People's Democratic Republic of Algeria
Ministry of Higher Education and Scientific Research

University of Abderrahmane Mira - Bejaia
Exact Sciences Faculty
Computer Science Department

MASTER PROFESSIONAL THESIS

In

Computer Science

Option

Software Engineering

Theme

Towards forest fire prediction in Algeria

Presented by

MERAD Chaïma

SACI Amina

Defended on September, 2022. Examining Committee composed of:

Supervisor	Mrs EL BOUHISSI Houda	MCA	U. A/Mira Bejaia
Examiner 1	Mrs ALOUI Soraya	MCA	U. A/Mira Bejaia
Examiner 2	Mrs BATTAT Nadia	MCB	U. A/Mira Bejaia

Bejaia, Sept 2022.

Dedications

“

TO my dear parents, for all their sacrifices, their love, their tenderness, their support and their prayers all throughout my studies,

TO my dear sister Amira, for her endless encouragement, and her constant support.,

TO my dear brother Djawad, for her support and constant motivation.,

TO all my family, my friends and all those who have helped me and contributed to near or far to carry out this work. TO my partner and friend Amina, with whom I took a lot of fun to work.

”

- Chaïma

Dedications

I dedicate this modest work

To my dear mom, Lila who always encouraged and supported me with these beautiful words in gold and these precious douaas, thank you for always being by my side mom.

To my dear dad, Mohamed who has always been the example of a magnificent father for me, with all this advice and encouragement, and his confidence in me, thank you for being the best dad.

To my saida tent, which I have always considered as a second mom for me, thank you for all.

To my sister litycia, your presence has always been a strength for me to continue to succeed thank you my best.

To my little brothers amine and nassim who gave me the example of the true brothers of the world with their patience and encouragement.

To my partner, and my friend Chaïma, that I took great pleasure in working with her and sharing this wonderful work.

To all my cousins kamy and sonia who always help me with their encouragement

To all my friends houda, feriel, zina, warda, nihad and ryma for their support and love And to all who have intervened in the realization of this work either from near or far thank you for all.

Amina

Acknowledgement

First of all, we thank God the Almighty for giving us the strength to survive, patience, will and courage during these long years of study.

We would like to express our gratitude to **Pr. EL BOUHISSI Houda**. We thank her for the quality of her exceptional supervision, her rigor and availability, her help and her advice.

We extend our sincere thanks to all the professors, speakers and all the people who, through their words, writings, advice and criticisms have guided our reflections and agreed to meet us and answer our questions during our research.

We would also like to thank each of the members of the jury for their interest in this work and for agreeing to evaluate it.

Abstract

Forest fires are one of the most frequent subjects and address in recent years in the world generally and in Algeria in particular, it is considered among the greatest dangers that have threatened the world in recent years and as the Algeria is among the countries most affected by these fires, the significant risk rate caused by them has generated remarkable concern by the stakeholders and which calls into question the means of protection against these strong fires and their usefulness to protect the environment of the country from these disasters.

The field of prediction is such a vast field and several techniques could be identified in an effective way in this field and the most popular are those which are based on artificial intelligence and among these methods Machine learning and these different tools and models which make it possible to achieve reliable and robust systems in different discipline of life and especially better known in the field of prediction and in our case for the prediction of forest fires.

This thesis is a summary of the different aspects and approaches related to forest fires and proposes a new approach to solve this problem by using one of the machine learning methods which is logistic regression, explaining to us this approach which is sacred to the prediction of forest fires in the two regions of northern Algeria Bejaia and Sidi Bel Abes using a set of data obtained from a site called kaggle, analyzing this dataset, building our system based on the chosen model and finally having an effective prediction system The power of computer systems today has opened new doors for new solutions in the different areas of life but also in the field of forests and these risks. So forest fire prediction systems are among the most reliable and useful solution today to fight against these fires.

Keywords : Machine learning, Kaggle, datasets, forest fire prediction, Logistic regression, Intelligence artificial.

Résumé

Les incendies de forêt sont l'un des sujets les plus fréquents et abordés ces dernières années dans le monde en général et en Algérie en particulier, il est considéré parmi les plus grands dangers qui ont menacé le monde ces dernières années et comme l'Algérie est parmi les pays les plus touchés par ces incendies, le taux de risque important qu'ils occasionnent a suscité une inquiétude remarquable de la part des intervenants et qui remet en cause les moyens de protection contre ces forts incendies et leur utilité pour protéger l'environnement du pays de ces catastrophes.

Le domaine de la prédiction est un domaine tellement vaste et plusieurs techniques pourraient être identifiées de manière efficace dans ce domaine et les plus populaires sont celles qui sont basées sur l'intelligence artificielle et parmi ces méthodes Machine learning et ces différents outils et modèles qui permettent pour réaliser des systèmes fiables et robustes dans différentes disciplines du vivant et surtout mieux connus dans le domaine de la prédiction et dans notre cas pour la prédiction des feux de forêt.

Cette thèse fait une synthèse des différents aspects et approches liés aux feux de forêt et propose une nouvelle approche pour résoudre ce problème en utilisant une des méthodes d'apprentissage automatique qui est la régression logistique, nous expliquant cette approche qui est sacrée à la prédiction des feux de forêt. les incendies dans les deux régions du nord de l'Algérie Bejaia et Sidi Bel Abes en utilisant un ensemble de données obtenues à partir d'un site appelé kaggle, en analysant cet ensemble de données, en construisant notre système basé sur le modèle choisi et en ayant enfin un système de prédiction efficace La puissance des systèmes informatiques aujourd'hui a ouvert de nouvelles portes pour de nouvelles solutions dans les différents domaines de la vie mais aussi dans le domaine des forêts et de ces risques. Ainsi, les systèmes de prévision des incendies de forêt sont aujourd'hui parmi les solutions les plus fiables et les plus utiles pour lutter contre ces incendies.

Mots clés : Apprentissage automatique, Kaggle, ensembles de données, prédiction des incendies de forêt, régression logistique sion, Intelligence artificielle

---

Contents

List of Figures

List of Tables

Abreviations list

IT	*Information Technologie*
AI	*artificial intelligence*
PDF	*Portable Document Format*
SHP	*Shape*
CSV	*Comma Separated Values*
SPSS	*Statistical Package for the Social Sciences*
XLSX	*Excel Xml Spreadsheet*
XLS	*Excel spreadsheet*
DD	*Day*
MM	*Month*
YYYY	*Year*
Temp	*temperature*
RH	*Relative Humidity*
Ws	*Wind speed*
FFMC	*Fine Fuel Moisture Code*
DMC	*Duff Moisture Code*
DC	*Drought Code*
ISI	*Initial Spread Index*
BUI	*BuilUp Index*
FWI	*Fire Weather Index*
GUI	*Graphical User interface*
OS	*Operating System*

JSON	*JavaScript Object Notation*
SQL	*Structured Query Language*
TXT	*Text*
TK	*Tool Kit*
Tcl	*Tool Command Language*
API	*Application Programming Interface*
MATLAB	*MATrix LABoratory*
NUmPy	*Numerical Python*
2D	*2 Dimentional*
DM	*Data Mining*
ANN	*Artificiel Neural Network*
CNN	*Convolutional Neural Network*
SciPy	*Scientific Python*

Chapitre 1

General introduction

1.1 Indroduction

Forest fires are the most common hazard in forests. They constitute a threat not only for the forest richness but also for the whole of the fauna and flora regime, and cause a very serious imbalance of the biodiversity and the ecology of a region.

According to a study by the Food and Agriculture Organization of the United Nations (FAO) : "about 5 percent of the world's forest area is affected by fires each year". Unfortunately, the Mediterranean basin is no exception, as more than 55,000 fires per year cover an average of 500,000 to 700,000 hectares of forest, causing enormous ecological and economic damage and loss of life.

Algeria is the tenth largest country in the world by area and the largest country on the African continent after Sudan, with a total area of 2,381,741 square kilometers, a geographical location that endows it with a special climate and ecological diversity. Algeria favors the development of a very rich and diverse flora, in fact one of the most diverse and pristine in the Mediterranean basin.

However, the region is experiencing very severe soil degradation and desertification representing a late stage, the degradation of this natural environment is manifested in the reduction of biological potential and disruption of ecological and socio-economic balance.

The Mediterranean region, Algeria is one of the countries where the scientific community knows very little about the forest fire problem. These fires are particularly destructive due to the scarcity of forests and the threat of desertification. Algeria alone has 4.1 million hectares of forest, with an afforestation rate of 1.76 percent. However, the cluster of fires that occurred one after the other at intervals of less than 10 years had a catastrophic impact on the ecological scale.

Forest fires in Algeria have damaged more than 32,000 hectares tracks annually despite the various prevention and control plans put in place, the situation is getting worse and worse every year and is causing forest degradation and directly affecting the environment and the local economy. conventional fire prevention methods and the control does not respond reliably to this problem with a gigantic consumption of time and a hyper high complexity rate which poses great disadvantages to its use and makes the work more complicated as possible.

The various conventional methods of fire prevention used in Algeria take a long time and are not always reliable. This field requires observation techniques and analysis such as remote sensing and geographic information system, these techniques are very effective and fast in the development and risk assessment and fires in forest areas.

Forest fires are one of the major risks that we regularly face, their control is very difficult and they can have devastating effects. Forest fire prevention and detection strategies have varied over the years. Nevertheless, these techniques remain not very effective because of the very important response time and their complexity, which shows the need to develop new techniques that respond to these shortcomings. Currently, after the evolution of technology and information systems, the exact prediction of forest fires depends on the

remote sensing system, which plays a very important role in the detection of this great phenomenon.

In this thesis, we studied the most important works regarding forest fires, their causes, and their impact on the life of living beings. In addition, we have proposed an effective solution that can contribute to resolve this big problem, which affects the forests of our country.

1.2 Problem statement

Every year Algeria suffers from a dark period with forest fires that threatens the safety of our country. Despite all the efforts made to avoid them by the protection services and up to now this problem remains a great risk, which causes the environment of the country and even the security of its inhabitants.

The damage and danger left behind by these fires worries officials and associations in the country who are trying to find immediate solutions to put an end to this disaster by providing all the necessary equipment.

The experience of all these years proves that despite the immediate intervention of the protection services, it still generates a significant rate of damage ; the country is still threatened by these forest fires.

For this reason, building a forest fire prediction system seems a very good solution, to prevent the risks and reduce the damages.

1.3 Objective and contribution

The objective of this dissertation is mainly to predict forest fires in Algeria. By developing a prediction system to classify the possibility of forest fire into two categories : fire, non-fire and represent the different techniques and approaches proposed and their results.

Our goal is to implement a prediction system based on supervised machine learning, to predict forest fires. This model is evaluated on the basis of a series of forest fires in Bejaïa and Sidi-Bel-Abbes taken in recent years.

1.4 Work Methodology

Our work focuses on a set of steps, which are presented in the following :

Research and analysis : Concerns establishing a state of the art of the most important studies proposed in the field of forest fire prediction systems. A comparative study of these studies is addressed.

Problem and solution identification : Involves the problem statement and the appropriate solution.

System implementation : Refers to the implementation of the proposed system and its different functionalities.

1.5 Thesis organization

The remainder of this thesis is organised as follows :

The second chapter : is devoted to the fields of prediction and machine learning. We will present the different levels of prediction, problems related to prediction and maching learning for forest fire prediction.

In the third chapter : we overview the most important related works according to forest fire prevention. we will present this in a table that contain the outline of each synthesized document, then we will proceed to an analysis comparative approaches to related documents and our approach.

The fourth chapter : focuses on experimenting with our proposal. It presents the different aspects related to the implementation of system that we will develop.

The fifth chapter : we present the programming tools and the implementation of our application, presentation of the interfaces and the execution results, as well as the software chosen for the implementation of our approach.

The sixth chapter :concludes the thesis and opens other perspectives for future work.

Chapitre 2

General information on prediction

2.1 Introduction

In recent years, data has become a key driver of economic growth and the foundation on which industries are built. For many, knowing what to do with this complex raw material is now a major organizational challenge. This is where actuaries Intervene. By relying on predictive modeling, they can help turn big data into big business.

Predictive modeling analyzes datasets to identify significant interdependencies. It then resorts to these interdependencies to better predict outcomes and make informed decisions more quickly and deployable. It relies on historical information to describe past interdependencies from which to draw trends for the future. These trends can be applied to several areas, such as forest fire prediction.

Predictive modeling is inspired by many disciplines, including statistics, modeling, optimization, cluster analysis, study market and computer programming. Its application rests usually on considerable computing power and overlaps areas such as self-learning and artificial intelligence.

Predictive modeling is used in the most diverse fields and specialties : the insurance sector, the financial sector, telecommunications, science, e-commerce, customer relationship management or business intelligence. For economic applications, these prognoses can be used as decision-making bases in budget planning and the assessment of opportunities and risks. A well-known use of predictive modeling is the calculation of risks in the context of life insurance. Rather, in the scientific context, it is a matter of confirming or refuting theories that describe the behavior of an object in a specific specialty area using data.

Predictive modeling is a mathematical process that aims to predict future events based on past behavior. This is the main function of predictive analytics applications ; it works by analyzing data to identify patterns, then using those patterns to create algorithms that data scientists train. Predictive modeling is used in a wide variety of applications, from meteorology to fraud detection and security management.

There are many different modeling methods and algorithms. Some of the most popular include : decision trees, time series analysis, neural networks, linear regression, and logistic regression.

In this second chapter we will present some definitions of the field of study that is predictive modeling, its types, its difficulties, the problems related to this field and Machine Learning and its types.

2.2 Prediction

Predictive modeling is a set of methods to collect and analyze defined data, so as to interpret it to deduce predictions about future trends, future events or consumer behavior in the future. Predictive modeling thus gives rise to prognoses that must however be considered as probabilities and not as certain predictions, which will necessarily materialize. The plausibility of the results of the prognostic models is to be reported statistically.

Their probability is then possible depending on the size of all the data studied. Thus, the greater the number of data analyzed, the more the results of the prognostic models can be considered as possible and accurate results. However, there is no guarantee as to the actual occurrence of the output data.

Several fields and various specialties regularly use predictive modeling, such as the insurance sector, the financial sector, e-commerce, telecommunications, science, customer relationship management or business intelligence. Predictive modeling also gives rise to prognoses that can be used as decision-making bases for assessing opportunities and risks in the context of budget planning. Predictive modeling is thus commonly used to calculate risks in the context of life insurance. In the scientific field, this science is used to confirm or refute theories using data from a specific area of expertise.

Generally speaking, predictions can be made using a large number of different regressions analyzes and statistical models. When predictive modeling is applied to the IT field, it includes data mining and machine learning. It is then necessary to extract relevant input data in the form of large data sets, and on the other hand, to establish prognostic models capable of self-learning, which are then able to integrate automatically new data to the forecasts already established.

2.3 Machine Learning

Machine learning can be defined as an artificial intelligence technology that allows machines to learn without having been previously programmed specifically for this purpose.Machine learning is explicitly linked to Big Data, since in order to learn and grow, computers need data streams to analyze, on which to train.[13]

The first machine learning algorithms are not new, since some were designed as early as 1950, the best known of them being the Perceptron.[20]

Machine learning is a modern science that allows you to discover patterns in one or more data streams and make predictions based on statistics. Clearly, machine learning is based on data mining, allowing pattern recognition to provide predictive analysis. Machine learning systems are divided into two main categories : supervised and unsupervised systems.

2.3.1 Supervised Learning

Machine learning algorithms that implement learning per pair of inputs/outputs. Is to teach an algorithm to come to a specific conclusion based on historical data. For example, if the question is ”is this customer going to leave ?”, [17] an analyst can look at historical data about customers the company has lost before, and then train an algorithm to determine which customers are most likely to leave based on that data. To do this, the analyst creates a training data set with a known result (i.e. a lost or not lost customer) that the algorithm then uses to create a predictive model based on historical data.

Currently, supervised machine learning is the most widely used method and often the most effective. This type of learning corresponds to our problem, because we know the output data.

Supervised learning consists of establishing rules of behavior at from a database containing examples of cases already tagged. More specifically, this database is a set of couples entered- outputs (Xi, Yi) 1 i n random. The objective is then to learn to predict, for any new X input, the Y output.

This method is carried out in two phases : the learning phase and the inference phase. During the learning phase, the model optimizes its parameters in order to match the input and output we provide. During the inference phase, the Model uses previously optimized parameters to predict output from a provided input. The model can learn two types of outputs : continuous (regression model) or discrete (classification model).

During a classification, the different classes represented by the output are categories. In the case of a regression, the output can take a continuous set of values. In the case of a classification, the output of the model will be a vector of the size of the number of classes and the selected class will be the one with the highest value in the vector. In the case of a regression, the output will be directly the expected value.

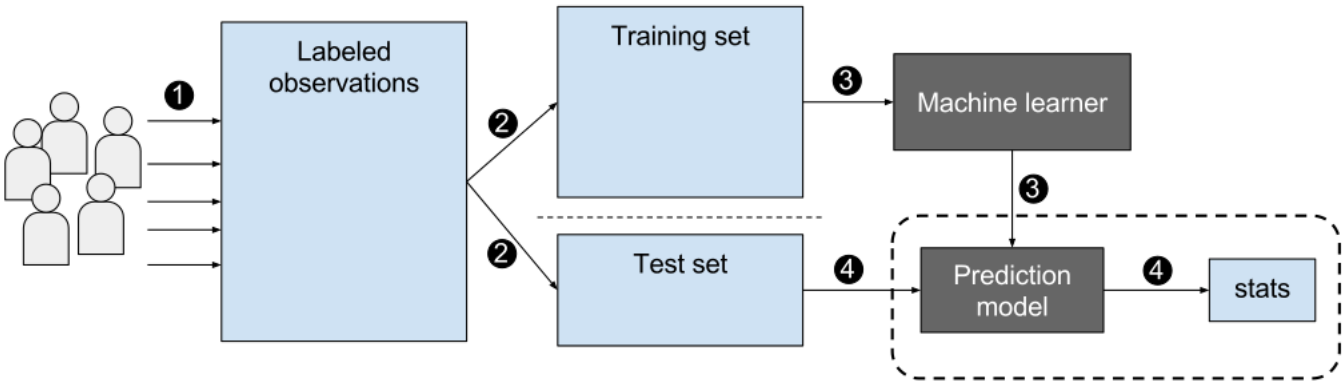

Fig. 2.1 : Supervised learning workflow.

2.3.2 Unsupervised Learning

Is teaching an algorithm to look for similarities or trends in data and group things together based on that information, without being told what to look for.

Unsupervised learning theory deals with the case where only the Xi1 i n inputs are available, without the outputs. The most important problem then consists of partitioning the data. The aim is to group the observations into different groups in such a way that the data of each subset share common characteristics.

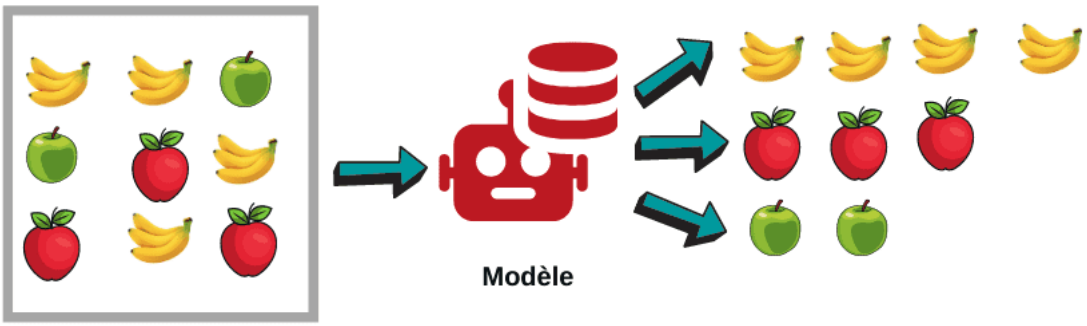

FIG. 2.2 : Unsupervised learning. [9]

2.3.3 Semi-supervised Learning

Semi-supervised learning methods combine labeled and unlabeled data. Algorithms of this type feed on certain information through labeled categories, suggestions and examples. Then they create their own labels by exploring the data on their own, following a rudimentary scheme or the indications of data scientists.[25]

2.3.4 Reinforcement learning

Reinforcement learning algorithms are based on reward and punishment systems. The algorithm is assigned a goal and seeks to get closer to it to get a maximum reward. He relies on limited information and learns from his previous actions. These algorithms can depend on a schema (a model) ; they must then follow predefined steps and the number of errors and trials is limited. Others do not rely on a diagram and interpret with each new attempt.

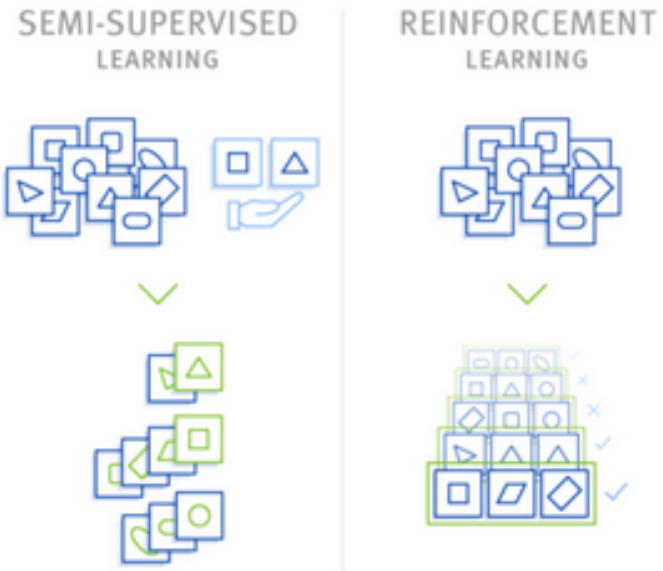

FIG. 2.3 : Semi-Supervised and Reinforcement Learning. [8]

2.4 Types of Predictive Modeling

Predictive models use known results to develop (or train) a model to predict the values of different or new data. Modeling gives results in the form of predictions that represent a probability of the target variable based on the estimated weight of a set of input variables.

There are two types of predictive models. Classification models predict class membership. For example, you try to find out if a person is likely to leave, if they will react to a solicitation, whether or not they have a credit risk, etc. Typically, the results of the model are of type 0 or 1, with 1 denoting the event you are targeting. Regression models predict a number, such as the revenue a customer will generate in the coming year or the number of months after which a component will fail. The three most commonly used predictive modeling techniques are decision trees, regression, and neural networks.

2.4.1 Decision trees

Decision trees are classification models that divide data into subsets according to categories of input variables. They make it possible to understand a person's reasoning and are in the form of a tree structure, each branch of which constitutes a choice between several alternatives and each sheet, a classification or a decision. This model looks at the data and tries to find the variable that divides it into the most different logical groups possible.

A supervised algorithm that predicts a category based on historical data. Easy to understand and interpret, decision trees are widely used, they handle missing values correctly and are very convenient for preliminary sorting of variables.

2.4.2 Regression

A supervised algorithm that predicts a value or number based on historical data. Example : Based on location, area and other factors, a regression algorithm can predict the value of a home.

Regression (linear and logistical) is one of the most common methods in statistics. Regression analysis evaluates the relationships between variables. Designed for continuous data that is supposed to follow a normal distribution, it detects key trends in large datasets and is often used to determine the influence of specific factors.

With regression analysis, it is a question of predicting a number, called a response variable or variable Y. In linear regression, an independent variable is used to explain and/or predict the outcome of Y. Multiple regression uses two or more independent variables to predict this outcome.[18] In logistic regression, unknown variables of a discrete variable are predicted based on the known value of other variables. The response variable is categorical, which means that it can only take a limited number of values. In binary logistic regression, a response variable has only two values, of type 0 or 1. In multiple logistic regression, a response variable can have several levels, for example low, medium and high, or 1, 2 and 3.

2.4.3 Neural networks

Neural networks are sophisticated techniques capable of modeling extremely complex relationships. Their power and flexibility make them successful, thanks in particular to their ability to manage the non-linear relationships of data, which are becoming more frequent with the increase in the volume of data collected. They are often used to confirm results using simple techniques, such as regression and decision trees.

Neural networks are based on the recognition of trends and on certain artificial intelligence processes that "model" parameters in graphical form. They work well in the absence of known mathematical formulas to relate inputs to outputs, in cases where the forecast matters more than the explanation, or when there is a lot of training data.

2.5 Challenges When Deploying a Predictive Analysis Model

Because predictive analytics is a sophisticated capacity, implementing it is likewise complicated and difficult. Companies that adopt a traditional approach to predictive analytics (that is, treating it like any other sort of analytics) frequently run into roadblocks.

2.5.1 Data-Related Challenges

Data allowing predicting intent can come from a variety of data sources : Large structured data from backend systems with detailed attributes and functionality. Unstructured data from social media. This complexity can create some challenges when deploying a predictive model. To overcome them, the predictive analytics process must ensure the quality and quantity of training data.

2.5.2 Data Quality / Cleansing

Errors such as inconsistent data, duplicates, logical conflicts, and missing data can appear when uploading data to a data warehouse database. Data consolidation should eliminate noise.

2.5.3 Over-Cleansed Data

If there are strict rules and the data is over cleaned, the data does not match the actual data, making the training data inaccurate and the prediction model unreliable.

2.5.4 Old Data are Out of Date

Most of the time, user master data, such as customers, appears frequently in backend systems. These can be separate entities that point to the same client. Old data should be marked as old and kept out of reach, since forming a model with old data can lead to incorrect conclusions about predictions.

2.5.5 Not Enough Data

The reliability of the model depends greatly on the size of the training data set. according to this rule, the larger the training dataset, the greater the reliability. This is a basic machine learning challenge called cold start, where the model is not effective at first due to lack of data. This can be found in the case of a new product, a new group of customers, etc.

2.5.6 Too Much Data

Statistically, at some point, providing the model with data does not increase the accuracy of the predictions. This can spend unnecessary resources and computational time.

2.5.7 Overestimate Surveys

Predictive models typically rely on data from surveys or submitted forms. However, the success rate of the surveys is low, which can lead to unrepresentative training data and an unreliable prediction model.

2.5.8 Model Related Challenges

The choice of model type, and the choice of input functions, is main to the construction of a predictive model.

2.5.9 Model Complexity

To cover a wider range, models can incorporate a lot of input variables building a complex and expensive model, and therefore difficult to monitor and adjust, making them powerless to predict outcomes. The most important thing is not the quantity but rather the precise selection of input variables which requires a thorough understanding of the domain knowledge, objectives and data on which the model will be executed.

2.6 Machine learning as a fire prediction tool

Today, firefighting professionals are making an alarming observation : the ever-increasing number of fires, ever more intense, have become very difficult to control. We are talking about mega fires, almost impossible to put out.

Faced with these phenomena that have become unmanageable, prevention becomes the ideal solution : by predicting the occurrence of fires, we can anticipate their management and control them before they degenerate, while taking into account their natural role in regulating vegetation.

Just as humans learn from experiences, AI models rely on learning from past events. First, the researchers give the algorithm a history of data describing the characteristics of past fires as well as the weather and vegetation conditions associated with them. Using mathematical principles, and relying on this set of data, the algorithm will look for correlations between these environmental factors and the different types of fires. He thus builds what is called a model, which will serve as a reference for scientists to make predictions about future fires. To do this, they provide the model with forecasts and estimates of weather and plant conditions in the areas observed. On this basis, the AI returns the probability of a fire occurring, as well as its characteristics.

Through these predictions, firefighting teams will be able to adapt their interventions in the field.

2.7 Conclusion

Nowadays, research on predictive modeling is very important. Most domains create different types of data and need Analyze this data to predict and learn from it and make decisions that are beneficial.

In this chapter, we have presented what predictive modeling is, its types, machine learning and its different types, and finally talk about the challenges when deploying a predictive analytics model.

In the next chapter, we will draw up a state of the art of the main approaches relating to the domain of the prediction of forest fires. A comparative study will be prepared for the main work already done.

Chapitre 3

Stat of the art

3.1 Introduction

Forest fires are a global disaster, causing economic and ecological damage and threatening human life. Millions of hectares of forests are destroyed every year around the world.

Algeria is one of the countries affected by this phenomenon. Deforestation and climate change have led to an increasing number of forest fires in recent decades. These fires have a major impact on the climate, causing widespread health, social and economic problems.

Existing wildfire forecasting systems rely on hand-crafted features and require expensive field instrumentation and maintenance. Wildfire detection and prediction is becoming a very important issue in avoiding the devastation caused by this disaster. Researching new fire detection and prediction methods as an alternative to older methods is becoming an emergency that will provide an automated and intelligent way to detect wildfires without human intervention.

3.2 Related works

Forest fire prediction is such an interesting and hot topic, several studies have been proposed in this area to better understand the phenomenon and suggest ways and solutions to tackle it. These works can be divided into several categories.

The Forest Fire Prediction Using Image Mining Techniques includes using satellite imagery for data mining to create predictive models. The second category of approaches uses machine-learning algorithms for fire prediction. These works use different methods and algorithms of machine learning to perform tasks or make predictions based on data. The third category of forest fire prediction models uses deep learning algorithms, which is based on artificial neural networks. Finally, the forest fire prediction models using data mining algorithms that are based on algorithms that deal with dataset.

3.2.1 Forest fire prediction using Machine learning

The authors Abid and Izeboudjen,(2020)[11] , present a predictive model based on the decision tree for forest fires prediction in Algeria. The data used is collected from two regions of north Algeria : Sidi-Bel-Abbes and Bejaia.The Meteorological data with three attributes that influence the fire occurrences are used, namely : temperature, relative humidity (RH) and wind speed.

Results shown that the decision tree is suitable for this purpose, since its gives significant performances and it can be translated to rule based, hence its hardware implementation will be relatively simple and will requires fewer resources when implemented as an IP core.

Another work proposed by Mohajane et al.,(2021)[14] , which involves the creation of forest fire maps in order to protect the functionalities of forest ecosystems as

well as their precious benefits for people's well-beings.The data used is an ensemble of historical data and previous reports used for preparing for study the area. Data on fire occurrence exploited from the fire information and resources management system, two types of satellites Landsat8-OLI images are used to cover the area of study, the digital elevation model obtained from the shuttle radar topographic mission. Meteorological and topographic data are used and calculated using surface tool in spatial analyst tool.

This map generated are so useful and effective management tools for analyzing and developing forest fire management and strategies also can be used for all areas suffered of the fires forest around the world .

Preeti et al.,(2021)[19] , describes a system that brings together Kaggle's data set which consists of meteorological data, they carried out the exploration analysis than pre-processing where they try to remove noisy data and convert this categorical data to numerical data so that it is easy to understand this dataset. After the pretreatment technique, then the location of the hotspot is identified based on weather conditions data available in the dataset, then apply the models to predict the chances of occurrence of a fire and send the notification at the nearest station.

The data employed in this proposal is meteorological data such as : rain, temperature, humidity, and the wind. Progress reads inputs and grows an abstraction-assisted regression model, time and weather variables. After data collection, data pre-processing takes where the dataset should be trained in the standard format. After data preparation, the appropriate model should be selected based on the data set. They use regression techniques used for prediction, random forest (RF), decision Tree (DT) and Support Vector Regression (SVR) and Naive Bayes.

Liqinget al.,(2017) [22], propose a solution based on machine learning algorithms using logistic regression models and other methods.

In this proposal, the methods of spatial overlay analysis, binomial logistic model and Kriging interpolation were used to study the relationship between forest fire occurrence and terrain factor such as altitude, slope and aspect, vegetation factor and meteorological factor including precipitation, average temperature, daily average wind speed and average relative humidity. Based on the data of fire and non-fire points under different ratios, a binomial logical model was established to predict the probability of forest fire occurrence and fire danger rating in Lijiang area.

This paper randomly selects fire points and non-fire points to construct sample data, and choose the best performance model to predict forest fire danger in Lijiang area through test of model fitting at different ratios of fire and non-fire points. In addition to data of fire points, a percentage of random points need to be created as non-fire points.

3.2.2 Forest fire prediction using Deep learning algorithms

Singh et al. ,(2021)[23] , propose a new method namely parallel SVM for reliable performance of forest fire prediction. The data used is collected from the Indian meteorological department which consist of weather data from the Indian region. Different

frameworks are used for implementation of the system which is Django framework for the user interface, apache spark framework for big data analysis.

This type of solution can help very well the detection of the fires before it destroys the whole forest and simplified the prediction of this fire forest.

3.2.3 Forest fire prediction using data mining algorithms

Stojanova et al.,(2006)[24],propose a system that predicts the forest fires in Slovenia using different data mining techniques. The study uses predictive models based on Geographic Information System (GIS) forest structure : the weather forecast model - Aladdin and MODIS satellite data.

On these datasets, they applied the logistic regression and decision trees (J48), as well as forests, bagging and boosting decision trees, in order to obtain predictive models of fire occurrence. The best results in terms of predictive accuracy were obtained by bagging the decision trees.

3.2.4 Forest fire prediction using image mining algorithms

Cortez and Morias,(2008)[12] , advance solution, which is based on SVM is capable of predicting small fires, which constitute the majority of the fire occurrences, and is still useful to improve firefighting resource management.

Forest fires are a major environmental issue, creating economical and ecological damage while endangering human lives. Fast detection is a key element for controlling such phenomenon. To achieve this, one alternative is to use automatic tools based on local sensors, such as provided by meteorological stations.

Data Mining (DM) approach is explored to predict the burned area of forest fires. Five different DM techniques, such as Support Vector Machines (SVM) and Random Forests, and four distinct feature selection setups (using spatial, temporal, FWI components and weather attributes).

Compares to other methods, this solution offer a very good result in the prediction of the small fires which can accumulate into large forest fires I there is no intervention, and again with a simplified use of metrological data which is not expensive compared to the data interpreted by satellite and other expensive means.

The authors Yang et al., (2021)[26], suggest a system named Agni which represents a cost-effective fire prediction system for developing countries.

In this work Landsat 7 satellite images, retrieved from Google Earth Engine are used as inputs to the neural network, for ground truth labels. With an integration of artificial intelligence (supervised learning to train a neural network data) with remote sensing . This model uses readily available remote sensing data in the form of satellite images to predict fire hotspots, and he performs consistently well under extensive evaluations. Also it can lead to reliable and cost-effective forest fire prediction systems.

3.3 Comparative study and analysis

In this part, we will bring much closer to our subject by presenting the different approaches relative to forest fires and their main characteristics with a comparative study that regroups the most relevant propositions of the researchers by associating a summary for each approach used. On the other hand, we will also discuss the results obtained for the forest fire prediction systems already implemented.

The table below summarizes the main characteristics of the approaches cited above. The table contains seven columns that indicate a comparison criterion as follows :

- **Approach** column designates the approach of each paper.
- **Category of the approach** column indicates the category of the approach used.
- **The Data source** column specifies the data sources used.
- **The Output** column indicates the final output of the approach.
- **Used Technique** column designates the methods used for the forest fire prediction process.
- **The Supported Tool** column indicates whether the approach has been implemented with a software tool
- **The Advantages** column presents the main advantages of the approach.

Approach	Category of the approach	The Data source	The Output	Used Technique	The Supported Tool	The Advantages
Abid and Izeboudjen,(2020)	Machine learning	Dataset of Sidi-Bel Abess and Bejaia. Dataset of weather elements.	The boosting of decision tree (AdaBoostM1). Generation of a predictive model with the aim being its implementation in hardware.	The Decision tree algorithm The binary DT classifier.	Yes	Good results
Cortez and Morias,(2008)	Image Mining	real-world data of Portugal forest fire data of the Montesinho natural park type of vegetation meteorological data	automatic detection tools	satellite data. infrared/smoke scanners . local sensors. Data Mining algorithms. Features selections.	NO	Real-time data Low costs data
Yang et al,(2021)	Machine learning	Remote sensing data. Fire Information	Hospot.	Performance of Model. custom loss function,	NO	Requiring less storage of data. cost-effective fire prediction system

TAB. 3.1 : State of the art of related works

Singh et al. ,(2021)	Deep learning	Weather data Information about the forest fire	Predector tab Alert tab Intensity tab Validation tab	Django framework PySpark module	YES	More efficient and reliable model
Mohajane et al.,(2021)	Machine learning	Historical data Fire occurance data Fire informations for resource management system		Digital elevation model Normalized difference vegetation	NO	immediate prediction and identification model.
Liqinget al.,(2017)	Machine learning	forest fire statistics, vegetation type, and meteorological data and DEM data of fire occurrence area.	binomial logical model	The grid method random point method spatial overlay analysis, Kriging interpolation analysis, Logistic regression model	YES	using weather factors to perform modeling. Study the relationship between the occurrence of forest fires and the terrain factor.
Preeti et al.,(2021)	Machine learning	meteorological data (Temperature, Relative Humidity andWind Speed)	almodels to predict the chances of occurrence of fire and send the notification to the nearest station.	DecisionTree, Random Forest, Support Vector Machine, Artificial Neural networks (ANN) algorithms	NO	Solves the problem of computer file accuracy.

TAB. 3.2 : State of the art of related works(suite)

Stojanova et al.,(2006)	Data Mining	GIS data Multitemporal MODIS data Meteorological ALADIN data	predictive models of fire occurrence	logistic regression, random forests, decision trees (J48), bagging and boosting ensemble methods.	YES	Produce very accurate probability. The observation can be much easier to find many approximate rules. Decision trees shows the best results.

TAB. 3.3 : State of the art of related works(suite)

Through these commonly used various techniques and algorithms, artificial intelligence has been effectively applied to all aspects of life, especially in the field of prediction.

In this section, several papers are covered that explain in detail forest fire prediction methods that can help produce interesting results. Among these methods, we find that machine learning plays an extremely important role and is identified by these different techniques and algorithms, such as random forests, SVMs, regression trees, artificial neural networks (ANNs)...this Compared with the timely prediction of forest fires by responding to simple problems, the method results in high computational efficiency due to the application of various algorithms, thereby improving the performance and robustness of the system. On the other hand, it brings inconvenience in terms of complexity and delivery volume to do good research and get expected good results.

There is also data mining on site, which will also be presented as one of the methods commonly used in the field of forecasting using these different algorithms. The main benefit of this approach is that it reduces solution time thanks to the use of several different statistical techniques to analyze the data, but the disadvantage is that it is computationally expensive.

On the other hand, we have deep learning which is considered to be one of the most useful methods along with machine learning and its importance is reflected in these different algorithms and techniques such as artificial neural network (ANN) and traditional neural network (CNN)) . . . which makes them valid and believable. Deep learning has several advantages as it allows to process unstructured data such as images and videos... and it also ensures superior quality of results obtained better than any other method, but it also knows to represent in large amounts of data Some problems with it, it requires

a lot of computing power, and even its cost is so high that it is instead reserved for big companies with franc budgets, which can be set up if that makes it less useful and unusable.

Since we also have image mining technology, as the name suggests, it is very suitable for image processing, whether satellite or other software implementation. In this type of method, the data used is usually images, which is why this method is so effective for its different object detection and image classification and grouping techniques. One of the advantages of this approach is that it adapts well to geometric and satellite data, but it also has many disadvantages as it requires a lot of processing in the database before use, which buys us a lot of time and reduces the system its effectiveness.

Our solution is to implement such an efficient and useful prediction system based on machine learning. And aiming to provide a useful and usable forest fire prediction system to detect forest fires and manage them in a profitable and efficient manner, logistic regression is one of the easiest algorithms to explain in machine learning, and improvised so Impressive results, especially since he is so time-flexible and responsive, doesn't require much computing power, and it allows us to get results with greater precision, which motivates us to choose it and use it, and most importantly, It is the one of the most efficient algorithms in the field of prediction . Compared with other algorithms that have been proposed in previous work, this algorithm provides very good utility.

3.4 Conclusion

In this chapter, we have established a state of the art of the principles of the contributions in the field of forestry which represents related works that we have synthesized ; we have presented this in a table which contains the main lines of each approach synthesized, while following each work with a brief paragraph summarizing it.

In the next chapter, we present our contribution in detail.

Chapitre 4

Prediction of forest fires

4.1 Introduction

Algeria is a particular ecological area in the biosphere. However, the ignorance of wildfire disasters and the lack of funds to deal with these disasters has accelerated the process of desertification. In addition, the currently delicate Algerian forest needs to be protected because of its rich biodiversity and its impact on the country's socio-economic balance. For development programs to be successful and to combat environmental degradation, it is therefore necessary to find computer-aided solutions and develop an integrated and participatory approach for all relevant stakeholders. Therefore, there is interest in forest fire forecasting systems.

Forest fire forecasting has become an annual must-have activity as we cannot avoid these fires and this regrettable loss, human or material, at any time of the year.

A forest fire forecasting system may be the solution we've been looking for years, especially as computer technology evolves, it allow us to look at the problem in a way that traditional services never considered. These solutions can be so practical and reliable in the forest fire problem, if successful, we will have the opportunity to develop such reliable, useful and inexpensive systems to help solve the forest fire problem, why not put them in what others have introduced in our opinion Areas where we have no effective solutions

In this chapter, we will present in detail our approach that we used during our project as well as its different steps to make a fire prediction based on weather data. We start first by describing how to model a prediction model then we present the main steps of our proposal.

4.2 Modeling a prediction system

Our project involves the prediction of forest fires,it includes a certain in-depth study of datasets typically located in certain meteorological parameters. Based on these parameters, we can assess whether the area is likely to be at risk of fire or not. For this purpose, there are several steps toperform in order to have a powerful prediction model ; these steps being the study of the different parameters and their influence statistics, the selection of algorithm models to use, the training of these algorithm models on the dataset and the evaluation of the obtained results.

Classification techniques are an important part of learning applications. For our study, we focus on logistic regression as a simple model widely used in classification problems. Our interest in logical regression is explained by the fact that this model makes it possible to model binary variables or sums of binary variables.

4.2.1 Classification model

A classification model is a model that classifies each instance of a datasetin a category. There are two main types of classification model :

Binary : there are only two possible categories, such as 0 or 1, Pass or Fail, etc. For example : the potability of awater (potable or non-potable).

Multi-class : there are at least three categories, for example : the weather on a given day(rain, sun or snow).

Logistic regression can be applied to both types of classification. In our case, we used the binary because the principle is to predict whether a selected area is at risk of forest fire or not.

4.2.2 Logistic Regression Principles

The principle of the logistic regression model is to relate the occurrence or the non-occurrence of an event at the level of explanatory variables.

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.

Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. [2] It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function,[4] which predicts two maximum values (0 or 1).

Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets.

Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification. The below image is showing the logistic function

- **Logistic Function (Sigmoid Function) :**

  The mathematical function used in logistic regression is Sigmoid function and she serves to map the predicted values to probabilities.

  It maps any real value into another value within a range of 0 and 1. The value of the logistic regression must be between 0 and 1, which cannot go beyond this limit, so it forms a curve like the "S" form. The S-form curve is called the Sigmoid function or the logistic function.[3]

  In logistic regression, we use the concept of the threshold value, which defines the probability of either 0 or 1. Such as values above the threshold value tends to 1, and a value below the threshold values tends to 0.

  Figure 4.1 illustrates the sigmoid function.

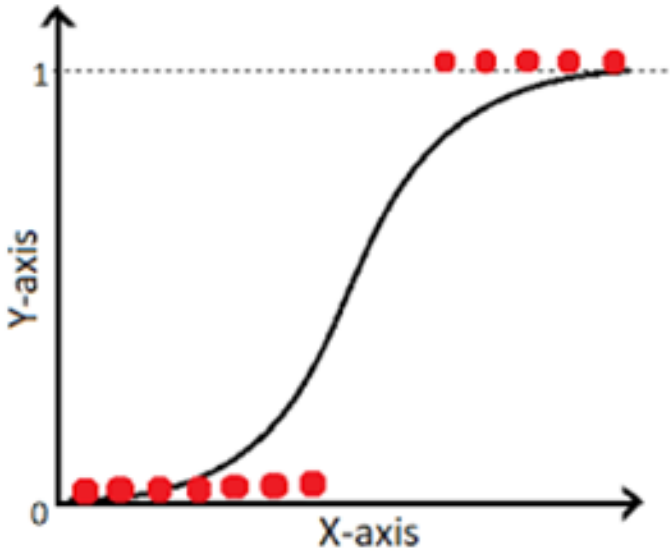

FIG. 4.1 : Sigmoid Function

- **Assumptions**

  In the logistic regression we have this two point to take in consideration : The dependent variable must be categorical in nature. The independent variable should not have multi-collinearity.

- **The mathematical Equation**

  The Logistic regression equation can be obtained from the Linear Regression equation.[5] The mathematical steps to get Logistic Regression equations are given below :

$$y = \underset{0}{b} + \underset{11}{bx} + \underset{22}{bx} + \underset{33}{bx} + \ldots + \underset{nn}{bx} \tag{4.1}$$

  In the equation (4.1), y is the dependent variable and x1, x2,.....,xn are explanatory variables.

  In Logistic Regression y can be between 0 and 1 only, so for this let's divide the equation (4.1), by (1-y) :

$$\frac{y}{1-y} \tag{4.2}$$

  0 for y=0 and infinity for y=1

  But we need range between -[infinity] to +[infinity], then take logarithm of the equation it will become :

$$log(\frac{y}{1-y}) = \underset{0}{b} + \underset{11}{bx} + \underset{22}{bx} + \underset{33}{bx} + \ldots + \underset{nn}{bx} \tag{4.3}$$

  The equation (4.3) is the final equation for Logistic Regression.

4.3 Proposed model

The proposed model provides prediction results using the logistic regression algorithm. The figure 1 depicts the architecture of our proposal, which mainly involves 2 steps. The first step "Construction model " consists of different phases to build a system using training data. The second step " the prediction model " consiste of examination of system based by the model constructed and by introducing new data .

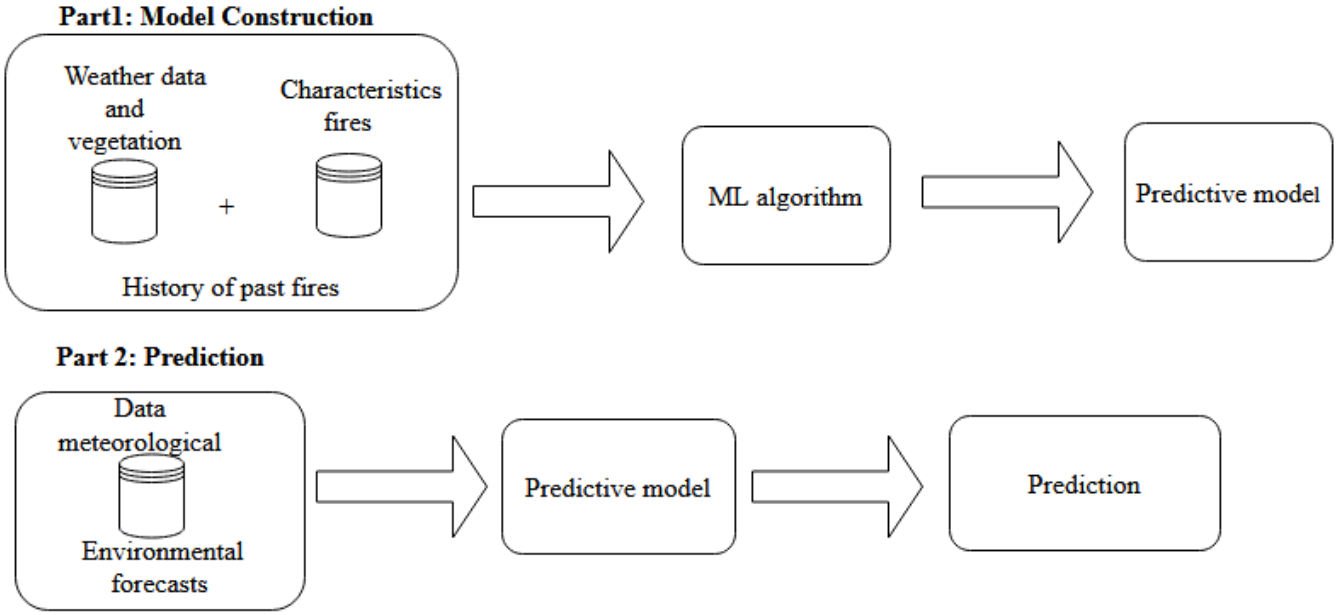

Fig. 4.2 : Prediction system Architecture

In the following, we will present the different steps of the proposed system in detailed.

4.3.1 Construction model step

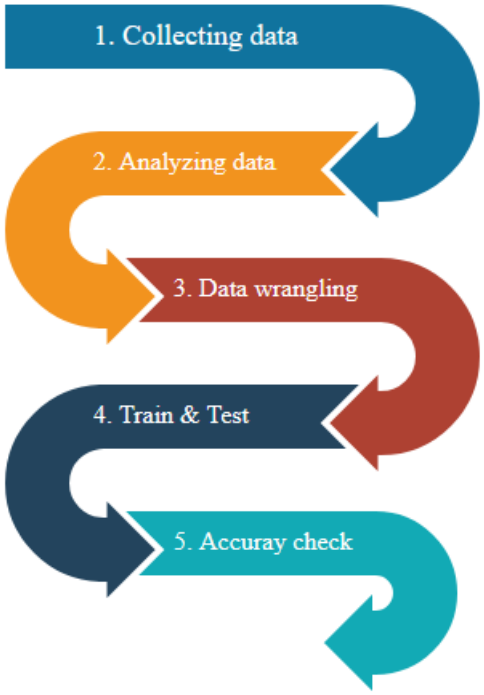

FIG. 4.3 : Construction model phase

Illustrations are included to explain the different stages of the model building phase.

This stage includs the realization of 5 stages, which are : Data collection this part collects the data needed for forecasting, Data analysis includes analyzing the datasets that need to be used, Data preparation includes preparing data to make it available for use, The train and test consists of training the model with the training set and testing our model with the test set, and finally the accuracy check we need to evaluate our system.

In this step, the system works as follows : it examines a data set in which each observation contains information about the response variable as well as the predictive variables.

This classification task is performed as follows : examine the data set enclosing the predictive variables and the response variable classified earlier. Therefore, it learns the combinations of variables associated with this or that response value. This data set is called the learning set.

This step includes selecting a function f, that is to say to find an evaluation of the Parameters xi. The selection of these parameters is carried out by a learning algorithm that receives the learning set as input. The set of Xi parameters resulting from learning is called a model.

1. **Data Collection**

   The first phase to start the construction of our model is collecting data, which is a fundamental and critical part of our project because everything else depended on it. We note that the quality of the collected data plays a very important role, where it is important to collect reliable data to ensure more interesting and accurate results.

The purpose of important goal of data collection is to ensure that information-rich and reliable data is collected for statistical analysis so that data-driven decisions can be made for research.

2. **Data Analysis**

Analyzing the dataset is all about getting an idea of what it looks like and getting ideas about it. We can find the correlation of each pair of characteristics (and the result variable), and visualize the correlations using a heat map, and creating different plot to check relationship between variables.

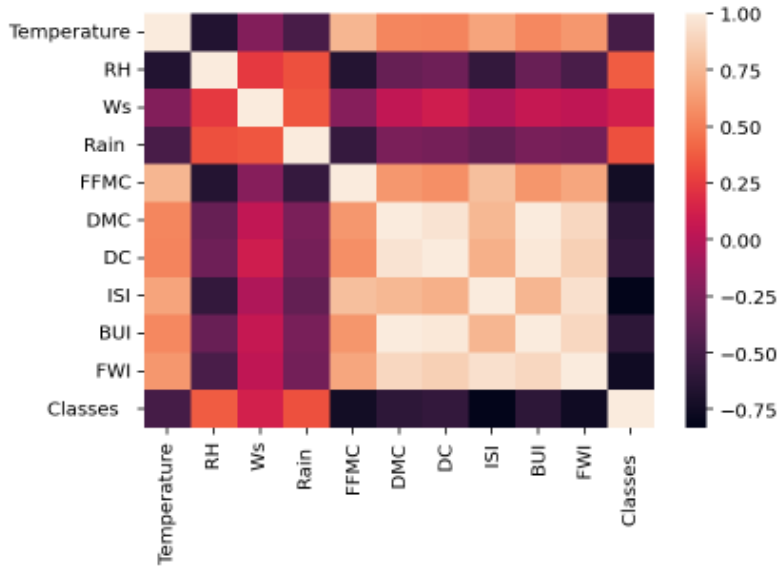

Fig. 4.4 : Heatmap of feature (and outcome) correlations

In the above the heat map of the features correlations above, brighter colors indicate more correlation. As we can see temperature, HR, Ws, Rain, FFMC, DMC, DC, ISI, BUI, FWI all have a significant correlation with the outcome variable. Also note the correlation between pairs of characteristics, such as temperature and rain, or RH and ISI.

3. **Dataset Preparation**

This phase concerns the data preparation to be ready for use. Clean up data by removing null values and unnecessary columns from the dataset. This pahse involves different features :

- **Data cleaning**

  The amount of data available to us is constantly increasing, and so is the risk of error. Therefore, we must rely on data cleansing to simplify our data management process.

  Data cleaning is the process of repairing or removing incorrect, corrupt, malformed, duplicate, or incomplete data from a dataset. When combining multiple

data sources, there are many opportunities for data to be duplicated or mislabeled. If the data is wrong, then the results and algorithms are unreliable, even if they appear to be correct. There is no absolute way to prescribe the exact steps in the data cleaning process, as the process varies from one dataset to another. But it's important to set a pattern for your data cleaning process so you know you're doing it right every time.

- **Data reducing**
  Data reduction is the process of reducing the capacity required to store data. Data reduction can improve storage efficiency and reduce costs. Storage manufacturers usually use raw capacity and effective capacity to describe storage capacity, that is, the reduced data.

  Data reduction is the conversion of empirically or experimentally obtained numerical or alphanumeric information into a corrected, ordered, and simplified form. Data reduction can have two goals : to reduce the number of datasets by eliminating invalid data, or to generate aggregated data and statistics at different levels of aggregation for different applications.

- **Data transforming**
  Data transformation is the process of changing the format, structure, or value of data, such as importing a database file, an XML document, or an Excel spreadsheet into another file. Transformation typically involves converting the original data source into a sanitized, validated, and ready-to-use format. Analyzing information requires structured and accessible data for optimal results. Data transformation allows organizations to change the structure and format of the original data as needed.

  Data transformation can be constructive (adding, duplicating, and duplicating data), destructive (removing fields and records), aesthetic (standardizing salutations or street names), or structural (renaming, moving, and combining columns in a database).

4. **Train and Test**

   In this phase, we build the classification model on the train data and predict the output on the test data. The model used is logistic regression. First, we created an instance called logreg and then use the fit function to train the model.

5. **Accuracy check**

   In this step we have calculate accuracy to verify the accuracy of results.and that to evaluate the final system

4.3.2 Prediction step

When you have a reliable model, you can make predictions about new examples. A prediction tool therefore corresponds to a trained machine.

In order to perform the main function of prediction, the system would examine new observations for which no information on the variable to be predicted is available, while using the model whose parameters were estimated in the learning phase. It would then assign classifications to the new observations based on the classifications in the learning set.

4.4 Conclusion

This chapter summarizes the methodology used in the development of prediction by describing the different steps that led to the final models and their performance.

The approach is inspired by the various works related to the prediction of forest fires and allows the use of several measures for classification.

In the next chapter, we will proceed to explain all aspects related to the implementation of the system.

Chapitre 5

Experimentation

5.1 Introduction

As part of our research, we processed the Annual Forest Fire Reports from previous years. This processing makes it possible to extract the data necessary for machine learning from our system. The input data we used are instances, dataset extracts to train and test our approach in real time on instances extracted from data from two regions of Algeria.

In this chapter, we will introduce the different aspects related to the implementation of our approach that we have developed, namely, technologies, software and languages selected using different data sources for the implementation of our approach.

5.2 Data set description

A dataset is a coherent set of data produced as part of the same project, on the same object of study and/or collected on the same place. A dataset can be a CSV file, TSV, an Excel spreadsheet, a table in an RDBMS, a document in a NoSQL database, an output of a service Web, etc. A dataset can be composed of a hundred files ; its size can reach a few tens of gigabytes.

Dataset used was made by "Faroudja ABID" [10] and download from the Kaggle [1] website and it is in CSV format. Dataset size is 14.3+ KB, includes 244 instances that regroup data of two regions of Algeria, namely the Bejaia region located in the northeast of Algeria and the Sidi-Bel-Abbes region located in the northwest of Algeria, 122 instances for each region. The dataset includes 11 attributes and 1 output attribute (class).

The Dataset involves twelve (12) columns :

- **Date :** (DD/MM/YYYY) Day, month ('June' to 'September'), year (2012) .
- **Temp :**in Celsius degrees : 22 to 42
- **RH :**in % : 21 to 90
- **Ws :**in km/h : 6 to 29
- **Rain :**total day in mm : 0 to 16.8 FWI Components
- **FFMC :**index from the FWI system : 28.6 to 92.5
- **DMC :**index from the FWI system : 1.1 to 65.9
- **DC :**index from the FWI system : 7 to 220.4
- **ISI :** index from the FWI system : 0 to 18.5
- **BUI :**index from the FWI system : 1.1 to 68

[1] https://www.kaggle.com

- **FWI :** from 0 to 31.1
- **Classes :**two classes, namely 'Fire' and 'not Fire'

The figure 1 provides an excerpt of the data set.

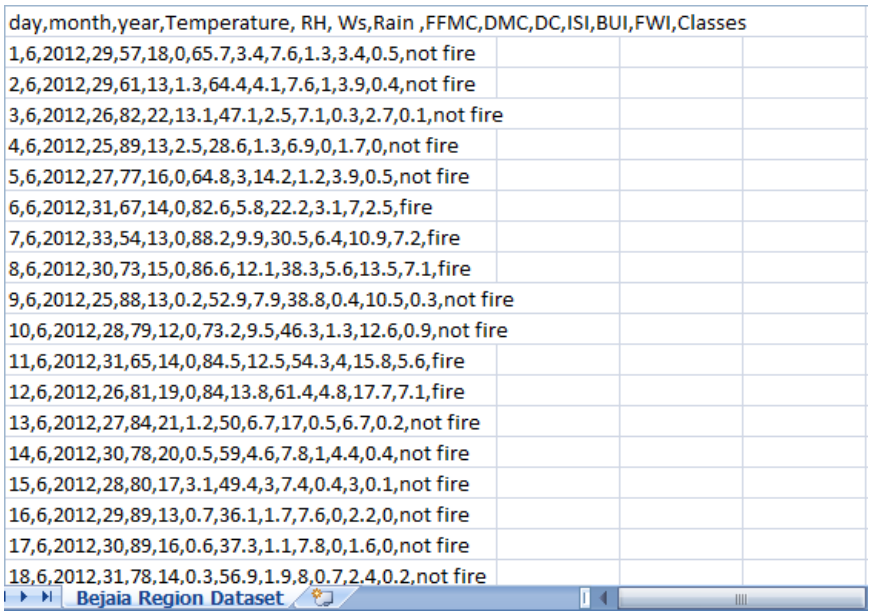

FIG. 5.1 : An excerpt of the data set.

5.3 Hardware environment

All experiments were carried out with an Intel Core i3 processor with a frequency of 2 GHz and 4 GB of memory under the Windows 10 platform.

5.4 Software environment

- **Anaconda**

  Anaconda [2] is an Open Source Suite or one that includes a series of applications, libraries, and concepts designed for developing Data Science with Python. In general the Anaconda Distribution lines are a Python distribution that functions as an environment manager, a package manager and has a collection of over 720 open source packages.

  Anaconda Distribution is grouped into 4 technological sectors or solutions, Anaconda Browser, Anaconda Project, Data Science Libraries and Conda. All these are installed automatically and in a very simple procedure.

[2] https://www.anaconda.com/

- **Anaconda navigator**

  The Anaconda browser is a graphical user interface (GUI) included with the Anaconda distribution that allows users to launch applications, but also manage conda libraries, environments, and channels without using the command line.

  Navigator can also access libraries in Anaconda Cloud or a local Anaconda repository to install, run, and update in the environment.

- **Jupyter Notebook**

  Jupyter[3] notebooks are electronic notebooks that, in the same document, can gather text, images, mathematical formulas and executable computer code. Originally developed for the Julia, Python and R programming languages (hence the name Jupyter), Jupyter notebooks support nearly 40 different languages.

  The cell is the basic element of a Jupyter notebook. It can contain text formatted in Markdown format or computer code that can be executed.

5.5 Programming Language

Python[4] is a cross-platform, object-oriented open source programming language. Thanks to specialized libraries, Python is used for many situations such as software development, data analysis, or infrastructure management.

It is an interpreted programming language. Python makes it easy and quick to create programs.

5.6 Python Libraries

- **Pandas**

  Acronym for Python Data Analysis Library, Pandas is a Python package that is compatible with other Python packages, such as NumPy, Matplotlib, etc. Its Installation is done by opening a command shell and invoking this command for Python 3.x : pip3 install pandas.

  The Pandas library provides users with the ability to manage large data sets. It provides tools to read and write data, clean and modify data.

  The Pandas package has the semantics of a spreadsheet, and it also works with different file types, such as xsl, xml, html, CSV, and TSV files. Pandas takes data stored in CSV or TSV files and provides a type of data called a Dataframe (similar to a Python dictionary) with extremely powerful features (similar to those of a spreadsheet).

[3] https://jupyter.org/
[4] https://www.python.org/

- **Numpy**

  The Numpy, or Numerical Python, It is an open source library for the Python language. This tool is used for scientific programming in Python, especially for data science, engineering, math or scientific programming. This library is useful for performing mathematical and statistical operations in Python. It is great for multiplication of matrices or multidimensional arrays. Integration with C/C++ and Fortran is very easy. Numpy arrays are more advantageous than the traditional python arrays because of its use less memory and storage space.

- **Tkinter**

  Tkinter, or "Tk interface", is a module of python that provides an interface to the tk GUI toolkit, developed in TCL (Tool Command Language) and multiplatform, with support for Linux, MAC OS and MS Windows. Tk is natively present in Linux and MAC OS, and can be easily installed on MS Windows, but it is not part of Python. Tkinter is part of Python, being called "Tkinter" in versions prior to 3, and "tkinter" in subsequent versions.[21]

- **Matplotlib**

  Matplotlib is one of the most popularly used data visualization libraries of python. This library was built by a John Hunter who is along with several contributors, and it had put in a greater amount of time into prompting this software used by every scientist and philosopher across the globe. Matplotlib is a graphics library for data visualization package in Python which encompasses as an integral aspect in the python data science stack and it is easily supported with NumPy, Pandas and other relevant libraries. [16]

- **Sklearn**

  Scikit-learn (sklearn) is a free Python library for ML. It provides many libraries of algorithms in its Framework to implement.

  These libraries are particularly useful for data scientists. It specifically includes functions for estimating random forests, logistic regression, classification algorithms, and support vector machines. It is designed to coordinate with other free Python libraries, including NumPy and SciPy.[15]

5.7 Application interfaces

In this part, we will show the different interfaces of our application and explain the usefulness of each of them.

Home page

It is the first user-accessible interface, with a bare menu ;the first "Home" menu contains 3 buttons :

- The”**Dataset**”button : clicking on it opens a window displayingthe files with the CSV extension of the data we used.

- The”**Exit**” button : clicking on it opens another window to confirm if you want to leave the application as in Figure.

- The”**Read the brief**” button : We can access to our brief by clicking on.

Figure 5.2: is the first interface when launching our application and the first menu that appears is the ”Home” menu.

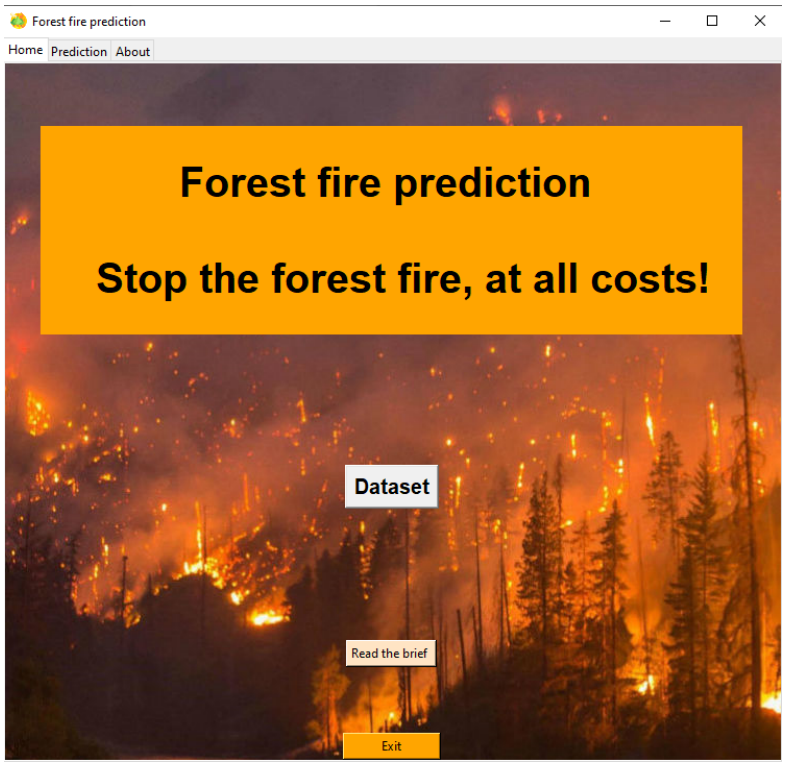

Fig. 5.2 : Home” interface

Figure 5.3: After clicking on the ”Dataset” button, a new window opens displaying the existing files in our hard disk, we select the desired Dataset, in our case it is CSV file named forestfires.csv and we import it for use in our evaluation system.

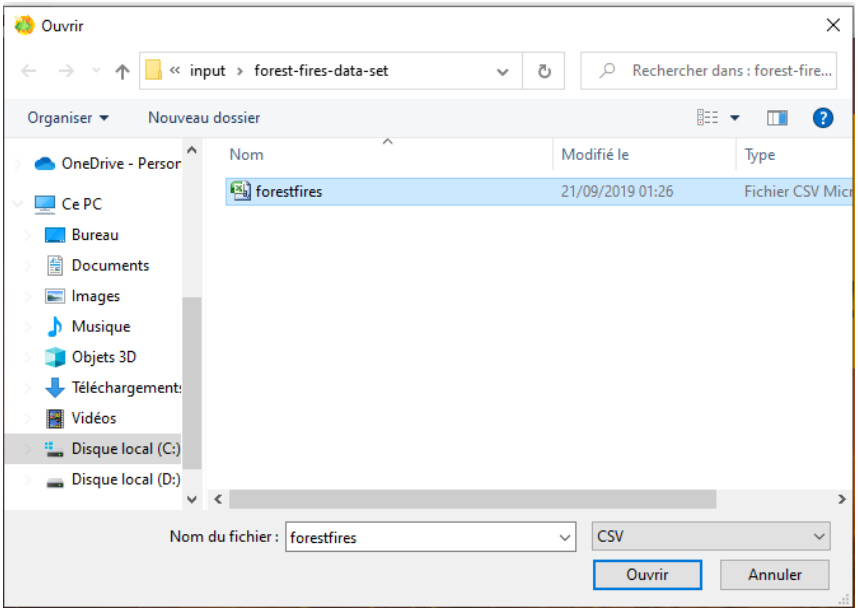

FIG. 5.3 : Dataset interface.

Figure 5.4: After clicking on the "Read the brief" button, a new window opens displaying the thesis on the prediction of forest fires in Algeria.

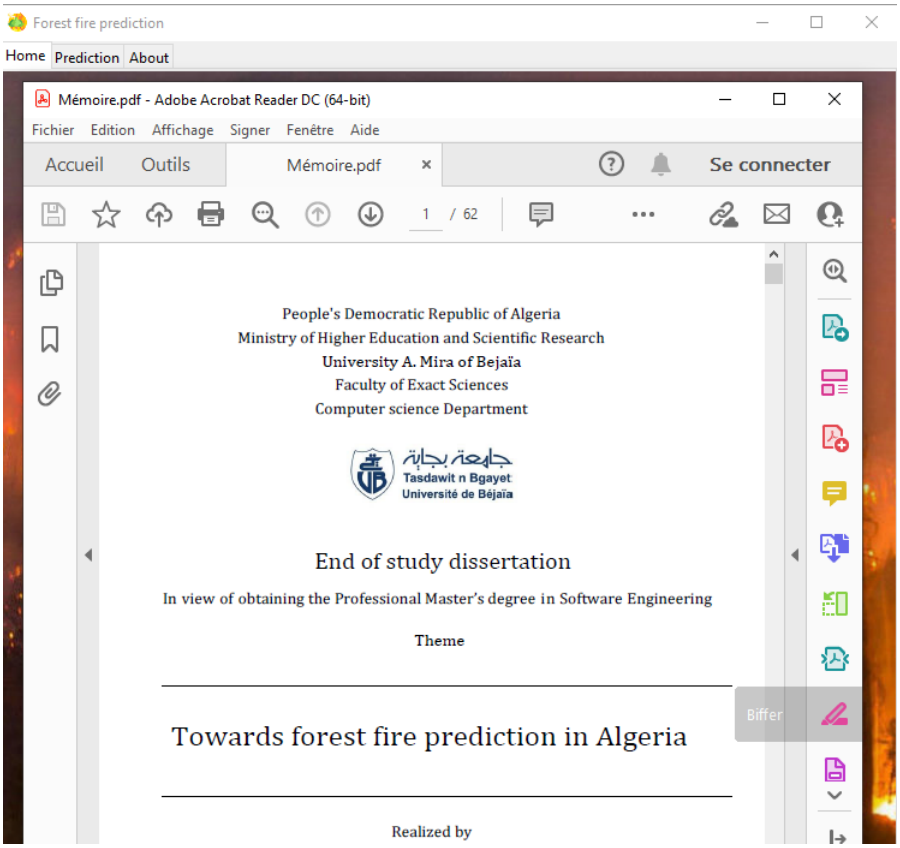

FIG. 5.4 : Interface to« Read the brief ».

Figure 5.5: After clicking on the "Exit" button, a new window opens displaying a dialog box to confirm if you want to exit the application.

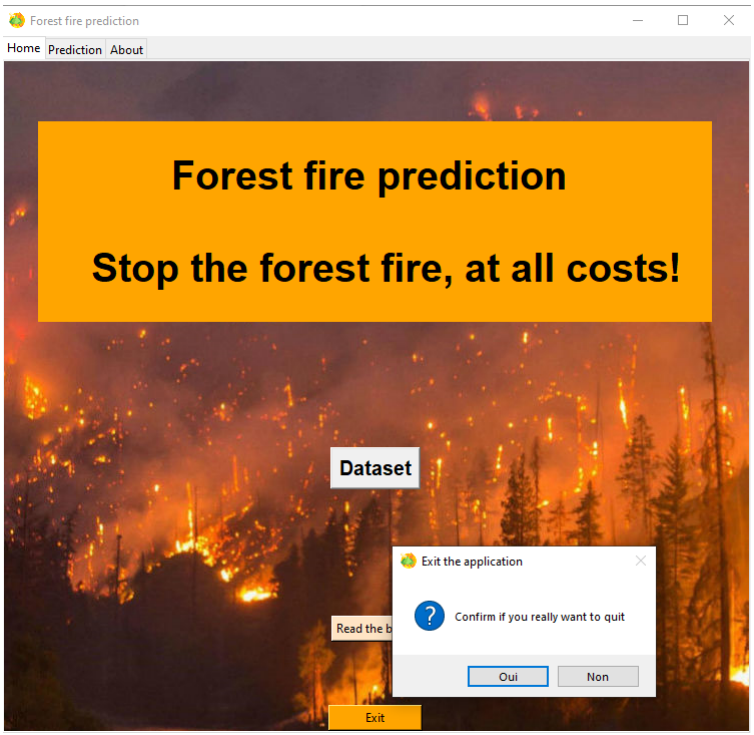

FIG. 5.5 : Interface to ”exit” the application.

Prediction page

Figure 5.6: Represents the interface of the second ”Prediction” menu of the application which consists of :

- Text fields where you can enter attributes to make the prediction.
- A ”Clear” button to clear the text fields of everything entered.
- A ”Predict” button to make a prediction.
- A ”help” button by clicking on it another window is displayed to see the information about the attributes
- An ”Exit” button to exit the application.

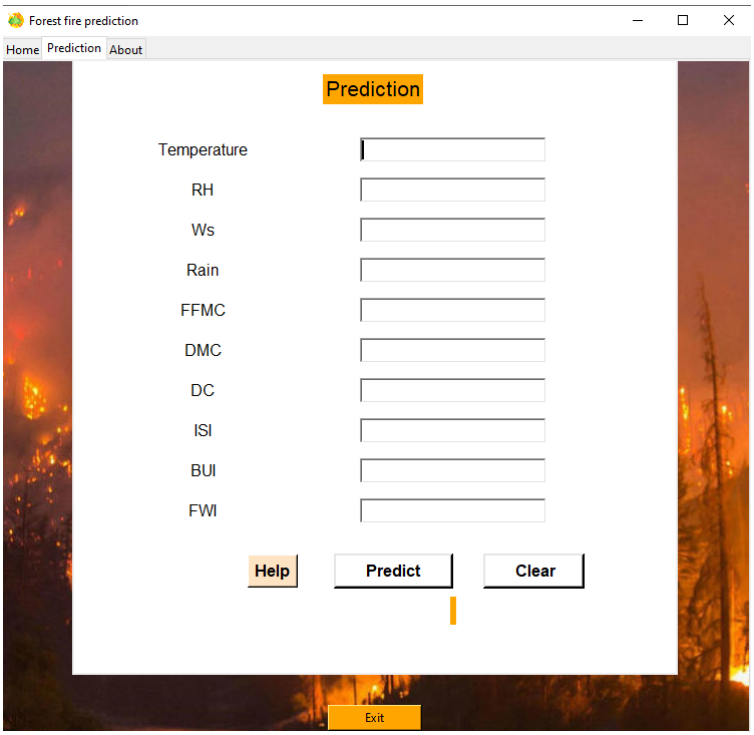

FIG. 5.6 : Interface of the "Prediction" menu.

Figure 5.7: shows an example of a « prediction » :

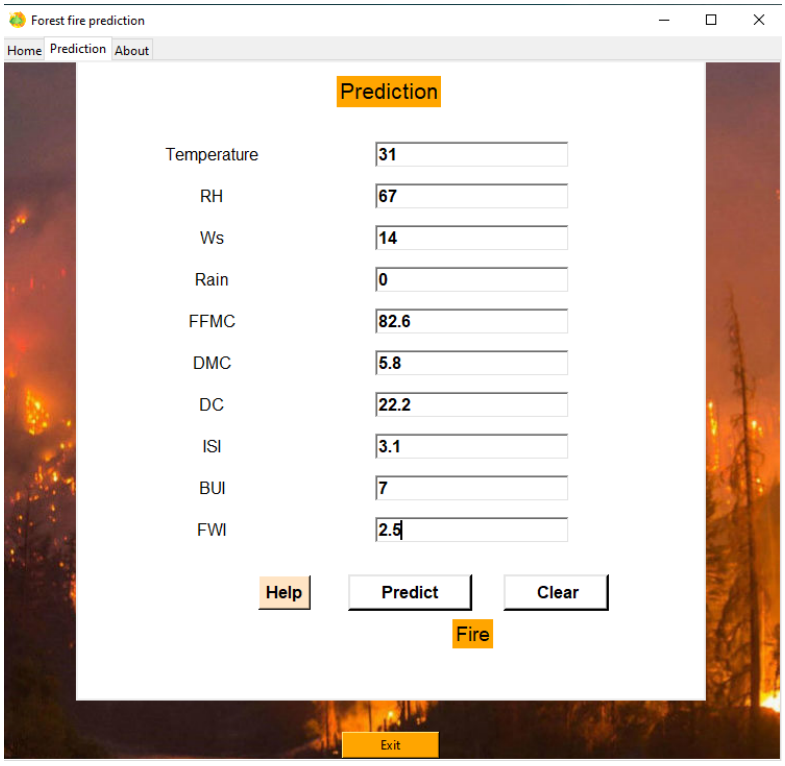

FIG. 5.7 : "Example prediction" interface.

Figure 5.8: shows the interface that is displayed to see the information about the attributes after clicking on "help".

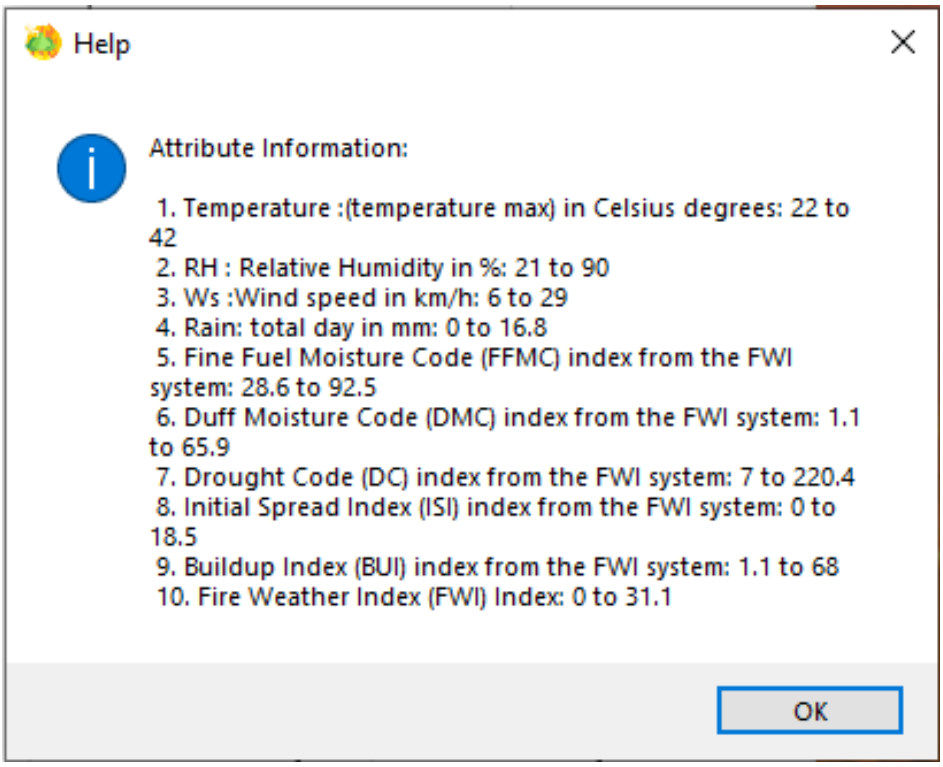

FIG. 5.8 : "Help" interface.

About page

This page represents a small description of our application, and the source of the Dataset used as well as the tools and techniques used to develop the forest fire prediction system, and the contact of the directors of this work.

Figure 5.9: shows the "About" menu interface.

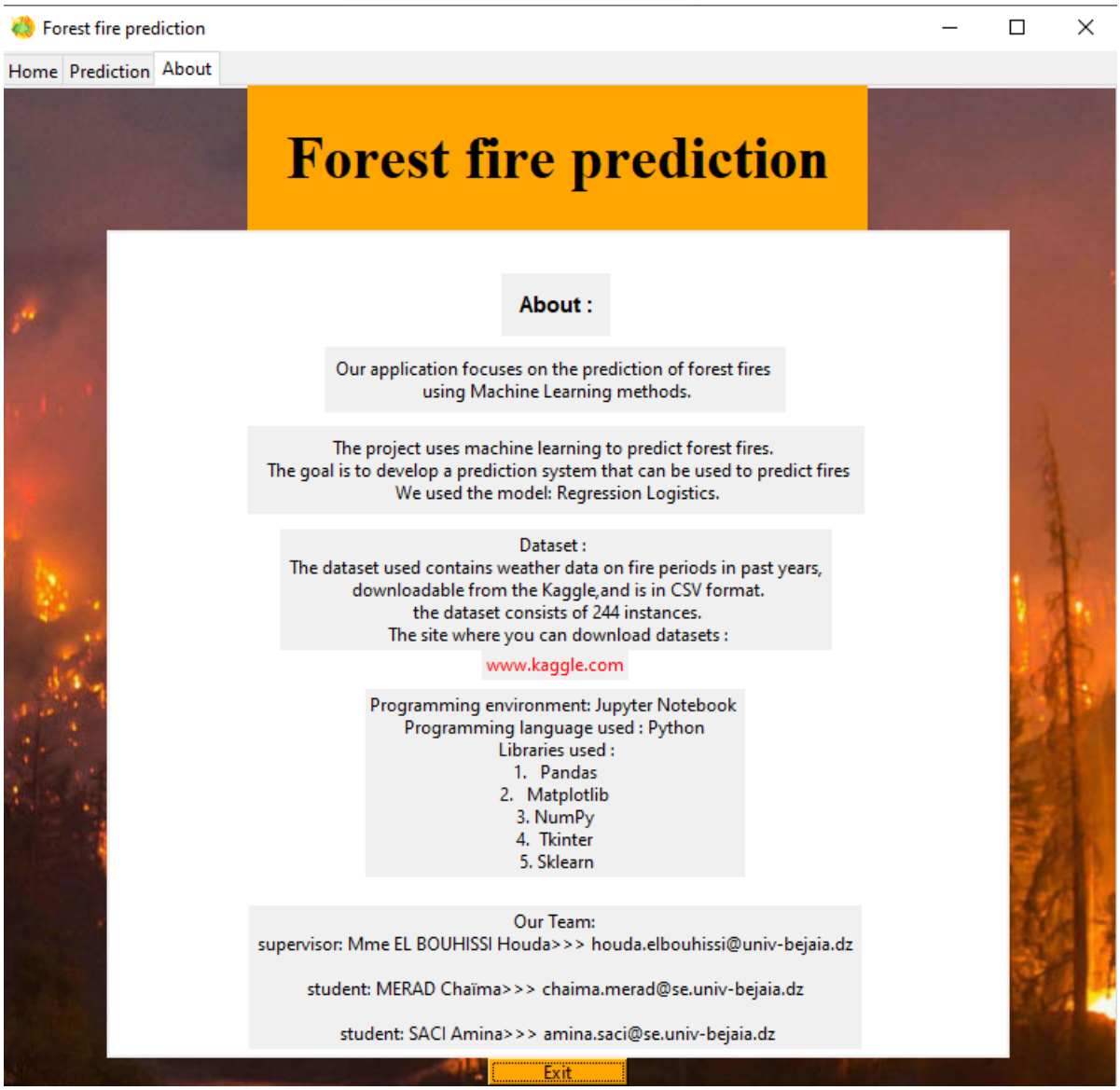

FIG. 5.9 : "About" menu interface..

5.8 Evaluation

The evaluation step is necessary to determine the efficiency of the prediction model. At this step, the performance of the calculations that have been performed will be tested with accuracy, precision and recall parameters. Evaluating a model is an essential part of building a model effective machine learning.

The confusion matrix summarizes the performance of a model when used in prediction. The structure of this matrix for a two-class classification problem is the following :

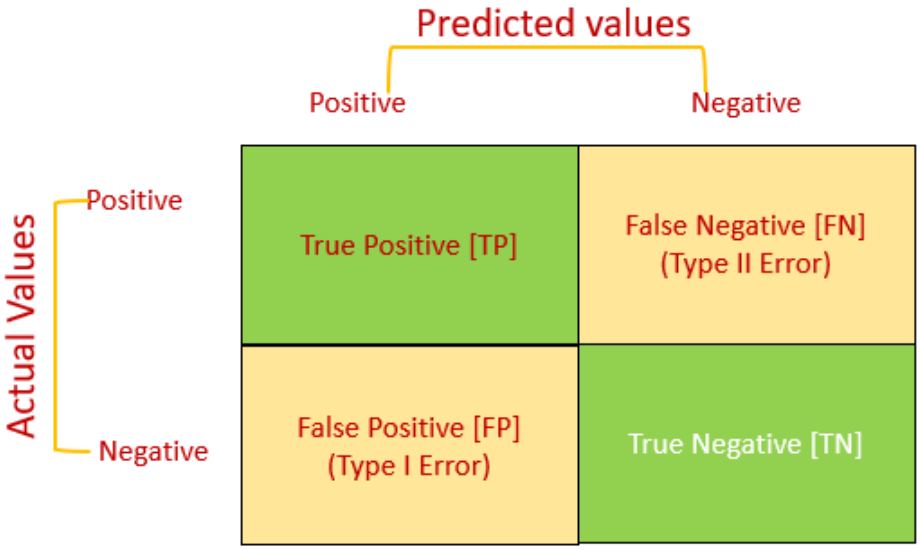

FIG. 5.10 : The confusion matrix

To fully understand how a confusion matrix works, it is important to understand the four main terminologies : TP, TN, FP and FN. Here is the precise definition of each of these terms :

- TP (True Positive) : Cases where the prediction is positive and the real value is actually positive.
- TN (True Negative) : Cases where the prediction is negative and the actual value is actually negative
- FP (False Positive) : Poorly predicted class 1 e cases where the prediction is positive, but the actual value is negative.
- FN (False Negative) : Cases where the prediction is negative, but the real value is positive.

From the confusion matrix, we can calculate the main measures of classifier performance. These are :

Precision : This is the probability that a positive predicted event is actually positive. Precision answers the question : "What proportion of positive identifications was actually correct ?".[6] The accuracy is calculated as follows :

$$Precision = \frac{TP}{TP + FP} \tag{5.1}$$

Recall (or sensitivity) : Proportion of well-classified items among those that are Positive. The recall answers the following question : "What proportion of real positive results have been identified correctly ?".[7] The recall is calculated as follows :

$$Recall = \frac{TP}{TP + FN} \tag{5.2}$$

Accuracy : Proportion of well-classified elements.[1]

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{5.3}$$

Figure 5.11 below represents the Precision, recall and accuracy for the Logistics regression classifier used in our model.

	precision	recall	f1-score	support
Positive	0.86	0.93	0.90	46
Négative	0.94	0.87	0.90	52
accuracy			0.90	98
macro avg	0.90	0.90	0.90	98
weighted avg	0.90	0.90	0.90	98

Fig. 5.11 : Evaluation of the logistic regression algorithm

Figure 5.12 shows the regression logistic model confusion matrix used, as we can see the TP is 43 and FP is 7, FN is 3 and TN is 45. If we calculate the precision and recall and accuracy from the confusion matrix, we will find the same results as the previous figure

```
array([[43,  3],
       [ 7, 45]], dtype=int64)
```

Fig. 5.12 : The regression logistic confusion matrix

Based on the obtained results, we conclude that the prediction model built for predicting forest fires provides interesting and conclusive results.

However, we can get an ideal model for predicting forest fires. For best results, it is suggested to increase the number of years it takes for training and testing.

5.9 Conclusion

In this chapter we have described the different interfaces that are there in our application with the navigation scheme between these different. We have given the hardware and software environments that have allowed us to carry out this work and to carry it out.

We have presented the essence of our work which is to create a forest fire prediction system. For the implementation, we chose two regions of Algeria, Bejaia and Sidi Bel abbes.

The instances have been classified as : fire, non-fire. Our system integrates in the field of artificial intelligence precisely (Machine Learning).

Chapitre 6

General conclusion

This work was carried out as part of our end-of-cycle master project in computer science Software Engineering option. It consisted of a machine learning approach to the prediction of forest fires in Algeria.

Predictive modeling is an emerging field, in recent years several researches are interested in the task of predictive analysis, especially in the forestry field. Our work focuses on predicting forest fires in the regions of Algeria.

The Dataset we used was taken from the Kaggle website containing fire data from two regions in Algeria : Béjaïa and Sidi-Bel-Abbes. The prediction was made on the meteorological data and then classified using an algorithm of the Machine Learning.

Machine learning is the most practical technique. So it would be much more easy to predict the possibility of a wildfire if a model was adopted to polarize them and learn from them. In this research work, forest fire data were analyzed and classified by logistic regression.

Our brief consists of six (6) chapters organized as follows :

In the first chapter, we defined our context and problematic as well as our objectives, we have also detailed our working methodology.

In the second chapter, we presented some definitions of the domain studies that is predictive modeling, its types, problems related to this field and Machine Learning and its types.

In the third chapter, we have established the state of the art that represents all the related works that we have synthesized, we have presented this in a table which contains the outline of each synthesized approach, while following each work by a brief paragraph that summarizes it, then we proceeded to an analysis comparative between the approaches of the related documents and our approach.

In the fourth chapter, we presented in detail the approach we used during our project as well as its different steps to make a prediction of fires from datasets.

In the fifth chapter, we have discussed the various aspects related to the implementation of the approach we have developed, namely, technologies, software and the languages chosen using different data sources for implementation of our approach.

Despite the difficulties we had during the realization of our work, such as the lack of databases, the lack of sources of information, we pushed the project as far as possible ; there are still many steps to add. In particular the implementation of some steps of our approach, namely, We are thinking of refining our approach and implementing it in better hardware and software conditions. Replace the logistic regression algorithms with other algorithms in order to improve the accuracy of the results obtained and build a more efficient and effective prediction system. Finally, apply the model really proposed on the ground and its exploitation by the Directorate of the Conservation of Forests.

The realization of this project was rich in lessons in several aspects. She allowed us to acquire new skills, and to put knowledge into practice theoretical ones that we have

acquired along our university curriculum. We have progress in many areas, notably in predictive modeling, programming in Python, state-of-the-art development...

Bibliographie

[1] Classification accuracy. https://www.sciencedirect.com/topics/engineering/classification-accuracy. Last accessed 20 August 2022.

[2] Logistic Regression , logistic regression principles. https://medium.com/analytics-vidhya/logistic-regression-c5a6c047363e. Last accessed 21 June 2022.

[3] Logistic Regression , sigmoid function or the logistic function. https://www.studocu.com/in/document/bharathiar-university/bachelor-of-computer-application/what-is-logistic-regression-and-what-is-it/22184122/. Last accessed 21 June 2022.

[4] Logistic Regression ,logistic function. https://pianalytix.com/logistic-regression-in-machine-learning/. Last accessed 21 June 2022.

[5] The mathematical equation. https://rs02.medium.com/introduction-to-logistic-regression-49904eb24b0f/. Last accessed 21 June 2022.

[6] Mesurer la performance d'un modèle ,precision. https://www.lovelyanalytics.com/2020/05/26/accuracy-recall-precision/. Last accessed 20 August 2022.

[7] Recall. https://www.practiceprobs.com/problemsets/evaluation-metrics-and-loss-functions/precision-and-recall/#. Last accessed 20 August 2022.

[8] Semi-supervised and reinforcement learningé. https://anexia.com/blog/en/machine-learning-for-beginners/. Last accessed 30 May 2022.

[9] Unsupervised learningé. https://phedone.com/fr/blog/artificial-intelligence-what-is-the-unsupervised-learning/. Last accessed 30 May 2022.

[10] Faroudja ABID. Algerian forest fires dataset data set. https://archive.ics.uci.edu/ml/datasets/Algerian+Forest+Fires+Dataset++#/. Last accessed 05 July 2022.

[11] Faroudja Abid and Nouma Izeboudjen. Predicting forest fire in algeria using data mining techniques : Case study of the decision tree algorithm. In *International*

Conference on Advanced Intelligent Systems for Sustainable Development, pages 363–370. Springer, 2019.

[12] Mauro Castelli, Leonardo Vanneschi, and Aleš Popovič. Predicting burned areas of forest fires : an artificial intelligence approach. *Fire ecology*, 11(1) :106–118, 2015.

[13] Thomas G Dietterich. Machine learning. *Annual review of computer science*, 4(1) :255–306, 1990.

[14] Ismail Elkhrachy, Quoc Bao Pham, Romulus Costache, Meriame Mohajane, Khalil Ur Rahman, Himan Shahabi, Nguyen Thi Thuy Linh, and Duong Tran Anh. Sentinel-1 remote sensing data and hydrologic engineering centres river analysis system two-dimensional integration for flash flood detection and modelling in new cairo city, egypt. *Journal of Flood Risk Management*, 14(2) :e12692, 2021.

[15] Raul Garreta and Guillermo Moncecchi. *Learning scikit-learn : machine learning in python*. Packt Publishing Ltd, 2013.

[16] John Hunter and Darren Dale. The matplotlib user's guide. *Matplotlib 0.90. 0 user's guide*, 2007.

[17] Rémy Kessler, Juan Manuel Torres-Moreno, and Marc El-Bèze. Classification thématique de courriels avec apprentissage supervisé, semi-supervisé et non supervisé. *les actes de VSST*, pages 493–504, 2004.

[18] Michael P LaValley. Logistic regression. *Circulation*, 117(18) :2395–2399, 2008.

[19] T Preeti, Suvarna Kanakaraddi, Aishwarya Beelagi, Sumalata Malagi, and Aishwarya Sudi. Forest fire prediction using machine learning techniques. In *2021 International Conference on Intelligent Technologies (CONIT)*, pages 1–6. IEEE, 2021.

[20] Vincenzo Riccio, Gunel Jahangirova, Andrea Stocco, Nargiz Humbatova, Michael Weiss, and Paolo Tonella. Testing machine learning based systems : a systematic mapping. *Empirical Software Engineering*, 25(6) :5193–5254, 2020.

[21] John W Shipman. Tkinter 8.5 reference : a gui for python. *New Mexico Tech Computer Center*, 54, 2013.

[22] Liqing Si, Lifu Shu, Mingyu Wang, Fengjun Zhao, Feng Chen, Weike Li, and Wei Li. Study on forest fire danger prediction in plateau mountainous forest area. *Natural Hazards Research*, 2(1) :25–32, 2022.

[23] Kajol R Singh, KP Neethu, K Madhurekaa, A Harita, and Pushpa Mohan. Parallel svm model for forest fire prediction. *Soft Computing Letters*, 3:100014, 2021.

[24] Daniela Stojanova, Panče Panov, Andrej Kobler, Sašo Džeroski, and Katerina Taškova. Learning to predict forest fires with different data mining techniques. In *Conference on data mining and data warehouses (SiKDD 2006), Ljubljana, Slovenia*, pages 255–258, 2006.

[25] Apprentissage Transductif and Arnaud Revel. Apprentissage semi-supervisé.

[26] Bo Zheng, Philippe Ciais, Frederic Chevallier, Emilio Chuvieco, Yang Chen, and Hui Yang. Increasing forest fire emissions despite the decline in global burned area. *Science advances*, 7(39) :eabh2646, 2021.