens is the hottest city in mainland Europe with an average temperature of 19.8° C. Its climate is featured by prolonged hot, dry summers and mild winters with moderate rainfall. The months of July and August are characterized as the driest and most dangerous for fires. Furthermore, annual precipitation of Athens is lower than most other parts of Greece. The urban

area of Athens suffers from the phenomenon of urban heat island effect due to human activity. Some of the highest temperatures have also been recorded in the Attica basin with the highest ever recorded in Europe 48° C, on 10/7/1977 in the areas of Elefsina and Tatoi. The extreme climate data for the period 1955-2010 in Athens (Nea Filadelfeia meteorological station) is being depicted in the following table.

Climate data for Elliniko, Athens (1955–2010), Extremes (1961–present) [show]

Climate data for Nea Filadelfia, Athens (1955–2010) [hide]

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Year
Average high °C (°F)	12.6 (54.7)	13.6 (56.5)	16.0 (60.8)	20.3 (68.5)	26.2 (79.2)	31.4 (88.5)	33.8 (92.8)	33.6 (92.5)	29.2 (84.6)	23.5 (74.3)	18.1 (64.6)	14.1 (57.4)	22.7 (72.9)
Daily mean °C (°F)	8.8 (47.8)	9.3 (48.7)	11.3 (52.3)	15.3 (59.5)	21.0 (69.8)	26.0 (78.8)	28.3 (82.9)	27.8 (82.0)	23.4 (74.1)	18.4 (65.1)	13.7 (56.7)	10.2 (50.4)	17.8 (64.0)
Average low °C (°F)	5.4 (41.7)	5.5 (41.9)	6.9 (44.4)	9.9 (49.8)	14.2 (57.6)	18.7 (65.7)	21.3 (70.3)	21.2 (70.2)	17.6 (63.7)	13.8 (56.8)	10.0 (50.0)	6.9 (44.4)	12.6 (54.7)
Average precipitation mm (inches)	53.9 (2.12)	43.0 (1.69)	41.8 (1.65)	28.5 (1.12)	20.5 (0.81)	9.1 (0.36)	7.0 (0.28)	6.7 (0.26)	19.4 (0.76)	48.8 (1.92)	61.9 (2.44)	71.2 (2.80)	411.8 (16.21)
Average precipitation days	12.0	10.6	10.2	8.3	5.8	3.4	1.9	1.6	4.1	7.4	10.1	12.5	87.9
Average relative humidity (%)	74.4	72.0	68.4	61.7	53.4	45.7	42.9	45.4	54.6	66.1	74.5	76.2	61.3

Source: HNMS[81]

Figure 4: Extreme recorded climate data in Attica basin (time period 1955-2010).
https://en.wikipedia.org/wiki/Athens

Given the purpose of the present study were taken the weather conditions of Attica basin into consideration for implementing predictions of the probability of fire occurrence (experiment one), fire scale (experiment two) and the size of the burned area (experiment three). In particular, the dynamic meteorological parameters that analyzed are:

- mean, minimum, maximum temperature (° C)
- mean, minimum, maximum relative humidity (%)
- mean, minimum, maximum atmospheric pressure (hPa)
- daily rainfall (mm)
- mean wind speed (km/h)
- wind gust (km/h)

3.2 Wildfire dataset

For the realization of the present work, the daily weather variables that were retrieved from the National Observatory of Athens Institute of Environmental Research and Sustainable Development were pieced together with the officially recorded forest fire incidents from the files of the fire brigade in the region of the Attica basin. The collected data refer to the period 2010-2019 and for the fire-prone months May to August. The specific dataset is named "wildfire dataset" created in csv form and has the configuration of the following table by enforcing it in the python platform.

1 to 5 of 5 entries Filter

index	Date	mean_temperature	max_temperature	min_temperature	mean_RH	max_RH	min_RH	mean_pressure	min_pressure	max_pressure	daily_rainfall	mean_wind_speed	wind_direction	wind_gust	prefe
0	1/5/2010	20.2	20.3	20.0	45.9	63	31	1015.2	1016.5	1013.8	0.0	3.4	SW	9.6	attica
1	2/5/2010	20.6	20.7	20.4	47.5	65	31	1015.2	1016.6	1014.1	0.0	2.5	S	7.7	attica
2	3/5/2010	20.7	20.8	20.6	55.2	79	34	1013.6	1014.6	1012.5	0.0	2.3	S	6.9	attica
3	4/5/2010	20.6	20.7	20.4	54.0	77	35	1014.0	1014.9	1012.8	0.0	2.0	S	6.5	attica
4	5/5/2010	21.9	22.0	21.7	51.1	68	33	1013.1	1014.4	1010.9	0.0	1.6	WSW	5.9	attica

Show 25 per page

Figure 5 : "Wildfire dataset" including daily weather variables of Attica basin with corresponding officially recorded fire incidents.

In essence, a temporal alignment of the daily weather variables of the Attica basin with the officially recorded cases of forest fires took place. The wildfire dataset consists of 1400 entries and 19 columns namely date, mean-min-max temperature, mean-min-max relative humidity, mean-min-max atmospheric pressure, daily rainfall, mean wind speed, wind direction, wind gust, prefecture, municipality, start time of fire incident, burned area in acres and declaration of fire occurrence. In case of fire occurrence, it was denoted by 1 otherwise 0. In addition, some assumptions were taken during the composition of the mentioned wildfire dataset such as that in cases where we had a fire incident but negligible burned forest areas were considered as non-fire occurrence. Also, the data of the fires in the landfill were not included. Even for recorded incidents of fire at different times for the same area it was considered as the time of event the first reported time and as a burned area the total sum of incidents in acres. Finally, although administratively the islands of Aegina and Salamina belong to the Piraeus prefecture, however, for gathering more data on the fire occurrence, they were included as incidents that generally belong to the Attica prefecture.

By performing data exploration in this specific dataset was easily seen that there was not a null value while the type of data was a mix of floats and integers number except for the date, wind direction, prefecture, and municipality columns respectively which were objects (figure 6). Substantial information was extracted from the display of correlation matrix (figure 7), understanding the dependence among two weather variables and how they moved together. For instance, there was a positive correlation among mean-min-max temperature with mean wind speed and wind gust correspondingly. Similarly, the mean-min-max relative humidity had positive correlation with daily rainfall. The mean-min-max pressure appeared positive correlation with the mean wind speed and wind gust but negative correlation with the mean-min-max temperature and the daily rainfall. The latter means that when pressure increases the temperature or daily rainfall probably goes down. Of particular interest is the control of the variable of daily rainfall as in prolonged periods of drought there is a high risk of fires. Bearing in our mind the conclusions of [5],[6] that two weather variable such as the relative humidity and cumulative precipitation are enough for fire occurrence prediction, a

plot of minimum relative humidity and daily rainfall was formed (figure 8). A plurality of scatter plot graphs can be sought into {appendix code_1, code_6}.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1399 entries, 0 to 1398
Data columns (total 19 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Date              1399 non-null   object
 1   mean_temperature  1399 non-null   float64
 2   max_temperature   1399 non-null   float64
 3   min_temperature   1399 non-null   float64
 4   mean_RH           1399 non-null   float64
 5   max_RH            1399 non-null   int64
 6   min_RH            1399 non-null   int64
 7   mean_pressure     1399 non-null   float64
 8   min_pressure      1399 non-null   float64
 9   max_pressure      1399 non-null   float64
 10  daily_rainfall    1399 non-null   float64
 11  mean_wind_speed   1399 non-null   float64
 12  wind_direction    1399 non-null   object
 13  wind_gust         1399 non-null   float64
 14  prefecture        1399 non-null   object
 15  municipality      677 non-null    object
 16  start_time        1399 non-null   float64
 17  burned_area       1399 non-null   int64
 18  fire_occurrence   1399 non-null   int64
dtypes: float64(11), int64(4), object(4)
memory usage: 207.8+ KB
```

Figure 6: Wildfire dataset -data type

Continuing a further data exploratory analysis distribution graphs (figure 9) provided useful information. The highest mean temperature seemed to have been recorded of 27-29° C for 340 observations, the mean relative humidity of 42% for 310 observations, the mean atmospheric pressure of 1015 hPa for 390 observations and the mean wind speed 5 km/h for over 800 observations. Quite interesting was the daily rainfall graph where the absence of rainfall was recorded for over 1300 observations, an ominous element and indicative factor for fire occurrence incidents, {extra information in appendix code_1, code_6, code_7}.

As for the size of burned area in association with municipalities located in Attica prefecture a thought-provoking graph (figure 10) revealed that the Megara, Oropos and Penteli were considerably fire prone areas.

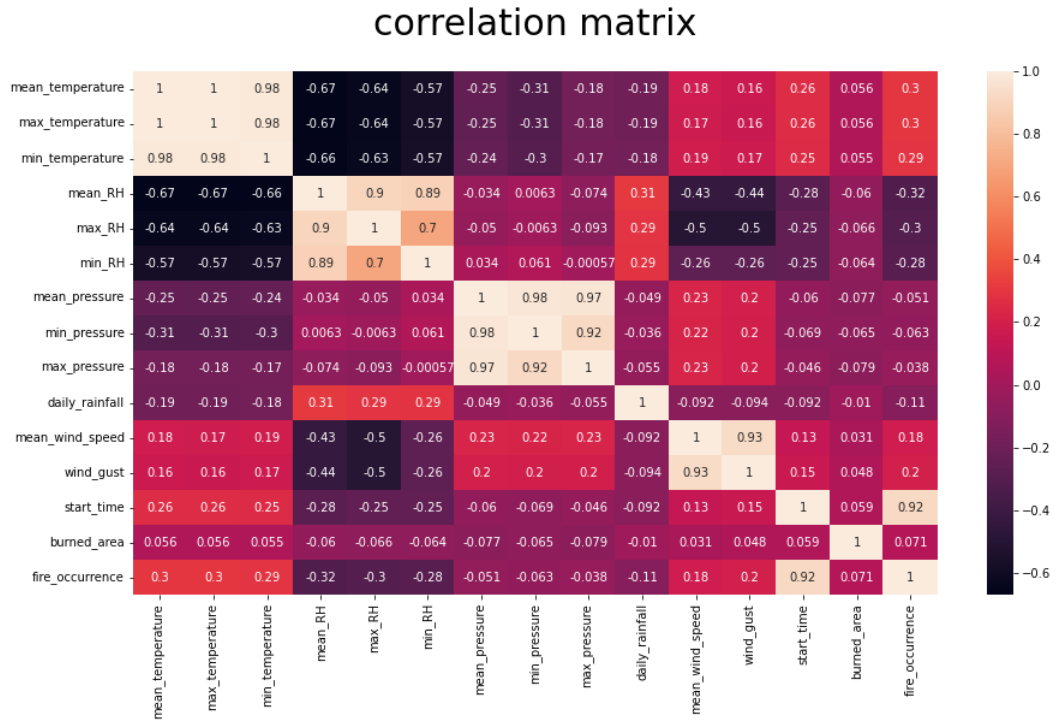


Figure 7 : Wildfire dataset -correlation matrix

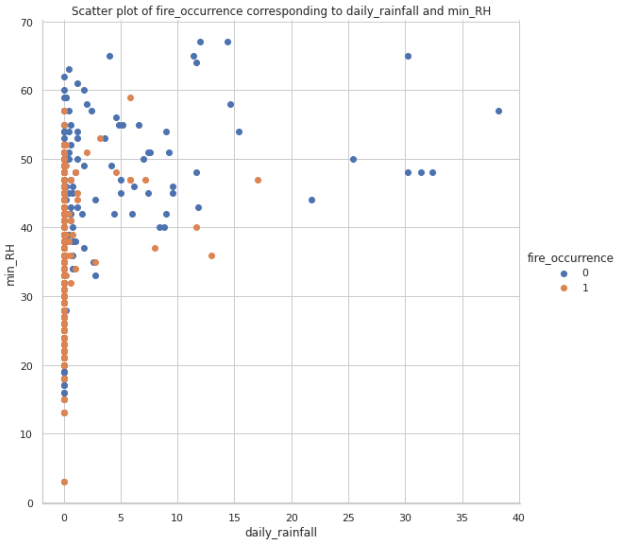


Figure 8: Fire occurrence incidents corresponding to daily rainfall and minimum relative humidity

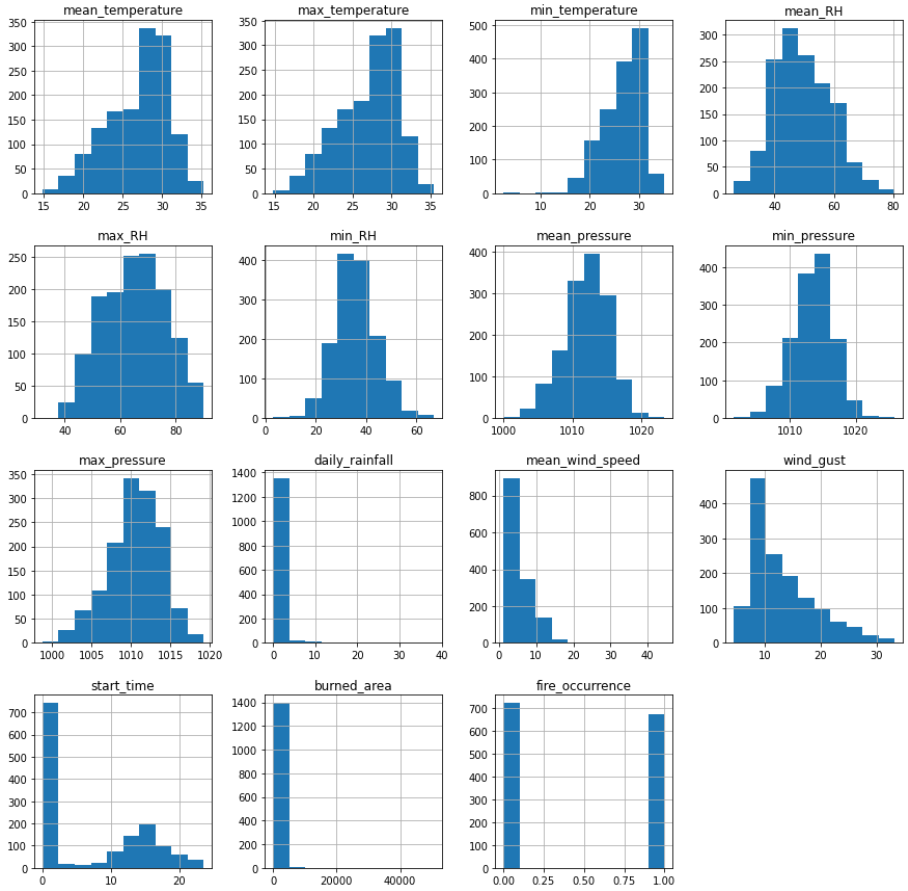


Figure 9 : Wildfire distribution graphs

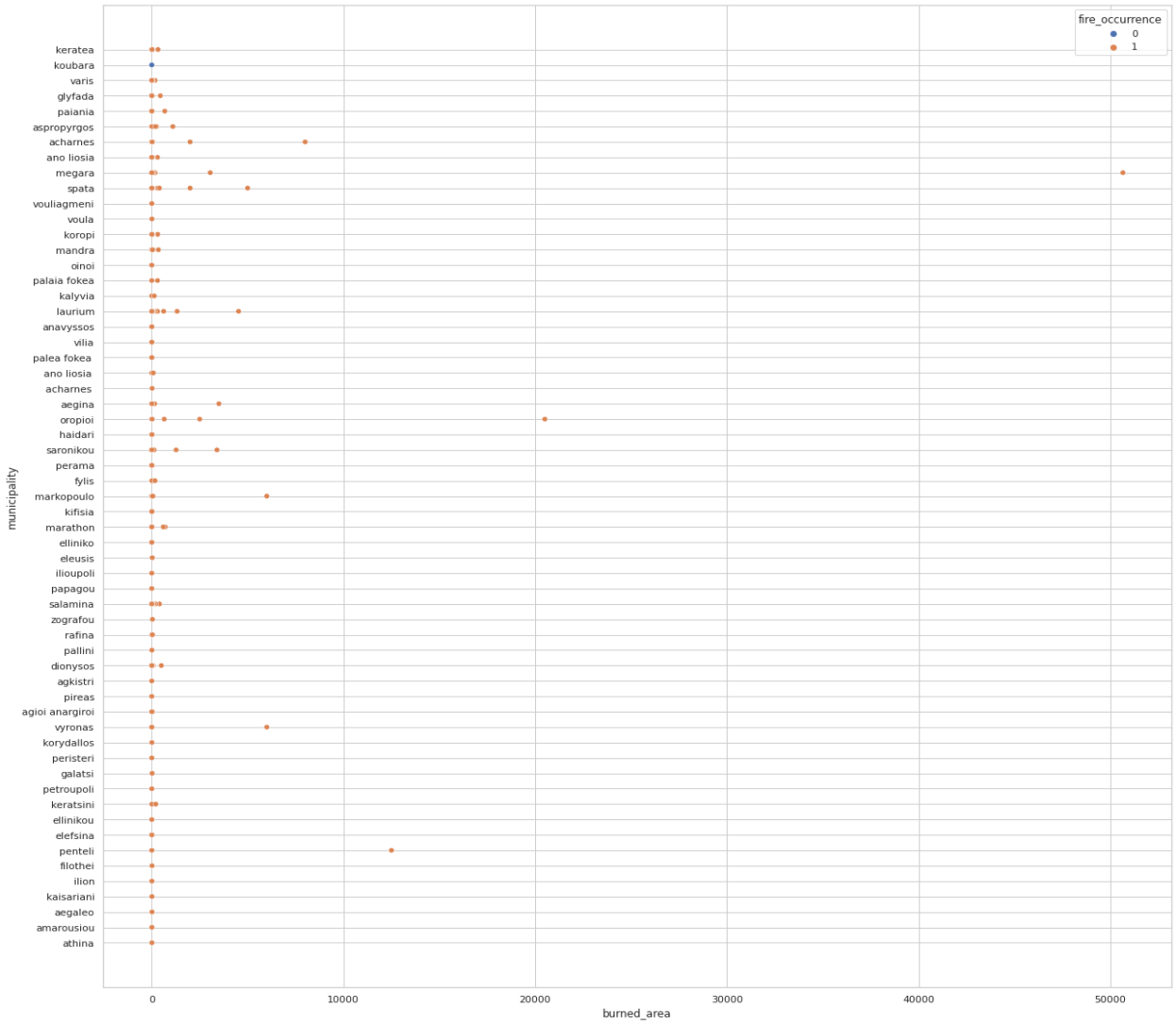


Figure 10: Size of burned area associated with the municipalities

3.3 Montesinho Dataset

The Montesinho natural park dataset is introduced for comparison with wildfire dataset to find out which performs better based only on dynamic weather variables. Montesinho dataset [15] comprises 517 entries and 13 columns namely two geographic features (X,Y), temporal parameters such month and day, Fine Fuel Moisture Code which influences ignition and fire spread, Duff Moisture Code and Drought Code which affect the fire intensity, Initial Spread Index that correlates with fire velocity spread, four weather variables (temperature, relative humidity, wind, rain) and the burned area which denotes the total burned area in hectares. Its form when it was enforced in python platform is being shown in figure 11. The data type is a mix of objects integer and float numbers (figure 12). For reasons of comparison of similar things between the two under consideration datasets the comparison was limited to the 4 weather variables as well as 2 best selected weather variables. The correlation matrix (figure 13) shown that daily rainfall

had a positive correlation with temperature, relative humidity and wind, while temperature had a negative correlation with relative humidity and wind respectively.

	X	Y	month	day	FFMC	DMC	DC	ISI	temp	RH	wind	rain	area
0	7	5	mar	fri	86.2	26.2	94.3	5.1	8.2	51	6.7	0.0	0.0
1	7	4	oct	tue	90.6	35.4	669.1	6.7	18.0	33	0.9	0.0	0.0
2	7	4	oct	sat	90.6	43.7	686.9	6.7	14.6	33	1.3	0.0	0.0
3	8	6	mar	fri	91.7	33.3	77.5	9.0	8.3	97	4.0	0.2	0.0
4	8	6	mar	sun	89.3	51.3	102.2	9.6	11.4	99	1.8	0.0	0.0

Figure 11: Montesinho dataset

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 517 entries, 0 to 516
Data columns (total 13 columns):
 #   Column  Non-Null Count  Dtype  
---  ------  --------------  -----  
 0   X       517 non-null    int64  
 1   Y       517 non-null    int64  
 2   month   517 non-null    object 
 3   day     517 non-null    object 
 4   FFMC    517 non-null    float64
 5   DMC     517 non-null    float64
 6   DC      517 non-null    float64
 7   ISI     517 non-null    float64
 8   temp    517 non-null    float64
 9   RH      517 non-null    int64  
 10  wind    517 non-null    float64
 11  rain    517 non-null    float64
 12  area    517 non-null    float64
dtypes: float64(8), int64(3), object(2)
memory usage: 52.6+ KB
```

Figure 12 : Montesinho dataset -data type

In accordance with the graphs (figure 14), it was pointed out that the temperature ranged from 17 to 21° C for 130 observations, the relative humidity from 30 to 40% for 130 observations, the wind from 5 to 5.9 km / h for 120 observations while the daily rainfall was zero for 510 observations. An exemplary plot between relative humidity and rain is being illustrated in figure 15 summarizing that in times of drought more fires occurred. Any extra information can be retrieved by Montesinho notebook, {appendix code_5, code_10}

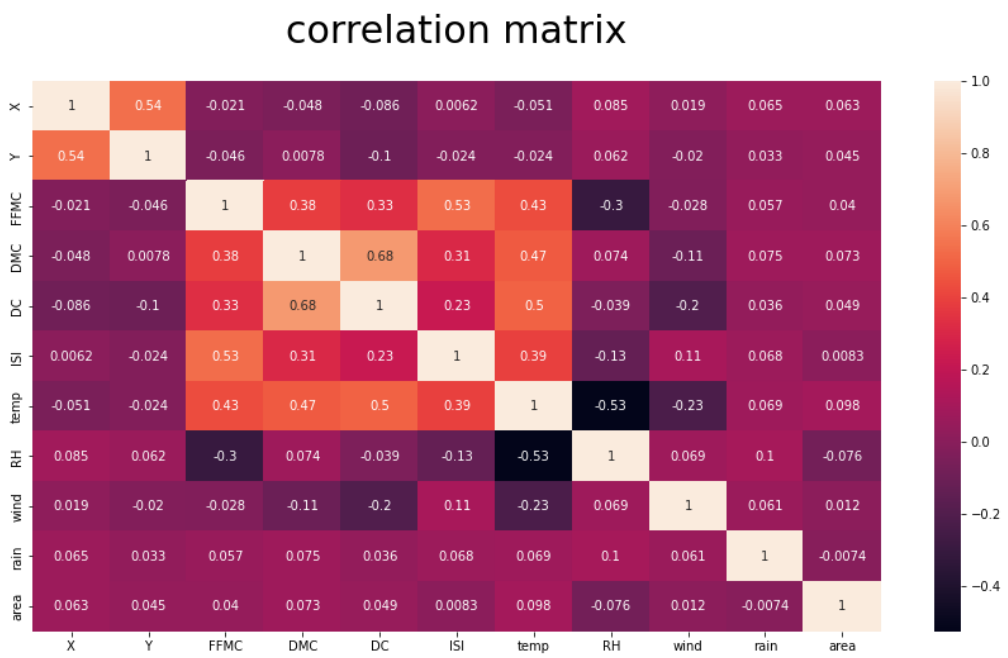


Figure 13 : Montesinho dataset -correlation matrix

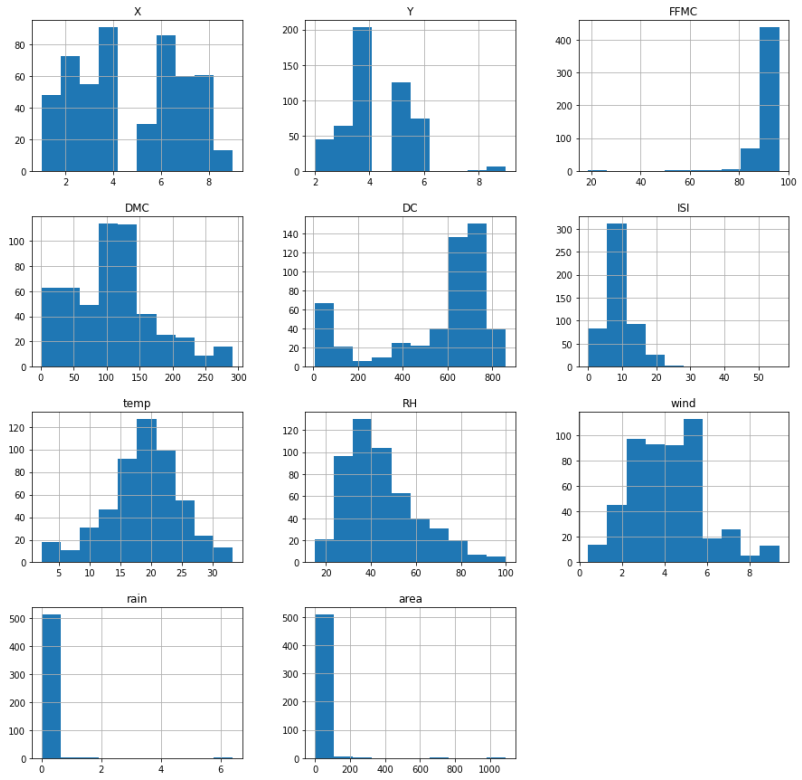


Figure 14 : Montesinho distribution graphs

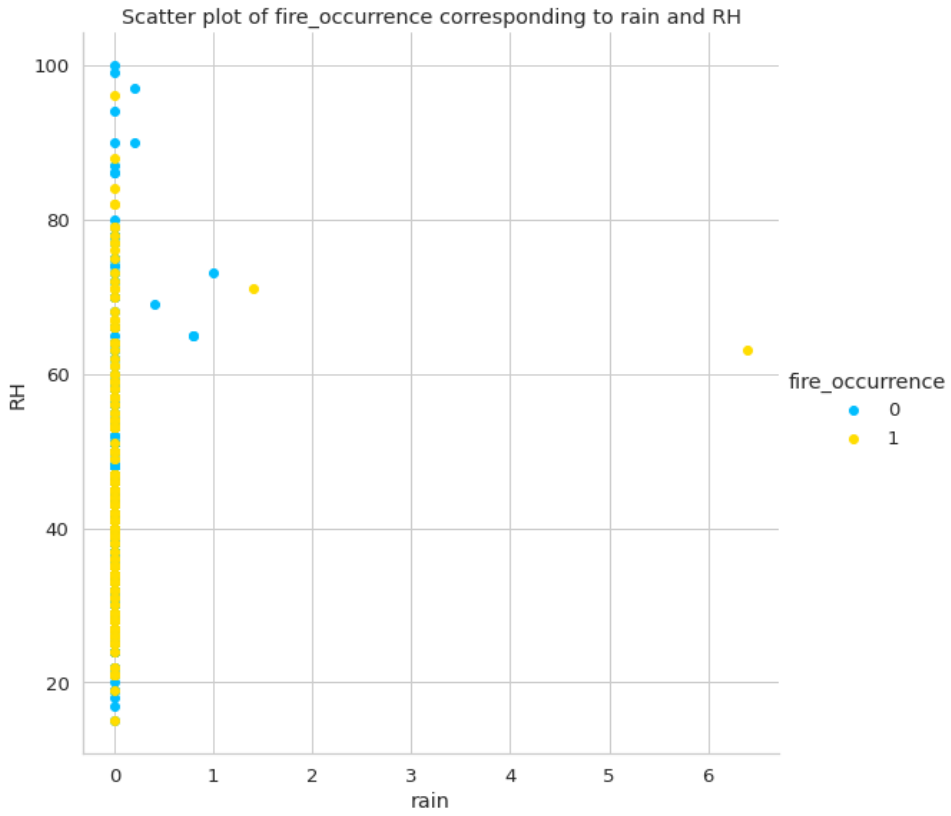


Figure 15 : Fire occurrence incidents corresponding to daily rain and relative humidity in Montesinho dataset.

CHAPTER 4 -- Machine Learning and experimental results

4.1 Definition

Machine Learning refers to the field of computer science, enabling them to learn automatically without specific programming and without human intervention by adapting their actions accordingly. Through exploratory data analysis and the use of computer algorithms they can learn and generate basic behavior patterns for different classes, train the data and generate predictions for them. The learning process begins with observations that are examples or empirical results so that patterns can be identified in the data and the best decisions can be made in the future based on the examples we have. Machine Learning and exploratory data analysis are concepts that complement each other since classification and regression are used in both before-mentioned concepts. So, Machine Learning on the one hand is linked to mathematical optimization techniques, and on the other hand does not lose the advantages that computers give [2]. Therefore, the creation of models or patterns from a dataset and a computer system, is called Machine Learning, with the most well-known techniques that have been developed and used depending on the nature of the problem to be classification and regression.

In the present work, supervised Machine Learning algorithms were applied to implement either classification or regression data prediction. Supervised Machine Learning algorithms construct functions that map given inputs to known desired outputs (training set) with the ultimate goal of generalizing this function to inputs with unknown output. A function is used to predict the value of a variable grounded on the values of a set of input variables. In general, the system is provided with a set of known examples, i.e. a set of situations into which the network may fall along with the results we want the network to give for these situations. As mentioned above in supervised Machine Learning methods, the learning algorithm takes as input the prior knowledge that exists about the problem and the training data, examines the hypothesis area, and returns the final hypothesis (model) as a result. It is therefore essential that prior knowledge and training data are effectively represented to enable the efficient use and production of new knowledge.

Having determined the dataset to be used (wildfire dataset) our purpose is through supervised Machine Learning to process the information contained in this dataset to acquire knowledge when interacting with it and the ability to improve the way is executed an action (hence the accuracy) through repetition. In other words, our goal is to create systems that can be trained from empirical past data (weather variables), in order to perform the work for which they are intended more effectively.

System architecture is being depicted in figure 16 where after splitting the data into training and test data, a feature extraction was carried out for choosing only the 12 daily weather variables (4 or 2 depending on the case study). Then, suitable classifiers /regressors models were selected on the basis of the dataset while by taking the testing data an evaluation of each model was implemented. In our study both classifiers and regressors were trained on up to 70% of the

available data and 30% was reserved for testing the classifiers and regressors correspondingly. Utilizing appropriate metrics such as accuracy or root mean square error of data prediction for each model, a comparative study was conducted of the applied different supervised Machine Learning algorithms with the view to ending up to the best model.

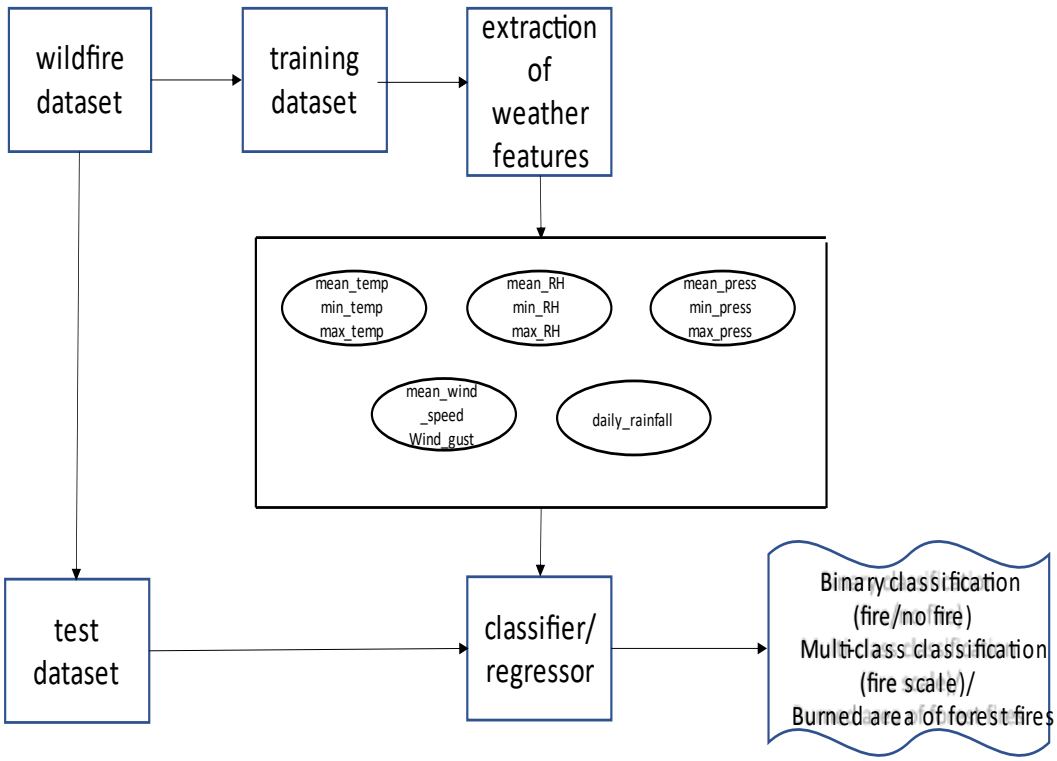


Figure 16 : Diagram of system architecture

4.2 Classification – Regression algorithms

In the current study a plurality of different classifiers was carried out for predicting the probability of fire occurrence and fire scale correspondingly and several regressors for predicting the size of burned area with respect firstly to the wildfire dataset and secondly to Montesinho dataset. The popular algorithms used to perform the classification and regression tasks are described in detail below.

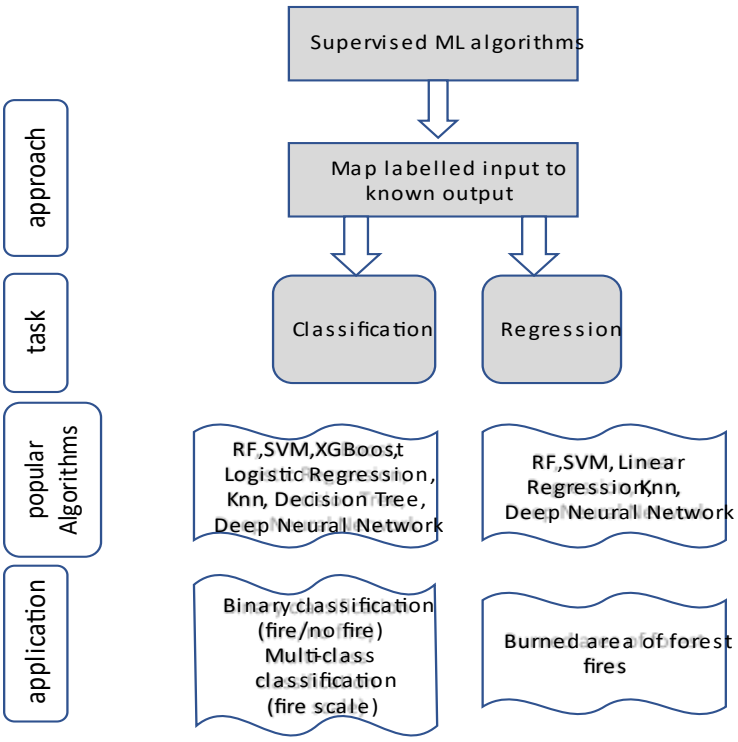


Figure 17 : Diagram showing the applied supervised Machine Learning algorithms for the applications of binary classification of fire occurrence, multiclass classification of fire scale and regression in terms of the size of burned area.

4.2.1 K -nearest neighbors

The K-nearest neighbor classifier is considered to be the simplest classification method as it is non-parametric. It is based on the principle that cases that are in the immediate vicinity of others have similar properties, so for the classification of unclassified cases it suffices to check the nearest neighbors. Consequently, their class is defined by the nearest neighbor's majority. In other words, the KNN algorithm uses 'feature similarity' to predict the values of any new data points. This means that the new point is assigned a value based on how closely it resembles the points in the training set. In the third experiment for predicting the size of burned area of forests fires (regression problem) the KNN was turned out to work best for 5 neighbors, {appendix code_8}.

4.2.2 Logistic Regression – Linear regression

Logistic regression is a powerful supervised Machine Learning algorithm utilized for binary classification problems suitable for investigating the non-linear effect of a dependent categorical variable with respect to the action of many independent variables. The range of logistic regression is bounded between 0 and 1. The difference with the linear regression is that the latter is associated with a model predicting the numerical value of a continuous response variable and

it also requires a linear relationship among inputs and outputs variables. Independent variables are those that can take any value and are used to predict the dependent variable. Therefore, a logistic regression is used for predicting the class of an observation whereas the linear regression focuses mostly on how the standard value of a dependent response variable changes. For the needs of the work the independent variables were considered the 12 dynamic daily weather variables while the dependent target variables depending on the task each time were the probability of fire occurrence (binary classification) and the size of the burned area in terms of the regression.

4.2.3 Support Vector Machines

Support Vector Machines is a popular supervised machine learning method, consisting of models and algorithms capable of analyzing any information and incorporating pattern recognition techniques for regression and classification problems. For example, in a binary classification problem, the SVM method seeks to create a maximum margin hyperplane that acts as a dividing boundary between the individual classes. The aim is for this hyperplane to be as far away as possible from the examples of the dividing classes. Furthermore, in linearly separable problems this hyperplane is defined by a finite number of instances of the training set called support vectors. The SVM classifier therefore tries to find a decision hyperplane that separates all the training examples in such a way that the examples belonging to the same category are on the same side of the hyperplane. Among all the possible hyperplanes, the one for which the distance from the nearest example is the maximum is sought. In addition, through kernel functions (polynomial, radial basis function, linear, sigmoid), SVMs can transform the initial case space so that non-linearly separable problems can be modified into linearly separable problems and finally solved with same methodology.

4.2.4 Decision Trees

Decision Trees are a dynamic and popular tool suitable for classification and regression tasks. When constructing decision trees the different attributes are evaluated retrospectively and that attribute is used in each node that separates the data better. That is to say, in decision trees we seek to separate the training sample using the features that work best for the task. With the ultimate goal of accurately capturing the input-output relationships using the smallest possible tree that avoids overfitting. In decision trees data is initially introduced along with the best feature into the root node so as to be separated according to their metrics (Gini Index). Thus, successive intermediate nodes are created with greater homogeneity till to the final nodes of the tree, where the predictions of a category or a numerical value are made.

The following figure 18 illustrates for the first experiment carried out in the present work, binary classification prediction of fire probability occurrence, the overall growth of decision trees. Having defined as metrics Gini index and maximum depth of the tree equal to 3, at the root node the weather feature that best separated the training sample (979) was the mean relative

humidity (X[3]). This resulted in two nodes, one with Gini of 0.48 and the other with Gini of 0.406. In the sequel different best features were utilized for splitting the intermediate nodes till to final nodes of the tree. The features that gave the best results (lowest Gini Index) were maximum temperature (X[2]) and mean temperature (X[0]), {appendix code_1}.

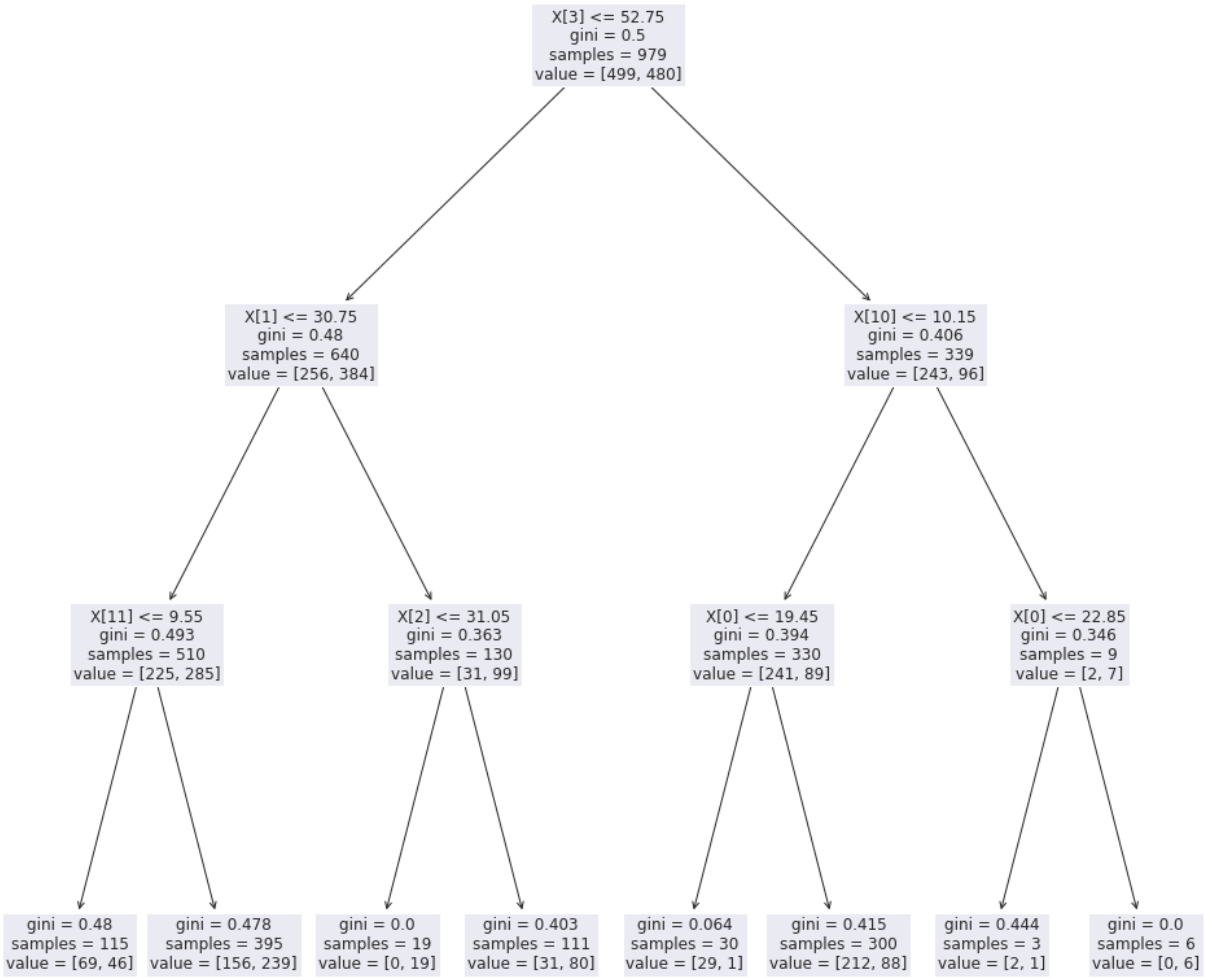


Figure 18 : Decision tree built for binary classification (experiment one – 12 weather variables) for the wildfire dataset.

4.2.5 Random Forest

Another great Machine Learning algorithm that is widely used in classification and regression is random forests since they can be employed both in categorical as well as continuous variables. In practice, this algorithm is essentially a collection of decision trees that run in parallel where in case of classification the predicted class is the most common class in node (majority vote) while in an event of regression the predicted value on a node is the average. Some basic hyperparameters applied to this algorithm are the method of collecting samples from a dataset (bootstrap), the number of trees in the forest (n_estimator), the maximum number of features taken into account for splitting a node (max_features), the maximum depth expressing the

number of levels in each decision tree (max_depth), the minimum number of data points placed on a node before splitting (min_samples_split) as well as the minimum number of data points allowed in a leaf node (min_sample_leaf).

More detailed, in the case of the first experiment (figure 19) of the fire probability prediction (binary classification, fire / no fire) for 12 dynamic daily weather variables, after hyperparameter's tuning by applying RandomizedSearchCV and GridSearchCV, a forest of 400 trees was developed, where a random sample of training data was deployed for each tree according to the bootstrap sampling method. This technique made it possible to create more than one set of training data from a single dataset, resulting in many different trees and therefore many training datasets. At each node, features were randomly selected from all 12 possible weather variables (at least 4 features), then the best splitting was found in the selected features, the forest was developed at a maximum depth of 50 levels, in the sequel the average of trees for new data predictions was calculated, giving the total prediction output by majority voting (classification) of all individually trained trees. In the case of regression, the average response would have been the prediction output, {appendix, code_1}.

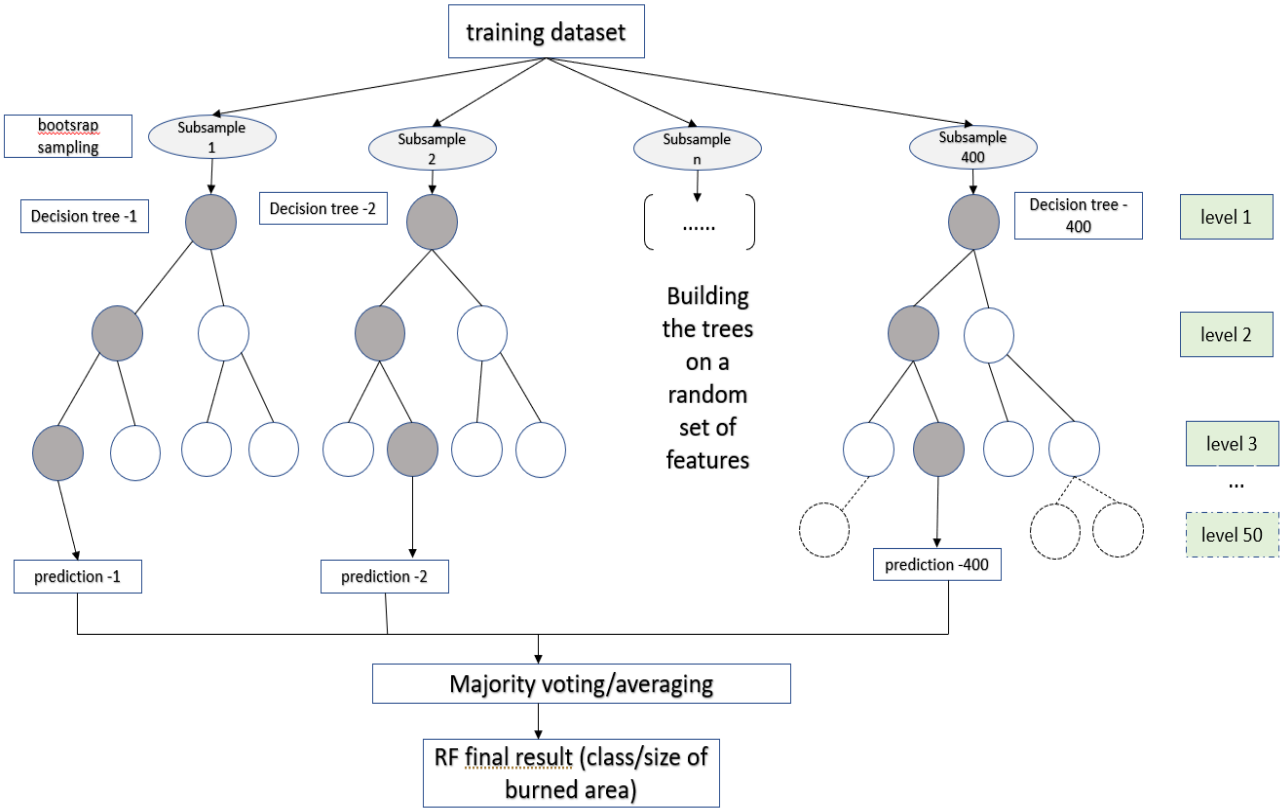


Figure 19 : Random Forest built for binary classification (experiment one -12 weather variables) for the wildfire dataset/ (alternatively for regression).

4.2.6 Extreme Gradient Boosting

Extreme Gradient Boosting (XGBoost) is an implementation of gradient boosting algorithm, a common technique in ensemble learning. This algorithm learns a model faster than many other

machine learning models and functions well on categorical data and limited datasets. In Gradient boosting the trees grow in a sequential way and turns the weak learners into strong learners by adding weights, while reducing the weights of the strong learners. Thus, each tree learns and is strengthened by the previous tree that was developed. In boosting technique, the new created models predict the residuals or errors of prior models, correct them and then added them together to make the final prediction. It is an extremely flexible and agile tool that can work through most forms of regression and classification. Some essential hyperparameters of XGBoost are the number of trees (n_estimator), learning rate, gamma which is the minimum loss reduction required to make a further partition on a leaf node of the tree, column subsampling (colsample_bytree) which is the subsample ratio of columns when constructing each tree, the maximum depth of a tree (max_depth), minimum sum of instance weight (min_child_weight) needed in a child and the used method to sample the training instances (sub_sample).

For instance, in case of the first experiment and for the prediction of the possibility of fire occurrence (binary classification, fire/no fire) for the 4 best selected weather variables (minimum temperature, minimum relative humidity, daily rainfall, average wind speed) the successive configuration of trees is being shown in figure 20. After hyperparameter's tuning 600 trees grew with learning rate 0.02 and maximum depth of tree to 4. Setting subsampling to 0.6 which denotes the fraction of observations to be randomly samples for each tree with the view to preventing overfitting, defining the minimum sum of weights of all observations required in a child to 1, gamma equal to 1 and finally denoting the fraction of columns to be randomly samples for each tree equals to 0.6,{appendix code_4}.

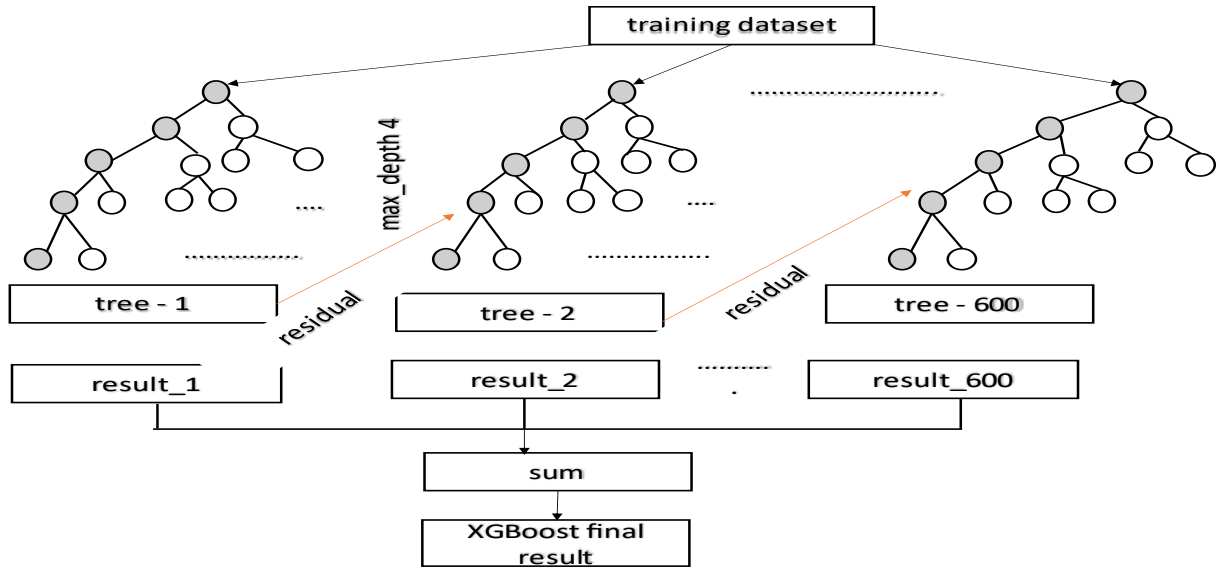


Figure 20 : Extreme Gradient Boosting built for binary classification (experiment one -4 best weather variables) for the wildfire dataset.

4.2.7 Artificial Neural Networks

The relationship among weather variables and depending each time on the experiments like binary classification (fire/no fire), fire scale and the size of burned area of forest fires is a nonlinear relationship. In case of the exact correlations between variables are not clear neural networks seem to be ideal for modeling complex relationships. A typical architecture of neural network is consisted of a plurality of connecting nodes separated into input, hidden and output layers. As more complex the problem is, more neurons and multiple layers need to use. The weather variables as inputs are inserted through input layers and distributed to each of the neurons in the next layer which is the hidden layer. The value from each input neuron is multiplied by initial random value of a weight. The resulting weighted sum is transformed via a transfer function and fed into the next layer. During training, the computed output is compared with the actual response of the inputs and the weights are modified so as to reduce the error function (gradient descent). The process is repeated many times till the error becomes minimized and tolerable. The designed topology of neural networks in the present study is a feed forward which adopts the back-propagation algorithm.

In particular, for the case of the first experiment of binary classification (fire / no fire) where a neural network architecture (figure 21) composed of three fully connected layers and ReLu as activation function acting on hidden layers was fed with two weather variables (minimum relative humidity, daily rainfall) as inputs through 128/12/8 connecting nodes. The activation function for output layer is Sigmoid, a very common function for classification. The default loss function binary_crossentropy was used taking the binary classification fire / no fire prediction into consideration while an Adam optimizer was utilized as a training optimizer for the suggested model to maximize its performance. During the training procedure a validation dataset was applied including cases that were not used in training whereas the rate and speed of the training were controlled by setting the learning rate to 0.01 and batch size equal to 100. The training process stopped at 50 epochs when the validation loss stopped improving, an indication that the network had a good generalization and an overfitting to training dataset had been eliminated. {appendix code_2}.

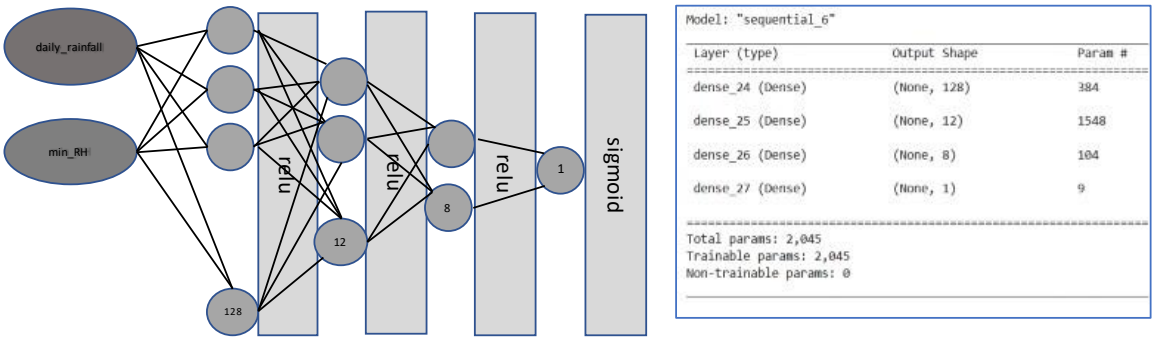


Figure 21 : Backpropagation Neural Network structure built for binary classification (experiment one -2 best weather variables) for the wildfire dataset.

4.3 Experimental results

The structure of the experimental part of the present study was implemented as follows:

- Experiment one: Prediction of fire occurrence probability (binary classification, fire/ no fire) for all weather variables - for the 4 best selected weather variables - for the 2 best selected weather variables, in association with the newly created wildfire dataset and comparison of the performances between the applied supervised Machine Learning algorithms. Comparison of the results with the known state of the art Montesinho dataset for 4 and 2 best weather variables respectively so that could be checked which dataset functioned optimally, {appendix: code_1, code_2, code_3, code_4, code_5}.
- Experiment two: Prediction of the fire scale (multiclass classification, low/medium/large/wildfire) according to the wildfire dataset, {appendix code_6}.
- Experiment three: Size of the burned area of forest fires prediction for all weather variables - the 4 best selected variables - the 4 manually selected variables identical to the Montesinho dataset according to the wildfire dataset and comparison of the applied Machine Learning algorithms (regression problem). Comparison of the results with the known state of the art Montesinho dataset for 4 weather variables so that could be checked which dataset functioned optimally, {appendix code_7, code_8, code_9, code_10}.

4.3.1 Experiment one: Binary classification fire/no fire

For binary classification, the metrics of performance are grounded on the four values of the contingency table (True Positive, False positive, True Negative, False Negative) acquired by putting in practice the classifier to the testing data. The standard performance metrics are defined as follows

Accuracy: the proportion of correct fire occurrence predictions (both true positives and true negatives)

Accuracy = (TP + TN)/ (TP + FP + TN + FN).

Precision: the proportion of true positives against all the positive predictions, both true positives and false positives

Precision= TP / (TP + FP).

Recall: the proportion of the true positives against all positives, the true positives and false negatives.

Recall =TP / (TP + FN).

The metrics of precision and recall are quite essential because a high precision means a small number of false alarms which is interpreted as a number of predicted fire occurrences that never broke out. On the other hand, a high Recall (or sensitivity for binary classification) denotes the probability that a fire outbreak is indeed predicted as positive. Non-predicted fire occurrences can have serious side effects in the environment and society and that is why they are very important in the sensitivity of fire occurrence predictions. Prior to the implementation of the first experiment, the distributions of fire / no fire observations were verified to have been balanced. Each fire incident was declared as 1 while no fire as 0, figure 22.

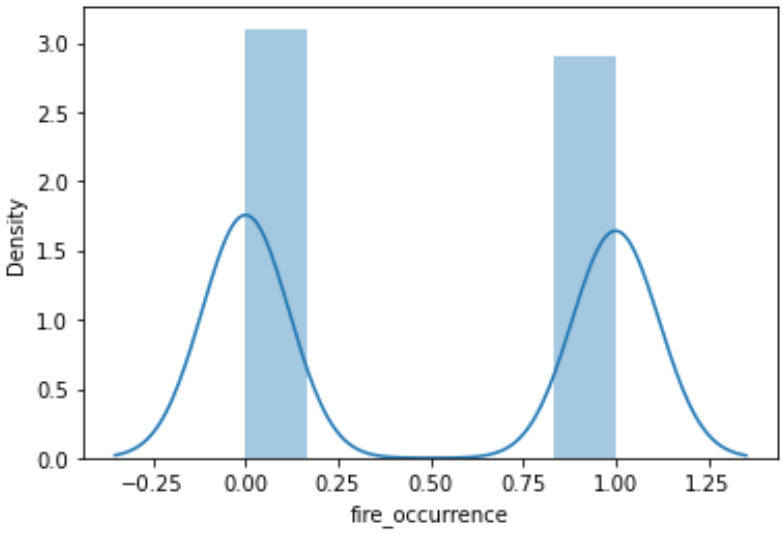


Figure 22 : Balanced data distribution of wildfire dataset for binary classification.

Therefore, by applying several supervised Machine Learning algorithms for conducting fire occurrence predictions for the whole 12 weather variables (mean, minimum, maximum_temperature - mean, minimum, maximum_ relative humidity -mean, minimum, maximum_(atmospheric) pressure, mean wind speed, wind gust, daily rainfall), the overall performance of each individual model is being depicted in the next table 2. From the below it can be clearly seen that the Random Forest model outperformed the other models, achieving 71% accuracy, precision for a small number of false alarms 69% and recall for a low number of unpredictable fires 67%, {appendix code_1}.

As mentioned before in order to construct the classifier model according to which the data will be classified, a dataset of known observations is used for model training (training set) and another set for its control (test set), so as to make it possible to classify future data. In cases where there is not a large dataset to train the model, a technique called N-fold cross-validation is used. According to it, the set of examples is divided into N subsets, and then each of them is used sequentially as a control set, while the remaining N-1 subsets are combined and used as a training set. At the end of the N trainings, the results are used to get an average accuracy for the model.

Choosing Random Forest as a best model for binary classification, hyperparameter tuning was executed with the view to optimizing its performance. The implementation was done by using initially RandomizedSearchCV algorithm and defining a grid of hyperparameter ranges (best

model	accuracy	precision	recall
RF	0.707	0.691	0.673
RF_tuned	0.700	0.675	0.689
SVM(rbf)	0.688	0.718	0.546
Knn	0.674	0.641	0.684
XGBoost	0.662	0.627	0.679
SVM(linear)	0.640	0.620	0.580
Logistic Regression	0.640	0.609	0.643
Neural Network	0.633	0.687	0.393
Decision Tree	0.633	0.592	0.689
SVM(sigmoid)	0.530	0.00	0.00
SVM(polynomial)	0.470	0.470	1.00

Table 2：Overall performance for each individual model using 12 weather variables by wildfire dataset for conducting binary classification.

parameters -n_estimators:400, max_features:auto, max_depth:90, min_sample_split:5, mean_sample_leaf:1, bootstrap: True), {appendix code_3}. In our case there were 4320 possible combinations but the benefit of RandomizedSearchCV was that it did not try every combination but selected at random a wide range of values so as to narrow down the range of values for each hyperparameter. Secondly a GridSearchCV applied focusing upon the best parameters of the RandomizedSearchCV so as to evaluate all the designated possible combinations (best parameters -n_estimators:400, max_features:12, max_depth:50, min_sample_split:4, mean_sample_leaf:1, bootstrap: True). The accuracy of RF_tuned marginally decreased to 70% but the recall (sensitivity) increased up to 69%. The confusion matrix on the test data with respect to RF and RF_tuned are being shown in figure 23 respectively.

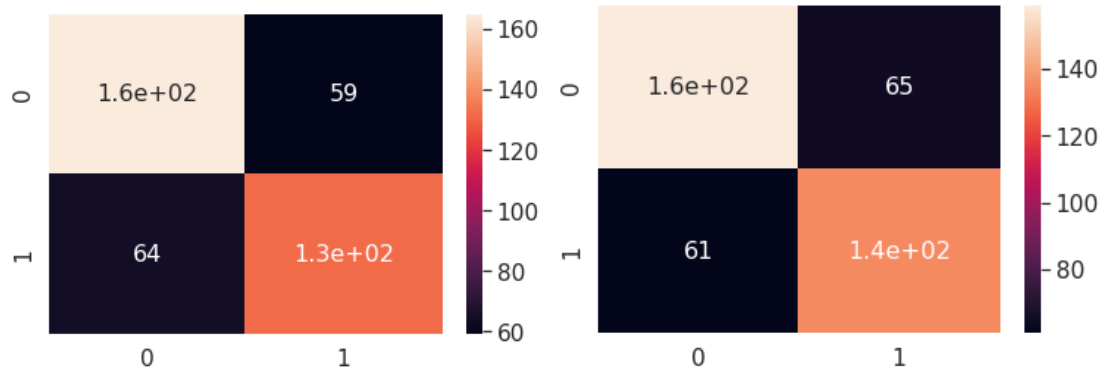


Figure 23 : RF and RF_tuned confusion matrices.

Comparing the two confusion matrices it is being noticed that after hyperparameters tuning the sensitivity (False Negatives:65) improved but at the same time the precision (False Positives: 61) decreased. Another interesting diagram (figure 24) is the following which illustrates the feature importance of the whole 12 weather variables by employing the Random Forest model. It explains which of the features of wildfire dataset are the most useful towards fire prediction binary classification. Consequently, for all 12 weather variables it is being revealed that the average and maximum relative humidity as well as the wind gust are of relative importance in predicting the possibility of fire occurrence.

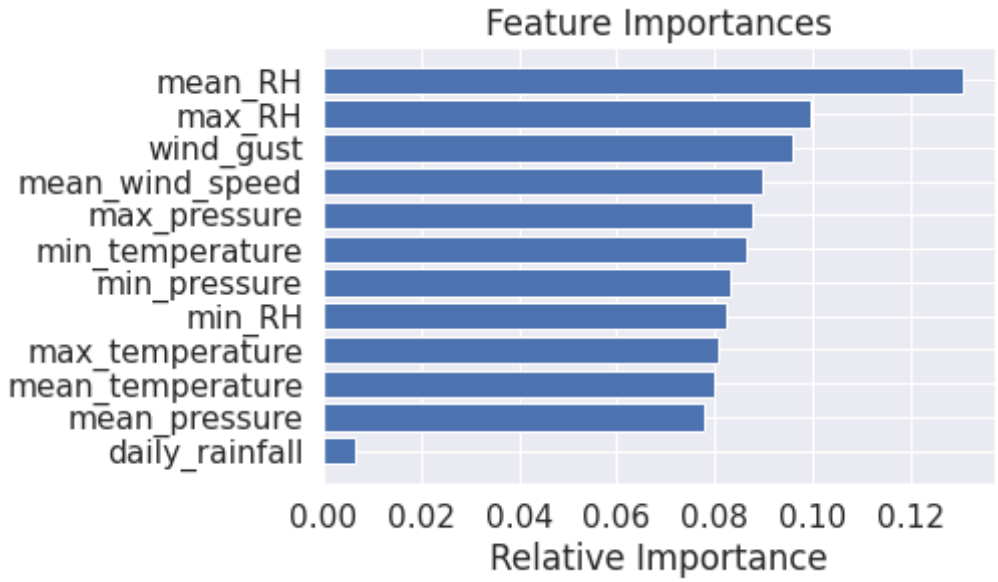
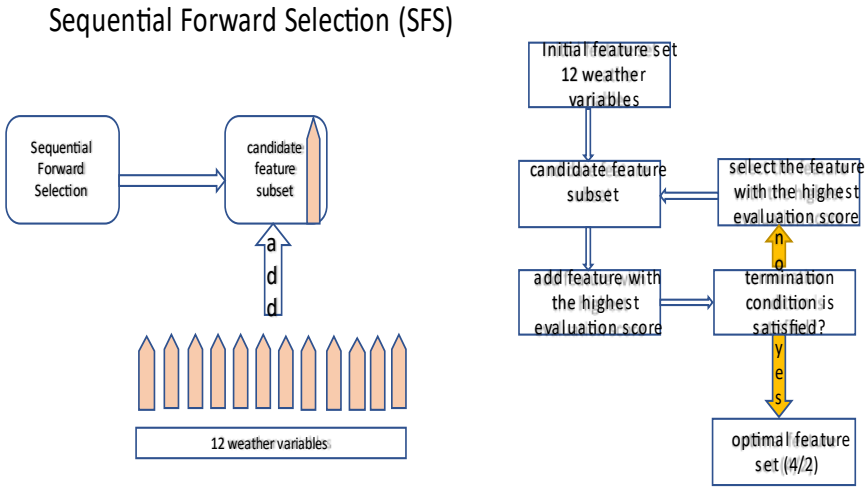


Figure 24 : Feature importance.

Quite intriguing conclusions can be also derived from the table 2. Of particular interest in terms of sensitivity were the models Decision Trees (69%), Knn (68%) and XGboost (68%) correspondingly. As for precision, Support Vector Machine(rbf) implemented to 72% while Neural Network to 69%.

4.3.1.1 Experiment one: Binary classification fire/no fire using the best 4 selected weather variables extracted from wildfire dataset

Then it was tried to find the 4 best weather variables by applying Sequential Forward Selection, an algorithm which is to automatically select a subset of features that is most relevant to the problem. As it can be seen from the next figure 25 the algorithm is initialized with an empty set and returns a subset of a predefined number of selected features. In the candidate feature subset, the performance of each feature is evaluated each time, with the consequence that the ones related to the best performance of the classifier are maintained.



1 to 4 of 4 entries Filter

index	feature_idx	cv_scores	avg_score	feature_names	ci_bound	std_dev	std_err
1	5	[0.59428571 0.67714286 0.54857143 0.65902579]	0.6197564469914 04	min_RH	0.08232891772343323	0.051359873156572 01	0.029652636592491885
2	5,9	[0.59142857 0.67714286 0.54571429 0.66475645]	0.6197605403192796	min_RH,daily_rainfall	0.0863340139980045	0.053858402741679 03	0.031095163321031665
3	1,5,9	[0.58857143 0.69714286 0.56 0.59598854]	0.6104257060990586	min_temperature,min_RH,daily_rainfall	0.08309513118875275	0.051837865913890474	0.02992860583960039
4	1,5,9,10	[0.57142857 0.66285714 0.6 0.6504298]	0.621178878428162	min_temperature,min_RH,daily_rainfall,mean_wind_speed	0.05952832573940743	0.037136006930999 04	0.021440483598240107

Show 25 per page

Figure 25 : Sequential Forward Selection – best 4 weather variables in wildfire dataset.

Hence, in accordance with the next comparative table 3 is being presented that the Extreme Gradient Boosting algorithm performed quite better than any other supervised Machine Learning algorithm. By employing the Sequential Forward Selection with the XGBoost classifier the features minimum temperature, minimum relative humidity, daily rainfall, mean wind speed were chosen as the 4 best features providing accuracy of 67%, precision of 65% and recall of 64% {appendix code_4}. In this alternative embodiment of using the best 4 chosen weather features was pointed out that hyperparameter tuning applying GridSearchCV (best parameters – colsample_bytree: 0.6, gamma:1, max_depth:4, min_child_weight:1, subsample:0.6) worsened the corresponding metrics, figure 26 confusion matrices of XGBoost and tuned_XGBoost. In the

event that our interest is focalized on high sensitivity, very good results (Recall: 86%) are being presented by the neural network for the same 4 weather variables, figure 26.

model	weather variables	accuracy	precision	recall
XGBoost	min_temp, min_RH, daily_rainfall, mean_wind_speed	0.6738	0.6545	0.6378
tuned_XGBoost	min_temp, min_RH, daily_rainfall,mean_wind_speed	0.6571	0.6368	0.6173
Neural Network	min_temp, min_RH, daily_rainfall, mean_wind_speed	0.6310	0.5695	0.8571
RF	mean_pressure, min_RH, daily_rainfall, mean_RH	0.6381	0.6089	0.6276
SVM (rbf)	mean_temp, mean_RH, wind_gust, mean_wind_speed	0.6400	0.6100	0.6300

Table 3：Overall performance for each individual model using 4 weather variables by wildfire dataset for conducting binary classification.

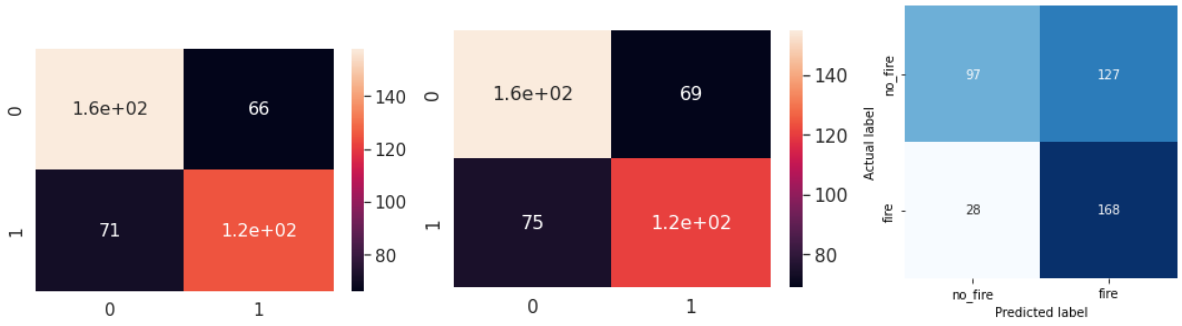


Figure 26 : XGBoost, tuned_XGBoost and Neural Network confusion matrices.

In the sequel based on the experimental performances and selected best 4 weather features extracted from the wildfire dataset as cited and in chapter 3, a comparison was made with known Montesinho dataset. This dataset appears to be balanced and appropriate for binary classification tasks such as fire occurrence predictions, figure 27. In Montesinho study case several supervised Machine learning algorithms developed for only 4 selected weather variables (temperature, rain, wind, relative humidity) concluding that the Extreme Gradient Boosting algorithm as in wildfire dataset achieved better accuracy of 56%, precision of 58% and recall of 41%, table 4 {appendix code_5}. Although Random Forest model succeeded better in sensitivity.

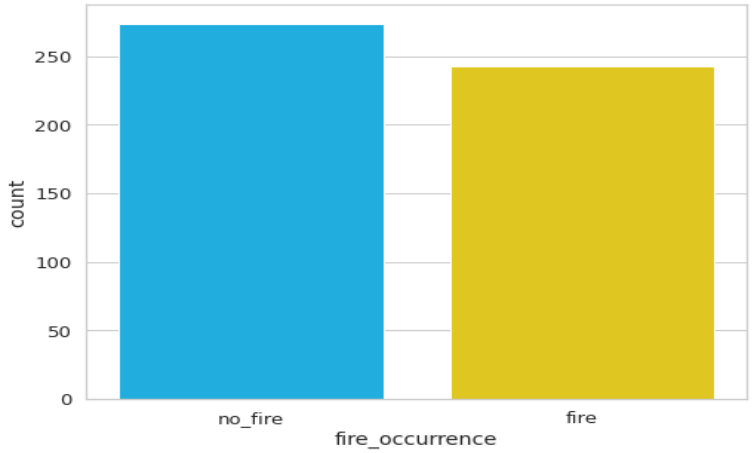


Figure 27 : Balanced data distribution of Montesinho dataset for binary classification.

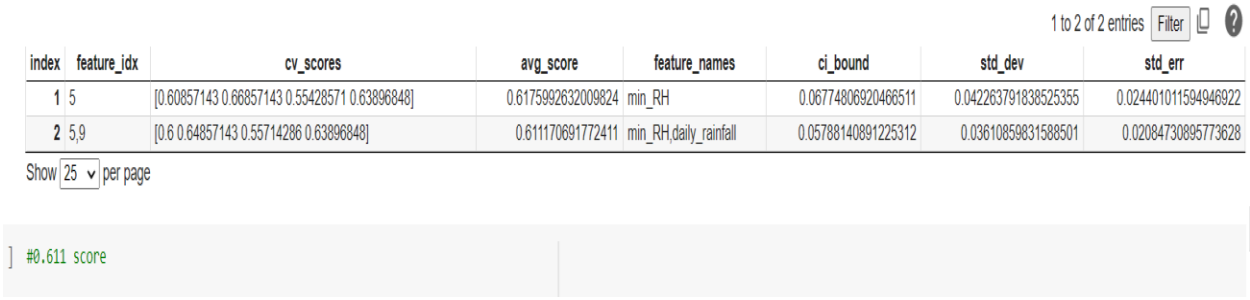
When comparing the 2 examined datasets, it is obvious that for the same applied XGBoost algorithm and for the partially identical 4 weather variables, the newly created wildfire dataset gives much better predictions of fire occurrence probability.

Montesinho_model	accuracy	precision	recall
XGBoost	0.558	0.582	0.410
RF	0.558	0.565	0.500
SVM(rbf)	0.500	0.0	0.0
Neural Network	0.538	0.588	0.256

Table 4：Overall performance for each individual model using 4 weather variables by Montesinho dataset for conducting binary classification.

4.3.1.2 Experiment one: Binary classification fire/no fire using the best 2 selected weather variables

In a similar vein carrying on experiments and trying to verify the critical conclusions of [5],[6] a challenge was designated to find the best 2 weather variables for the wildfire dataset. The final purpose was to be checked whether my experimental results coincided with referred prior art documents in terms of weather variables and if they were sufficient to predict the probability of fire occurrence. Implementing in the same way Sequential Forward Selection with an RF model for the selection of the best 2 weather variables, it emerged as with the state-of-the-art documents that they were the minimum relative humidity and the daily rainfall, figure 28 {appendix code_1}. Employing for the specific 2 weather variables a Radom Forest model and Neural Network as described in chapter 4.2.7 it resulted that the latter performed marginally better in terms of accuracy (63.5%) and precision (63%) but not for recall (54%), {appendix code_2}. In the event of our interest is centered around the sensitivity then the RF seems to be an ideal model, table 5 {appendix code_1}. In figure 29 is being illustrated The Neural Network confusion matrix on the test data and the training process.



1 to 2 of 2 entries Filter

index	feature_idx	cv_scores	avg_score	feature_names	ci_bound	std_dev	std_err
1	5	[0.60857143 0.66857143 0.55428571 0.63896848]	0.6175992632009824	min_RH	0.06774806920466511	0.042263791838525355	0.024401011594946922
2	5,9	[0.6 0.64857143 0.55714286 0.63896848]	0.6111706691772411	min_RH,daily_rainfall	0.05788140891225312	0.03610859831588501	0.02084730895773628

Show 25 per page

```
] #0.611 score
```

Figure 28 : Sequential Forward Selection – best 2 weather variables in wildfire dataset.

model	accuracy	precision	recall
Neural Network	0.6357	0.6272	0.5408
RF	0.6333	0.6071	0.6071
tuned_NN	0.6310	0.6881	0.3827

Table 5 : Overall performance for each individual model using 2 best weather variables by wildfire dataset for conducting binary classification.

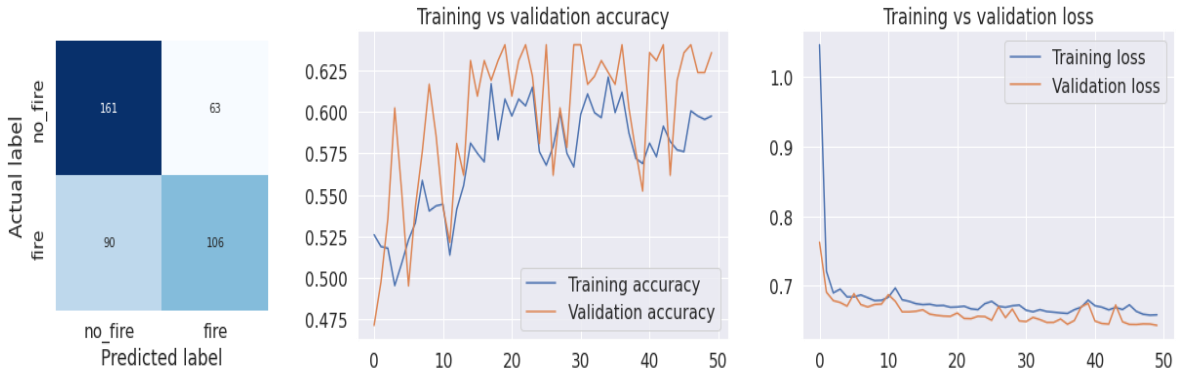


Figure 29 : Neural Network confusion matrix and training process.

In the effort to improve the performance of the neural network GridSearchCV was used for hyperparameters tuning. The accuracy reduced marginally, and the sensitivity deteriorated greatly. On the contrary, the precision improved significantly. The optimal resulting hyperparameter values were epochs: 50, batch size:10, learning rate: 0.001, activation function: sigmoid, number of neurons:100 and optimizer: Adam, {appendix code_3}.

By the same logic, deploying Sequential Forward Selection with either RF or XGBoost model in Montesinho dataset the best 2 selected weather variables turned out to be different features (rain and wind, figure 30), {appendix code_5} compared to wildfire dataset. In this case RF model proved to carry out slightly better than XGBoost while the neural network failed to provide reliable results, table 6. And in this case, it is clearly proved that wildfire dataset functioned much better for the two best selected weather variables in relation to known prior art Montesinho dataset.

	feature_idx	cv_scores	avg_score	feature_names	ci_bound	std_dev	std_err
1	(2,)	[0.5307692307692308, 0.5348837209302325, 0.527...	0.529979	(rain,)	0.005125	0.003197	0.001846
2	(2, 3)	[0.5461538461538461, 0.5271317829457365, 0.534...	0.533825	(rain, wind)	0.012487	0.00779	0.004497

Figure 30 : Sequential Forward Selection – best 2 weather variables in Montesinho dataset.

Montesinho_ model	accuracy	precision	recall
RF	0.538	0.568	0.320
XGBoost	0.532	0.551	0.346
NN	0.5	0.0	0.0

Table 6 : Overall performance for each individual model using 2 best weather variables by Montesinho dataset for conducting binary classification.

4.3.2 Experiment two: Fire scale prediction (multiclass classification)

In the second experiment, an attempt was made to forecast the fire scale. Given from the statistical data of wildfire dataset that the maximum burned area is 50650 acres, the scale of the fire was arbitrarily defined as:

small fire (class: 0), provided that the burned area <50 acres

medium fire (class: 1), provided that 50 < burned area < 500 acres

large fire (class: 2), provided that 500 < burned area < 5000 acres

and wildfire (class: 3), provided that burned area > 5000 acres.

The biggest problem that arose from the above predefined scaling was that the data in wildfire dataset appeared imbalanced. More specifically in a number of 1400 entries the data distribution regarding the scale was 1320 observations for small fire, 45 observations for medium fire, 15 observations for large fire and 19 observations for wildfire. The following figure 31 illustrates exactly the distribution of imbalanced data where the four classes were highly imbalanced.

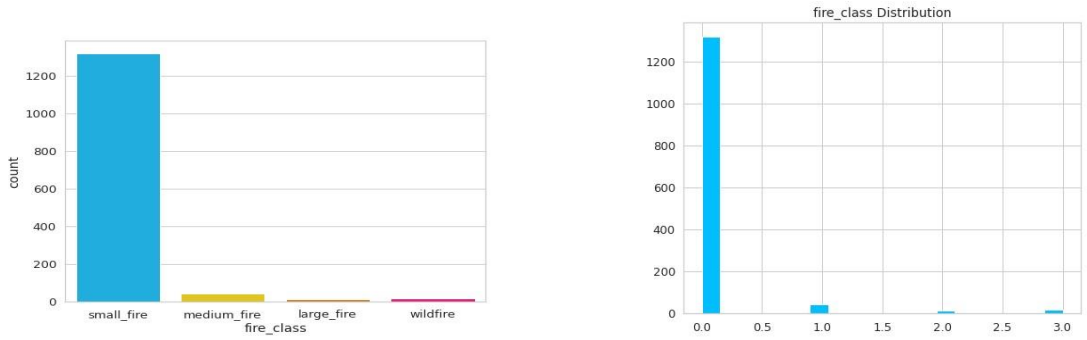


Figure 31 : Imbalanced data distribution of wildfire dataset for multiclass classification.

Data imbalance has consequences to lead classifiers to bias towards majority class at the expense of underrepresented minority class. To deal with this situation a random under sampling (figure 32) was applied with the aim of balancing class distribution by randomly removing majority class samples. By this approach it was managed to cut down the number of examples from the majority class and match them with the number of examples in the minority class. By using this technique is reduced the risk of bias toward the majority. However, a possible drawback might be that essential information is expelled during the transformation of majority class into equal with the rest minority classes, since we have a significant loss of data. In the current experiment during the application of the technique the 1400 observations were limited to 60 observations (44 training and 16 test).

Moreover, for the imbalanced wildfire dataset a new metric F1 score was introduced so as to compare the applied supervised Machine Learning algorithms. This metric considers not only the number of prediction errors but also the type of errors that are made. In essence the average of precision and recall is calculated

F1 score= 2* (precision*recall)/ (precision + recall).

By employing several Machine Learning algorithms, it was determined that the K nearest neighbor algorithm implemented better than the others with a F1 score of 45%, table 7, {appendix code_6}. As shown in Figure 33 for K=4 nearest neighbors achieved the minimum error, {appendix code_6}.

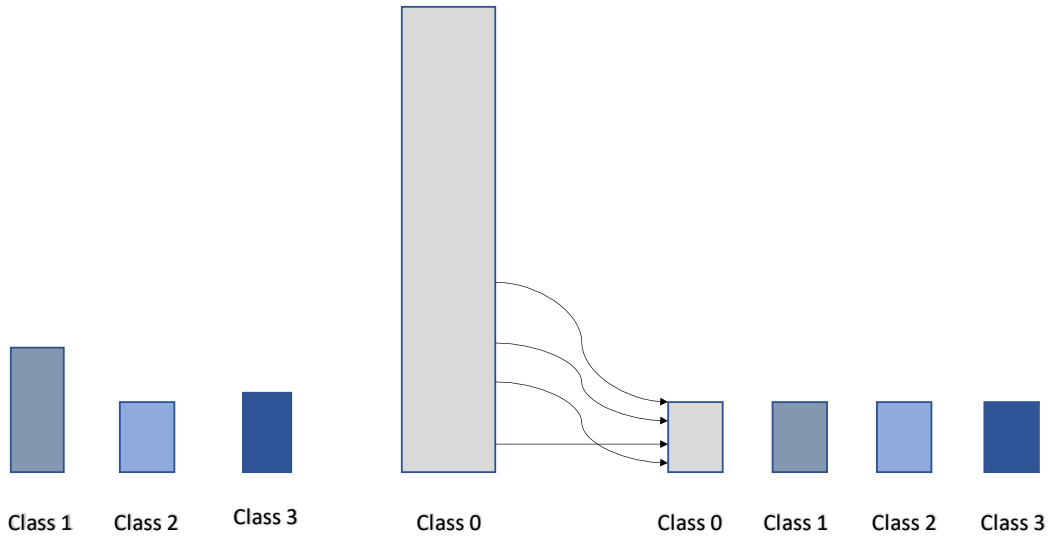


Figure 32 : Random undersampling.

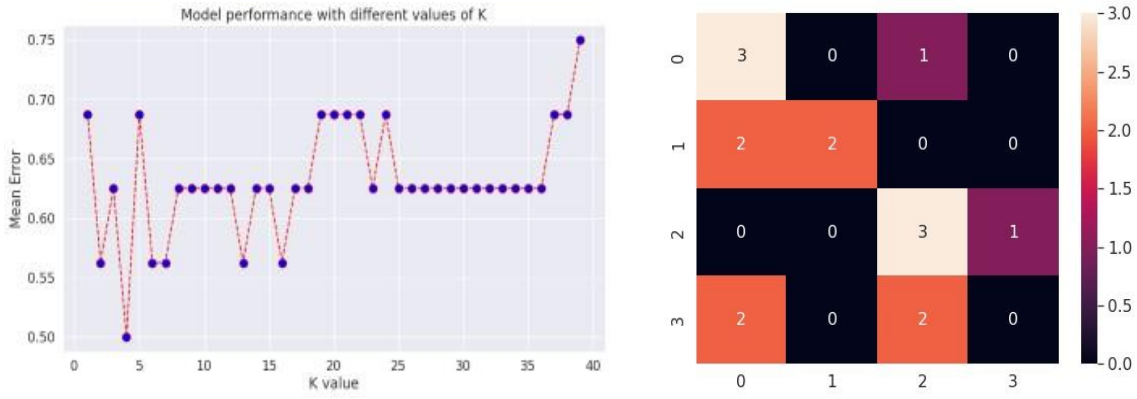


Figure 33 : Knn model performance for multiclass classification and corresponding confusion matrix.

model	accuracy	f1_score
Knn	0.500	0.453
Decision Tree	0.437	0.393
SVM(linear)	0.380	0.365
Decision Tree_tuned	0.375	0.380
RF_tuned	0.375	0.375
XGBoost_tuned	0.312	0.292
SVM(rbf)	0.310	0.210
SVM(sigmoid)	0.250	0.100
SVM(polynomial)	0.250	0.100

Table 7 : Overall performance for each individual model using 12 weather variables by wildfire dataset for conducting multiclass classification.

4.3.3 Experiment three: Size of the burned area of forest fires (regression problem)

As is well known the adverse effects of conflagrations have a great impact on our quality of life. Another important objective of this current study is the prediction of the size of burned area caused by fire occurrences. The main purpose of third experiment was to explore and evaluate supervised machine learning regressors with the view to accurately forecasting the burned area. Fruitful inferences were drawn based on the wildfire dataset by comparing the predictive performance of the regressors concerning all the weather variables and the best 4 of them. In the end, similarly with the first experiment, a comparison was carried out with the known prior art Montesinho dataset for the 4 selected weather variables. By scrutinizing the formed scatter plot graph (figure 34-last series) being shown the relationship between burned area and weather variables, it became obvious that burned area generally increased under

conditions of high mean temperatures and high mean atmospheric pressures respectively, low mean relative humidity, low daily rainfall and low mean wind speed as well as strong wind gusts.

Once again to begin with the exploration of wildfire dataset was seen that the distribution of data regarding the burned area was imbalanced. In figure 35 most fires presenting small and medium scale. In order to improve the symmetry of data a random under sampling was applied as in the second experiment limited the total 1400 observations to 101 (64 training and 37 test), {appendix code_7}. The overall performance of the applied regressors was computed by Root mean Squared error and r^2 score. Root Mean Square Error is an extension of the mean squared error and depicts how far apart the predicted values are from the actual values in a dataset, on average while r^2 score indicates the proportion of variance in the response variable of a regression model that can be explained by the predictor variables. In each case the desirable is to effect a high accuracy implying lower value of RMSE and a higher value of r^2 score. So, putting in practice several supervised Machine Learning regression algorithms the table 8 indicates that k nearest neighbors outperformed other models for the undersampled data. More detailed, for K equal to 5 nearest neighbors the model achieved RMSE of 1982.59 and r^2 score of 70%, figure 36b), {appendix code_8}. According to the data statistics the total burned area ranges from 0 to 50650 acres, a useful information for calculating the normalized value of RMSE. Figure 36a) indicates the actual burned area and burned area predicted by Knn model. The dash line shows the 1:1 correspondence.

4.3.3.1 Experiment three: Size of the burned area of forest fires using the best 4 selected weather variables extracted from wildfire dataset

In similar fashion by implementing Sequential Forward Selection with either RF or XGBoost model the best 4 selected weather features were computed as minimum temperature, mean relative humidity, minimum relative humidity and daily rainfall. The application of the relevant algorithms demonstrated that the models according to table 9 did not follow the trend of data and failed to fit them. An explanation of the negative results can be given to the fact that r^2 score value increases only by adding independent variables and not by subtracting {appendix code_9}.

model	r^2 score	RMSE	Normalized_RMSE
Knn	0.700	1982.59	0.04
Linear Regression	0.107	3421.77	0.07
Neural Network	-0.003	3625.70	0.07
SVM	-0.095	3859.67	0.08
RF	-0.137	3859.67	0.08

Table 8 : Overall performance for each individual model using 12 weather variables by wildfire dataset for predicting the size of burned area.

model	r^2 score	RMSE	Normalized_RMSE
Linear Regression	-0.02	3658.07	0.072
RF	-0.05	3718.07	0.073
SVM	-0.09	3788.84	0.075
Deep Neural Network	-0.02	3657.44	0.072
Knn	-0.60	4573.98	0.090

index	feature_idx	cv_scores	avg_score	feature_names	ci_bound	std_dev	std_err
1	9	[-284464.58178218 -232002.874255 -144284.44142688 -9014400.50314223]	-2418788.1001515724	daily_rainfall	6104846.464861641	3808307.9040861563	2198727.593581122
2	5,9	[-392151.64055541 -477965.32559945 -204707.366044 -8999840.8746753]	-2518666.301718541	min_RH,daily_rainfall	6000209.511909596	3743212.353019943	2161144.6596499966
3	1,5,9	[-278597.53581658 -495787.78575089 -185905.61530035 -9051752.43826373]	-2503010.843782885	min_temperature,min_RH,daily_rainfall	6063421.338041984	3782590.12055554	2183879.42433676
4	1,3,5,9	[-303506.33753911 -327049.54117924 -182146.3175421 -8985596.40755248]	-2449574.650953233	min_temperature,mean_RH,min_RH,daily_rainfall	6049610.6374265915	3773974.486445145	2178905.185663884

Table 9 : Overall performance for each individual model using the best 4 weather variables by wildfire dataset for predicting the size of burned area.

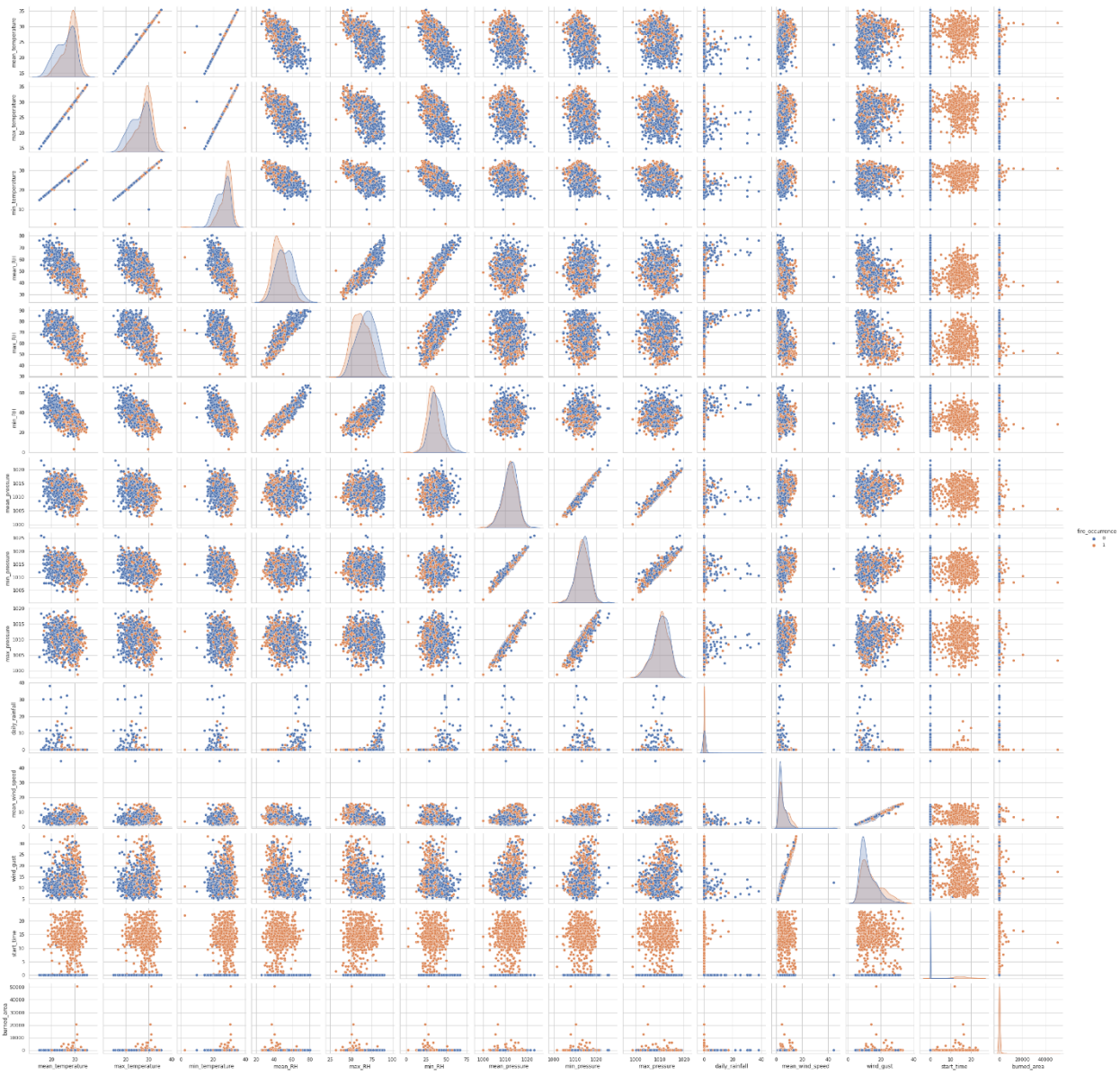


Figure 34 : Scatter plot graph – burned area Vs weather variables (last series).

Alternatively, a manual selection of the 4 weather variables (mean temperature, mean relative humidity, daily rainfall, mean wind speed) was attempted with the aim of being as close as possible to the weather variables with Montesinho dataset. There was a slight improvement in the forecast of size of burned areas by applying the Linear Regression model (r^2 score: 2%, RMSE: 3584.70) but not capable enough to reliably fit the dataset.

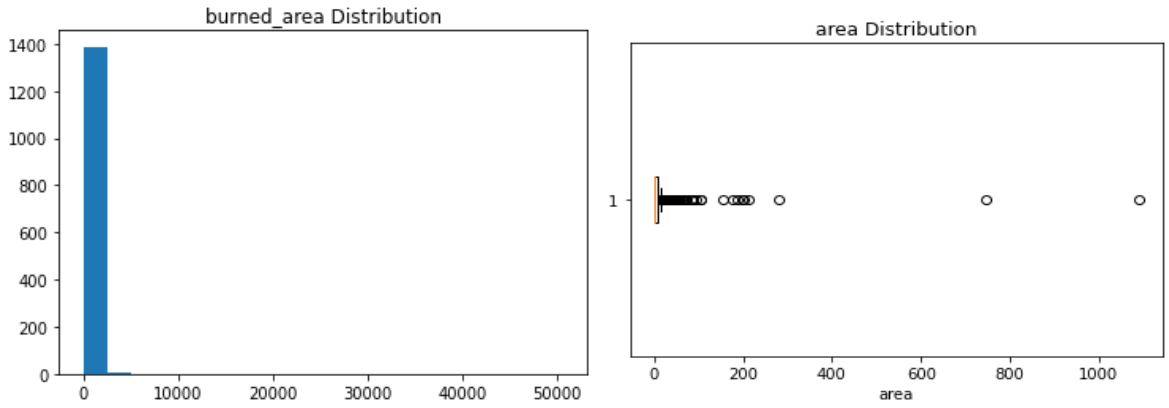


Figure 35 : Imbalanced data distribution of wildfire dataset for forecasting the size of burned area.

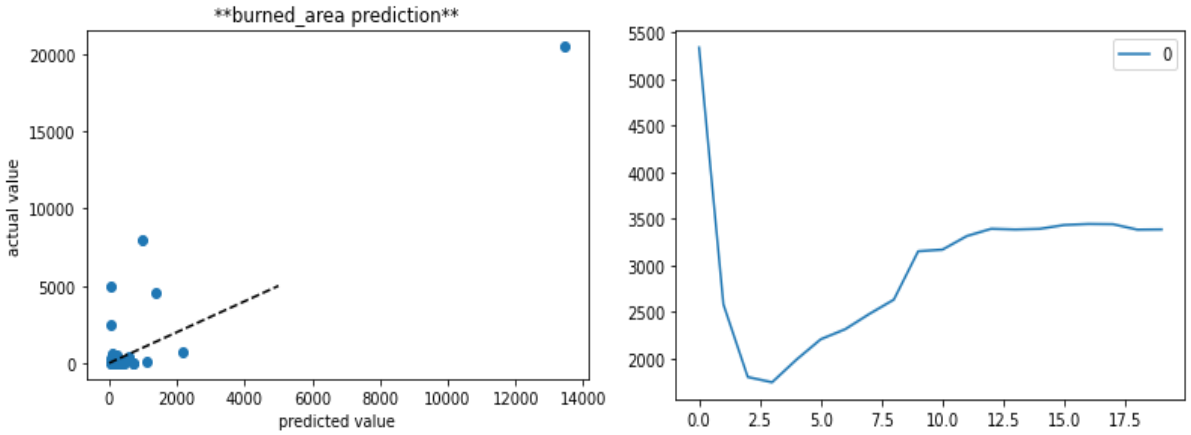


Figure 36 : a) Actual burned area and burned area predicted by the Knn model. b) RMSE values against K values.

In a similar manner as the first experiment a comparison was conducted with the Montesinho dataset for 4 selected weather variables. Likewise, the Montesinho dataset appeared to be imbalanced with the majority of burned area less than 200 hectares, figure 37. Before modeling the problem as a regression task there was a need to synthesize a balanced Montesinho dataset. By implementing random undersampling as in previous cases the number of observations limited to 85 (51 training, 34 test). After that by employing the supervised Machine Learning regressors was proved that for the 4 chosen weather features (rain, wind, temperature, relative humidity) the models failed to fit the data, table 10. The negative value of r^2 score could be possible attributed to the small dataset. Among the applied regressors Support Vector Machine seemed to perform better (r^2 score: -6%, RMSE: 69.06), {appendix code_10} , figure 38. What can be seen is that on the one hand random undersampling managed to provide symmetry to the data but the resulting small dataset both in wildfire and Montesinho dataset was an insurmountable drawback for fitting the data.

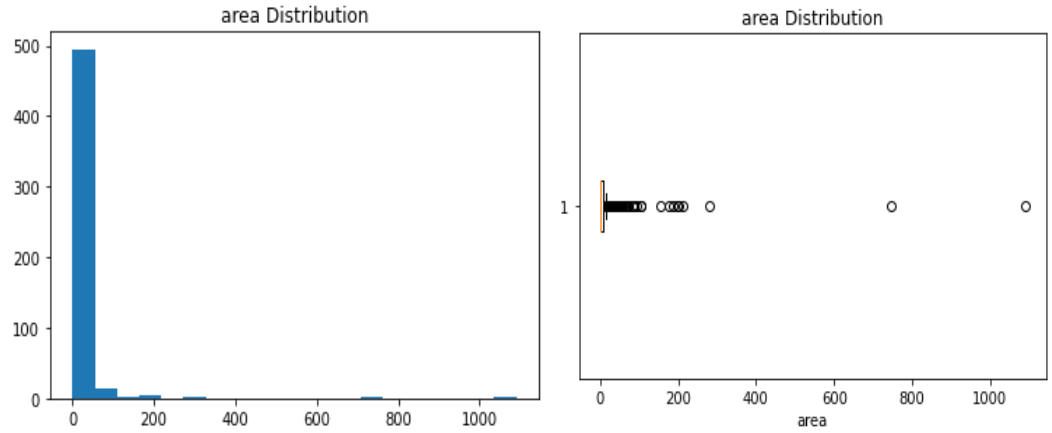


Figure 37 : Imbalanced data distribution of Montesinho dataset for forecasting the size of burned area.

Montesinho_model	r^2 score	RMSE	Normalized_RMSE
SVM	-0.059	69.065	0.06
Linear Regression	-0.221	74.14	0.07
Neural Network	-0.213	73.90	0.07
RF	-0.930	93.34	0.08
XGBoost	-1.341	102.67	0.09
Knn	-2.01	116.42	0.11

Table 10 : Overall performance for each individual model using 4 weather variables by Montesinho dataset for predicting the size of burned area.

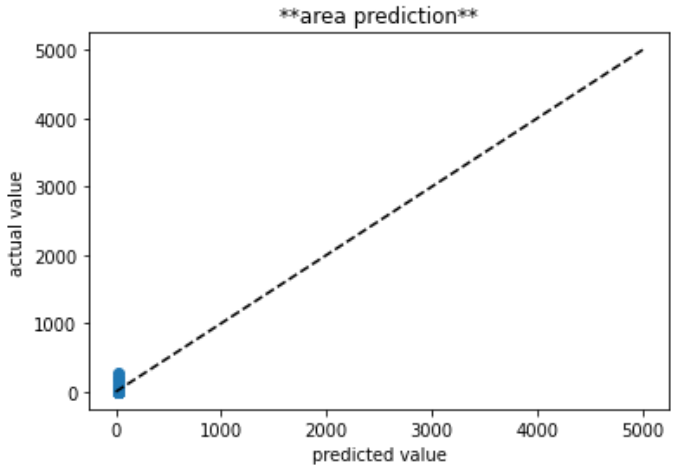


Figure 38 : Actual burned area and burned area predicted by the SVM model in Montesinho dataset.

CHAPTER 5 -- Conclusions and future challenges

5.1 Conclusions

This chapter summarizes the contribution of this postgraduate thesis investigating the influence of dynamic weather variables on fire occurrence probability, prediction of fire scale and the size of burned area in Attica basin. The deployment of the present study was based on the idea of applying as much as possible supervised Machine Learning algorithms concerning the task each time so that can be proved a useful guide for future research. A synthesis of a new dataset named "wildfire" was done including the exact prevailing weather conditions during the occurrence of forest fires for the period 2010-2019 and for the months from May to August in Attica basin. Three specific problems in a row were modeled using this concrete wildfire dataset 1) a binary classification (experiment one) where the scope was the prediction of fire occurrence probability 2) a multiclass classification work (experiment two) where the purpose was the fire scale prediction 3) a regression task (experiment three) where the main objective was the forecast of the size of burned area of forest fires. Then a comparative analysis was conducted to extract the best supervised Machine Learning models in relation with their performance. In the context of the research to analyze whether the quality of the created wildfire dataset was considered satisfactory, a comparison was made with the known prior art Montesinho dataset grounded always on weather features.

Firstly, taking all the weather variables (mean, minimum, maximum- temperature/ mean, minimum, maximum- relative humidity/ mean, minimum, maximum-(atmospheric) pressure/ daily rainfall/ mean wind speed and wind gust) into consideration (1400 observations) quite impressive results were acquired for binary classification and regression tasks respectively. In experiment one using as performance metrics accuracy, precision (directly correlated to the rate of false alarms in forecasting fire occurrences), recall (directly correlated to the rate of failed to predict fire outbreaks), the configured tuned Random Forest model (RandomizedSearchCV, GridSearchCV) achieved accuracy of 70%, precision of 67% and recall of 69%. Additionally, of the whole 12 weather variables in Random Forest model, the mean, minimum relative humidity and wind gust were of particular importance. Furthermore, in experiment two for implementing fire scale prediction (small fire, medium fire, large fire, wildfire), due to the fact the wildfire dataset appeared imbalanced a random undersampling technique was carried out limited the number of observations down to 60. An extra metric of F1 score was introduced in order to assure the correct comparison among the supervised Machine Learning algorithms providing as a best model the K-nearest neighbors (K=4) with F1 score up to 45%. Even more sensational in case of experiment three regardless the imbalanced dataset and by executing once again a random undersampling (101 observations) the K -nearest neighbors model (K=5) outperformed the others achieving r^2 score value of 70% and RMSE of 1982,59. In the mentioned third experiment it was clearly shown that despite the small under-sampled dataset the K nearest neighbors model managed to deliver significant efficiencies. All the above experiments demonstrated both the reliable quality of the newly created wildfire dataset and that the random undersampling

technique contributed positively to the prediction of fire scale and size of burned area of forest fires correspondingly.

Thereinafter, in first experiment for different combination of most relevant weather inputs variables a Sequential Forward Selection algorithm was employed for choosing the best 4 weather features (minimum temperature, minimum relative humidity, daily rainfall, mean wind speed) with the view to conducting a binary classification. According to the well-founded results the XGBoost classifier proved to outperform other classification models with accuracy of 67%, precision of 65% and recall of 65%. It is noteworthy that for the given 4 best weather variables the neural networks although had lower accuracy (64%) and precision (57%) showed much better sensitivity (86%). Respectively the known prior art Montesinho dataset for 4 weather variables (temperature, relative humidity, rain, wind) similarly proved that the XGBoost classifier performed better in terms of accuracy (56%) and precision (58%) but not in recall (41%) where Random Forest achieved 50%, (same accuracy 56% and lower precision 56%). One of the paradoxes was pointed out was the fact that tuned XGBoost classifier (GridSearchCV) in wildfire dataset deteriorated the results instead of improving them, (accuracy 66%, precision 64%, recall 62%).

In the event of predicting the size of the burned areas, the application of the Sequential Forward Selection algorithm computed as the best 4 weather variables the minimum temperature, the average relative humidity, the minimum relative humidity and the daily rainfall. In addition to the fact that due to the imbalance of the wildfire dataset a random undersampling was carried out again, the final results showed that the models (Linear Regression, Random Forest, Support Vector Machine, Neural Network, K -nearest neighbors) failed to fit the wildfire dataset. Even in the manual selection of the 4 weather variables (mean temperature, mean relative humidity, daily rainfall, mean wind speed) to be comparatively closer to the known Montesinho dataset there was a slight improvement in Linear Regression (r^2 score value of 2%, RMSE of 3584.70). Similarly, for the Montesinho dataset with 4 selected weather variables (temperature, relative humidity, rain, wind), a random undersampling was executed to provide symmetry to the dataset (observations limited from 517 to 85) but regression models (Linear Regression, Support Vector Machine, Neural Network, Random Forest, Extreme Gradient Boosting, K -nearest neighbors) flunked to fit the referred dataset. Somehow the SVM model seems to have better performance than the rest (r^2 score value of -6%, RMSE of 69.065). It is obvious that reducing the number of data to achieve symmetry in the dataset played a significant role in the performance of the models of predicting the size of the burned area for both wildfire and Montesinho datasets. Another possible cause for those poor outcomes could be the possible low correlation of weather variables with fire in terms of predicting the size of burned forest areas.

Forecasting fire outbreaks and the size of burned area according to the number of used weather variables was a continuous challenging task for the current study. In order to verify the results of [5],[6] that the 2 specific variables, the relative humidity and the cumulative

precipitation suffice to accurately predict of fire occurrence, a Sequential Forward Selection algorithm applied in wildfire dataset to extract the best 2 features and to check the outcomes achieved. Indeed, the minimum relative humidity and the daily rainfall were chosen as the best 2 weather variables. That is, the 2 main weather features that essentially determine the periods of drought in the Attica basin. Prolonged drought period means that the area under study is considered more vulnerable to fires. In this case study, it seems that the Neural Networks performed better accuracy (63.6%) and precision (63%) in relation to the Random Forest (63.3% - 61%) but not better sensitivity (NN: 54%, RF: 61%). Conclusions that show that Neural Networks and Random Forest models took into account the non-linear relationships between the independent weather variables, managing to create patterns between these weather features and fire occurrence probability so that they were able to generalize well.

5.2 Future challenges

The variety of alternative approaches presented in this study prepare the ground for further improvements and extensions in solving the problems considered. Taking into account the positive results obtained from the best applied supervised Machine Learning models on a case-by-case basis, the interaction of these models with variables other than weather could be investigated. A future project can be further expanded on other under examination factors such as topology, vegetation, the time of intervention and extinguishing, the starting point of the fire outbreak, the form of fire but mainly the human factor and activities, especially

- The topographic elements: The anaglyph of the Attica basin area contributes greatly to the spread of the fire, as there is intense mountain formation (Aigaleo, Parnitha, Hymettus, Penteli). In particular, the elevation associated with the vegetation and humidity as well as the slope of the ground affect seriously the speed of fire spread. Elevation and terrain slope are important factors for the evolution of forest fires, as they affect the growth of tree vegetation, but also determine the topoclimate. It seems, then, that the topography information of Attica basin location may create a potentially dangerous situation for forest fires and, therefore, the analysis of the above parameters is necessary to be considered in the analysis of fire data.
- Vegetation: The description of the vegetation in the geographical region of Attica leads to the conclusion that the ecosystems of the Mediterranean region are the most vulnerable to fires. It has been pointed out that the shrubby form of vegetation makes it more vulnerable to fires. Elements such as lawn density, vegetation density should be included as important causes of fire outbreaks.
- Starting point of fire and time of intervention and extinguishing: The recording of the locations where the forest fires break out, is considered essential for the organization of protection and extinguishment. These actions are determined by the intervention and extinguishment times. Statistics show that forest areas are the most common starting point for forest fires, but episodes in livestock farms are considered more catastrophic.

- The form of fire: Forest fires in relation to the way they spread and depending on the level that appear relative to the soil surface are divided into ground or underground, creeping, crowning and mixed fires. Ground fires move relatively slowly, burning the deposited biomass in the forest soil. Creeping fires cause burning mainly of plant elements of the subsoil consisting of shrubby vegetation and expanding relatively faster than the previous ones. Crowning fire spread through the canopy of trees and can carry relatively large jumps. Their expansion is quite fast. Finally, mixed fires combine all the above forms of spread and are more common.
- Human factor: Anthropogenic causes of fires (cigarette, malignant arson, pyromaniac energy, short circuit) are a field of further research and analysis. The way in which the human factor is determined according to the state of the art varies. Either as a population density of an area [12], or based on the determined distance from railways, road networks and settlements [16] or through Gross Domestic Product (GDP) density [10]. More and more studies converge to the fact that the Human-induced factors outweigh weather variables as main causes of fire occurrences [3].

Therefore, possibly a combination of alternative selected features for the Attica basin area with different supervised Machine Learning algorithms can lead to even greater accuracy in predicting fire occurrences, fire scale and size of burnt forests.

5.3 Future survey

During the study elaboration on the development of supervised Machine Learning algorithms for the prediction of fires based on weather variables, the idea of formulating a model that would be gradually released from the dependence of weather variables until their complete abolition emerged. That is to say, the generation of a model that would be trained in such a way as to be independent of sensory measurements. In essence, the deployment of a neural network that can integrate data from two sources. In the present project, a wildfire dataset was created, including the prevailing weather conditions during the fire occurrences in Attica basin. In a hypothetical scenario we could assume that this dataset was the first data source in a neural network. As a second data source we could receive a set of fire digital photographs/images corresponding to each officially recorded forest fire incident for the same period in Attica basin. When integrating this data into the neural network and training it to produce a kind of hybrid fire photo/image comprising the prevailing weather variables, figure 39. The new hybrid variables were then fed as data to a secondary neural network that would be trained based on them. So that the gradual supply of the second neural network only with digital fire photos/images to make it able to evaluate the possibility of how close a photo/image was without containing weather features to a photo/image comprising weather features. Therefore, depending on whether a threshold value was satisfied the weather variables of closer fire photo/image could be provided. This would mean the final independence from the weather variables where only by using a digital fire photo/image could the prevailing weather conditions be extracted. The recommended idea

is at an early stage and is simply formulated as a proposal for further research and implementation in the future by researchers in the field.

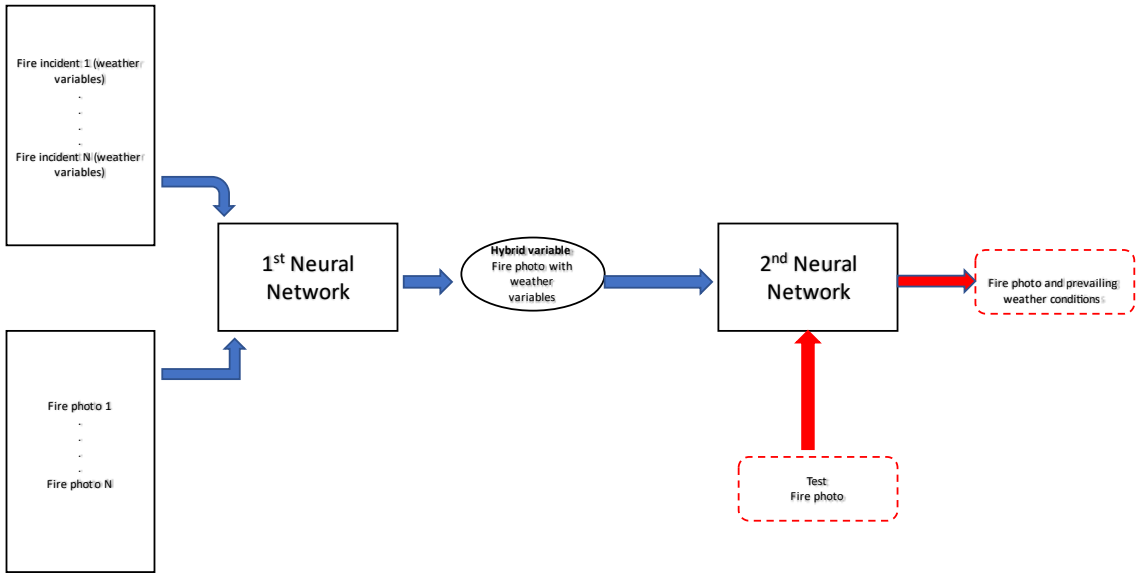


Figure 39 : Low level recommended architecture for generating a hybrid variable and a fire photo integrated with prevailing weather conditions.

Bibliography

[1] Tsagkari, Karetsos and Proutsos, "FOREST FIRE GREECE 1983- 2008," Institute of Mediterranean Forest Ecosystems and Forest Product Technology, Athens, 2011.

[2] Chariskos, "Spatial modeling and risk prediction of large forest fires in Greece," Democritus university of Thrace, department of forestry and environmental natural resources management, Orestiada, 2017.

[3] P. Jain, C. S. Coogan, S. G. Subramanian, M. Crowley, S. Taylor and M. D. Flannigan, "A review of machine learning applications in wildfire science and management," *NRC Research Press,* vol. 28, pp. 478-505, 2020.

[4] Y. Xiong, J. Wu and Z. Chen, "Machine Learning wildfire prediction based on climate data -Group 75".

[5] G. E. Sakr, I. Elhajj and G. Mitri, "Efficient forest fire occurrence prediction for developing countries using two weather parameters," *Engineering applications of Artificial Intelligence,* vol. 24, no. 0952-1976/Elsevier, pp. 888-894, 2011.

[6] G. Sakr, I. H. Elhajj, G. Mitri and U. C. Wejinya, "Artificial Intelligence for forest fire prediction," in *International conference on advanced intelligent mechatronics*, Montreal, Canada, 2010.

[7] D. Arias, F. Dorado, C. Ordonez and J. Perez, "Biophysical and lightning characteristics drive lightning-induced fire occurrence in the central plateau of the Iberian peninsula," *Agricultural and forest meteorology,* pp. 36-47, 20 May 2016.

[8] K. Blouin, M. D. Flannigan, X. Wang and B. Kochtubajda, "Ensemble lightning prediction models for the province of Alberta, Canada," *International journal of wildland fire,* vol. 25, no. Journal compilation IAWF 2016, pp. 421-432, 2016.

[9] E. Ashley, V. Beusekom, W. A. Gould, C. Monmany, A. H. Khalyani, M. Quinones, S. J. Fain, M. J. Andrade-Nunez and G. Gonzaez, "Fire weather and likelihood: characterizing climate space for fire occurrence and extent in Puerto Rico," Springer science and business media Dordecht, 2017.

[10] A. Aldersley, S. Murray and S. E. Cornell, "Global and regional analysis of climate and human drivers of wildfire," *Science of the total environment,* pp. 3472-3481, 18 May 2011.

[11] F. Guo, L. Zhang, S. Jin, M. Tugabu, Z. Su and W. Wang, "Modeling anthropogenic fire occurrence in the Boreal forest of China using Logistic Regression and Random Forests," *Forests,* pp. 1-14, 25 October 2016.

[12] L. Li, J. Ma and W. G. Song, "Artificial neural network approach for modeling the impact of population density and weather parameters on forest fire risk," *Wildland fire,* pp. 640-647, January 2009.

[13] T. Preeti, K. Suvarna, B. Aishwarya, M. Sumalata and Aishwarya S, "Forest fire prediction using Machine Learning techiques," in *2021 International conference on intelligent technologies*, Karnataka, India, 2021.

[14] D. Stojanova, A. Kobler, P. Ogrinc, B. Zenko and S. Dzeroski, "Estimating the risk of fire outbreaks in the natural environment," Springer, Kras, 2011.

[15] P. Cortez and A. Morais, "A data mining approach to predict forest fires using meteorological data," Guimaraes, 2007.

[16] Y. Xie and M. Peng, "Forest fire forecasting using ensemble learning approaches," in *The natural computing application forum 2018*, 2018.

[17] M. Bisquert, J. Sanchez-Thomas, E. Caselles and V. Caselles, "Application of artificial neural networks and logistic regression to the prediction of forest fire danger in Glicia using MODIS sata," *Wildland fire,* pp. 1-7, 2012.

[18] C. Vasilakos, K. Kalabokidis, J. Hatzopoulos, G. Kallos and Y. Matsinos, "Integrating new methods and tools in fire danger rating," *International journal of Wildland fire,* no. 16, pp. 306-316, 2007.

[19] A. Tohidi, N. McCarthy, Y. Aziz, A. Wani and T. Frank, "Fire forecasting," 2020.

[20] W. Minglang, "Modeling method of mountain fire prediction based on in-depth ntework learning," 2019.

[21] W. Dan, "Forest fire early warning method based on weather and remote sensing data," 2020.

[22] B. White, L. Jing and A. Adrian, "Method and system for predicting wildfire hazard and spread at multiple time scales," 2021.

[23] L. Jingsong, Y. Hua, Y. Shu, J. Min, W. Shuai, Y. Hui, J. Tao, M. Kangmin , L. Zhumao, N. Biao, Z. Wei, Z. Ziqiang and C. Shengzhi, "Power transmission line forest fire risk grade forecasting method based on gradient boosting tree," 2021.

[24] G. Shuchang, Y. Yi, L. Peng, G. Ruhui and Y. Zhida, "Convolutional neural network-based lightninh frequency forecasting method," 2021.

[25] G. Watt and E. Amjadian, "System and method for weather dependent Machine Learninng architecture," 2021.

Appendix

- **Code_1**: Implementation of RF, SVM, Knn, DT, XGBoost, Logistic Regression for 12 weather variables retrieved by wildfire dataset for conducting binary classification. Also, RF for the 2 best selected weather variables.

- **Code_2**: Implementation of Neural Networks for 12/2 weather variables retrieved by wildfire dataset for conducting binary classification.

- **Code_3**: Model tuning and hyperparameter optimization for RF with 12 weather variables and Neural Network for 2 weather variables, respectively.

- **Code_4**: Implementation of RF, XGBoost, SVM and Neural Network for the best 4 selected weather variables for conducting binary classification, XGBoost Model tuning.

- **Code_5**: Implementation of RF, XGBoost, SVM, Neural Network for 4/2 selected weather variables retrieved by Montesinho dataset for conducting binary classification.

- **Code_6**: Implementation of Knn, DT, SVM, RF, XGBoost for 12 weather variables retrieved by wildfire dataset for conducting prediction of fire scale.

- **Code_7**: Implementation of Knn, Linear Regression, SVM, RF, Neural Network for 12 weather variables retrieved by wildfire dataset for conducting prediction of size of burned area of forest fires.

- **Code_8**: Knn best regressor for 12 weather variables retrieved by wildfire dataset for conducting forecast of burned area.

- **Code_9**: Implementation of Knn, Linear Regression, SVM, RF, Neural Network for the best 4 selected weather variables retrieved by wildfire dataset for conducting forecast of burned area.

- **Code_10**: Implementation of Knn, Linear Regression,SVM, RF, XGBoost, Neural Network for the 4 selected weather variables retrieved by Montesinho dataset for conducting forecast of burned area.