THESIS

A platform and ontologies for environment data sharing and the use of Machine Learning models for wildfire ignition and prediction.

Jesús Noel Suárez Rubí

Brasilia, December 2022

FACULTY OF TECHNOLOGY

BRASILIA UNIVERSITY
Faculty of Technology
Department of Electrical Engineering

THESIS

A platform and ontologies for environment data sharing and the use of Machine Learning models for wildfire ignition and prediction.

Jesús Noel Suárez Rubí

Thesis submitted to the Electrical Engineering Department as a partial requirement to obtain the degree of PhD in Electrical Engineering

Examination Board

Paulo Roberto de Lira Gondim, D.C., FT/UnB ________________________________
Orientador

Joel José Puga Coelho Rodrigues, Ph.D, Instituto de Telecomunicações, Portugal ________________________________
Examinador externo

Sandra Sendra Compte, Ph.D., Universidade de Granada, Espanha ________________________________
Examinador externo

Adson Ferreira da Rocha, Ph.D, FT/UnB ________________________________
Examinador interno

FICHA CATALOGRÁFICA

RUBÍ, JESÚS NOEL SUÁREZ
A platform and ontologies for environment data sharing and the use of Machine Learning models for wildfire ignition and prediction. [Distrito Federal] 2022.
xvi, 123 p., 210 x 297 mm (ENE/FT/UnB, PhD, Engenharia Elétrica, 2022).
Thesis - Brasilia University , Technology Faculty.
Department of Electrical Engineering

1. IoT
2. Ontology
3. Wildfire
4. Machine Learning

I. ENE/FT/UnB
II. Título (série)

REFERÊNCIA BIBLIOGRÁFICA

RUBÍ, J.N.S. (2022). *A platform and ontologies for environment data sharing and the use of Machine Learning models for wildfire ignition and prediction.* Publication PPGEE .TD 191/22. Thesis, Department of Electrical Engineering, Brasilia University, Brasília, DF, 123 p.

CESSÃO DE DIREITOS

AUTOR: Jesús Noel Suárez Rubí
TÍTULO: A platform and ontologies for environment data sharing and the use of Machine Learning models for wildfire ignition and prediction.
GRAU: PhD in Electrical Engineering ANO: 2022

Jesús Noel Suárez Rubí
SQS 403 Bloco C apto 202, Asa Sul.
70237-030 Brasília - DF - Brasil

ACKNOWLEDGEMENTS

Agradeço a Deus por ter me dado a oportunidade de passar por esta experiência.

Ao meu orientador, pela sua paciência, pelo apoio, por ter me recebido como seu orientando e ter me guiado no desenvolvimento do projeto, Prof. Paulo muito obrigado por todo e por estar sempre presente e a disposição constante para ajudar. Vou sentir falta das nossas conversas 10 hrs da noite e dommingos 8hrs da manhà rsrsrs.

A minha esposa por estar ao meu lado, pela sua compreensão, pela sua ajuda e pela sua confiança em min. A meus pais, minha sogra e resto da família por ter me apoiado neste tempo que tenho estado longe deles, mas mesmo estando longe me fizeram sentir que estavam por perto, me cuidando e me acompanhando.

Às pessoas que conheci em Brasília e às que conviveram comigo no dia a dia fazendo me sentir como em família com elas.

Ao pessoal do Laboratório de Redes da UnB e ao Programa de Pós-graduação em Engenharia Elétrica da UnB. À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pelo apoio financeiro. A todos os que de algum jeito se interessaram pelo sucesso deste trabalho.

Muito obrigado, Noel.

RESUMO EXPANDIDO

Título: Uma plataforma e ontologias para compartilhamento de dados ambientais e uso de modelos de aprendizado de máquina para predição de incêndios florestais.
Autor: Jesús Noel Suárez Rubi
Orientador: Prof. Dr. Paulo Roberto de Lira Gondim, FT/UnB
Programa de Pós-Graduação em Engenharia Elétrica - PPGEE

Ecossistemas, assentamentos e vidas humanas são colocados em risco por incêndios florestais todos os anos, impactando a economia e o desenvolvimento socioeconômico. O Distrito Federal brasileiro, inserido no bioma Cerrado, vem apresentando um aumento desses fenômenos. No entanto, poucos estudos foram realizados na região.

Vários modelos têm sido propostos mundialmente para a predição da ocorrência e comportamento do fogo, permitindo a identificação dos fatores que os favorecem, os riscos e pós-efeitos. A aplicação direta de tais modelos na região do Distrito Federal é desafiadora devido às diferenças nas fontes de dados, características geográficas das regiões e indisponibilidade de dados em alguns casos.

Por outro lado, o uso de tecnologias de informação e comunicação e a ampla disseminação de equipamentos eletrônicos (por exemplo, redes de sensores e terminais celulares) são essenciais para o tratamento adequado de grandes volumes de dados com valor substancial para o desenvolvimento de cidades inteligentes. Particularmente, os dados ambientais de cidades inteligentes podem enriquecer os estudos sobre incêndios florestais. No entanto, propostas recentes têm enfrentado a mesma desvantagem, pois os dados são incompletos, seguem diferentes formatos de representação e até possuem diferentes conotações semânticas.

A heterogeneidade de objetos inteligentes conectados à Internet (ou seja, interfaces de rede, protocolos de comunicação, estrutura de dados, precisão de aquisição e semântica de dados) tem causado problemas de interoperabilidade, dificultando a eficácia dos sistemas de apoio à decisão intimamente relacionados à qualidade dos dados. A aplicação de algoritmos de big data e aprendizado de máquina para melhorar os processos relacionados a cidades inteligentes são alguns dos exemplos impactados negativamente pela falta de padrões.

As soluções para cidades inteligentes devem garantir a interoperabilidade desde a captura de dados até a extração e visualização do conhecimento por meio de tecnologias como Web Semântica e ontologias. Além disso, os componentes envolvidos devem incluir dispositivos IoT, gateways, computação em nuvem e em névoa para uma melhor aplicação das técnicas de análise de dados

Nesse sentido, esta tese propõe uma plataforma de cidade inteligente para monitoramento da qualidade ambiental baseada em tecnologias semânticas e ontologias, possibilitando um sistema de coleta e compartilhamento de dados multidefinição e multiprotocolo. Ela também implementa uma metodologia para a extração de *insights* sobre os dados coletados e um mecanismo para cálculos baseados em nuvem e névoa. Além disso, são propostas ontologias para a representação semântica e definição do esquema de armazenamento considerando cidade inteligente, internet das coisas florestais (IoFT) e terminologia relacionada ao fogo.

Oito modelos de aprendizado de máquina foram comparados na predição do risco de incêndios florestais na região mencionada. Eles consideraram correlações entre condições climáticas, localização espacial, características topográficas, características antropogênicas e ocorrência de incêndios e um conjunto de dados enriquecido com dados abertos do governo brasileiro composto por observações sobre 16 características climáticas de cinco estações de monitoramento, e dados de satélite sobre incêndios ocorridos nas últimas dois décadas. Características topográficas, hidrográficas e antrópicas, como Índice de Vegetação por Diferença Normalizada (NDVI), índice de urbanização e distância a rios/estradas também foram consideradas. De acordo com os resultados, o risco de incêndio pode ser previsto com 99% de precisão e os modelos se mostraram mais sensíveis ao NDVI, pressão atmosférica e umidade relativa, conforme demonstrado por um estudo sobre o impacto das feições.

Outro conjunto de dados foi compilado a partir de dados abertos do governo brasileiro para a predição do comportamento dos incêndios florestais e usado para o treinamento de vários modelos de aprendizado de máquina que consideram o ponto de ignição do fogo para prever as áreas que serão impactadas. Inclui observações sobre características climáticas de cinco estações de monitoramento e dados de satélite sobre incêndios ocorridos nas últimas duas décadas, enriquecido com características topográficas, hidrográficas e antropogênicas. De acordo com os resultados, o modelo AdaBoost previu a área afetada pelo incêndio florestal com 91% de precisão, mostrando melhor desempenho do que Random Forest (RF) 88%, Artificial Neural Network (ANN) 86% e Support Vector Machine (SVM) 81%. Um método "wrapper" permitiu o cálculo da importância das variáveis e a definição de um ranking para identificar o quanto uma variável influencia o risco e o avanço do incêndio.

Como resultado, a plataforma de monitoramento ambiental foi desenvolvida e testada quanto à predição de propagação e comportamento de incêndios florestais em um momento específico e/ou em regiões específicas para auxiliar os órgãos de gestão de incêndios a minimizar os danos causados. Tal estudo de caso mostrou a aplicação do aprendizado de máquina como o principal fator para melhorar os estudos de risco e comportamento de incêndio, impactando diretamente na sustentabilidade dos ecossistemas e promovendo diversas melhorias no estudo de incêndios na região do Distrito Federal.

Palavras-chave: Internet das Coisas, Ambiente, Incêndios Florestais, Ontologias, Predição, Aprendizado de Máquina.

ABSTRACT

Ecosystems, settlements, and human lives are put at risk by forest fires every year, impacting economy and social-economic development. The Brazilian Federal District, inserted on the Cerrado biome, has shown an increase in such phenomena. However, few studies have been conducted in the region.

Several models have been proposed worldwide for the prediction of fire occurrence and behavior, and identification of their conditioning factors, risks, and post-effects. The direct application of such models in the Federal District region is challenging due to differences in data sources, geographic characteristics of the regions, and unavailability of data in some cases.

On the other hand, the use of information and communication technologies and the broad dissemination of electronic equipment (e.g., sensor networks and cellular terminals) are essential for the adequate treatment of large volumes of data with substantial value for the development of smart cities. Particularly, environmental smart city data can enrich wildfire studies. However, recent proposals have faced the same downside, since data are incomplete, follow different representation formats, and even have different semantic connotations.

The heterogeneity of intelligent objects connected to the Internet (i.e., network interfaces, communication protocols, data structure, acquisition precision, and data semantics) has caused interoperability problems, hindering the effectiveness of decision-support systems closely related to the quality of data. The application of big data and machine learning algorithms for improving smart city-related processes are some of the examples negatively impacted by the lack of standards.

Solutions for smart cities should grant semantic interoperability from data capture to knowledge extraction and visualization through technologies such as Semantic Web and ontologies. Moreover, the components involved should include Internet of Things (IoT) devices, gateways, cloud, and fog computing for a better application of data analysis techniques.

In this sense, this thesis proposes a smart city platform for environment quality monitoring based on semantic technologies and ontologies, enabling a multi-definition and multi-protocol data collection and sharing system. It also presents a methodology for the extraction of insights into the collected data and a mechanism for cloud- and fog-based computations. Moreover, ontologies are proposed for the semantic representation and storage scheme definition considering Smart Cities (SC), Internet of Forestry Things (IoFT), and fire-related terminology.

This study compares eight machine learning models that predict wildfire risk worldwide so that they can be adopted in the aforementioned region. They considered correlations among climate conditions, spatial location, topographic features, anthropogenic characteristics, and fire occurrence. A dataset enriched with Brazilian governmental open data was comprised of observations on 16 climate features of five monitoring stations and satellite data on fires occurred over the

past two decades and topographic, hydrographic and anthropogenic features, such as Normalized Difference Vegetation Index (NDVI), urbanization index, and distance to rivers/roads. According to the results, fire risk can be predicted with 99% accuracy and the models showed more sensitive to NDVI, atmospheric pressure, and relative humidity, as demonstrated by a study on the impact of features.

Another dataset was compiled from Brazilian governmental open data for the prediction of the wildfire behavior and used for the training of five Machine Learning models that consider the fire point of ignition to predict the areas that will be impacted. It includes observations on climate features from five monitoring stations and satellite data on fires that occurred over the past two decades and was enriched with other topographic, hydrographic, and anthropogenic features, such as urbanization index, distance to rivers/roads, and Normalized Difference Vegetation Index (NDVI). According to the results, AdaBoost model predicted the area affected by the wildfire with 91% accuracy, showing better performance than Random Forest (RF) 88%, Artificial Neural Network (ANN) 86%, and Support Vector Machine (SVM) 81%. A wrapper method enabled feature importance calculation and definition of a rank that determines the influence of a variable on the fire risk and its advance.

As a result, the environment monitoring platform has been developed and tested regarding the prediction of both spread and behavior of wildfires at a specific time and/or in specific regions for helping fire management agencies minimize the damages caused. Such a case study showed the application of machine learning as the main factor for improving fire risk and behavior studies, directly impacting the sustainability of ecosystems and promoting several improvements in the study of fires in the Federal District region.

SUMMARY

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

Symbols

$aspect_{100}$	Aspect in 100-m resolution
$aspect_{30}$	Aspect in 30-m resolution
$\hat{B}_j$	Coefficient of $j-th$ feature in a linear regression model
C	A conceptualization represented by <D, W, R>
D	Domain
I	Intersection of predicted and known scars in SS calculation.
K	Known scar area (real sample) in SS calculation.
$R^2_{X_j\|X_{-j}}$	Coefficient of determination of the regression equation of the first step
m	Number of features considered after the application of a feature selection method
n_f	Number of features collected
n	Sample size
P	Predicted scar area in SS calculation.
RH	Relative Humidity
R	set of relations that represents a domain D
$r_{x,y}$	Correlation between variables x and y
S	Observations temporal serie
sd	Initial date considered by downsampling algorithm
s_{30}	30-m resolution sectors set
s_{100}	100-m resolution sectors set
$slope_{100}$	Slope in 100-m resolution
T	Temperature
W	Represents all existing concepts in D
ws	Downsampling Windows Size
x_i	Sample i of variable x
$\bar{x}$	Mean of variable x
y_i	Sample i of variable y
$\bar{y}$	Mean of variable y

ACRONYMS LIST

Acronyms

5G	5th generation mobile network
AdaBoost	Adaptive Boosting
AMQP	Advanced Message Queuing Protocol
ANN	Artificial Neural Network
API	Application Programming Interface
AQM	Air Quality Madrid
AUC-ROC	Area Under the Curve Receiver Operating Characteristics
BRT	Boosting Regression Trees
COAP	Constrained Application Protocol
DOCSIS	Data Over Cable System Interface Specification
EISCO	Environmental Indicators SC Ontology
eMBB	Enhanced Mobile Broadband
ENVO	Environmental Ontology
EQMS	Environment Quality Monitoring Station
FD	Federal District
FPR	False Positive Rate
GCIO	Global City Indicators Ontology
GPSCO	General Purpose SC Ontology
HTTP	Hyper Text Transfer Protocol
IoFT	Internet of Forestry Things
IoT	Internet of Things
IRI	International Resource Identifier
ISO	International Organization for Standardization
JSON	Java Script Object Notation
LST	Land Surface Temperature
LTE	Long Term Evolution
ML	Machine Learning
MLP	Multi Layer Perceptron
mMTC	Massive Machine Type Communications
MODIS	Moderate Resolution Imaging Spectroradiometer
MQTT	Message Queue Telemetry Transport
NDVI	Normalized Difference Vegetation Index
OWL	Ontology Web Language
QoL	Quality of Life

RBF	Radial Basis Function
RDF	Resource Description Framework
ReLU	Rectified Linear Unit
REST	Representational State Transfer
RF	Random Forest
SC	Smart Cities
SenML	Sensor Measurement List
SMSE	Structural Mean Square Error
SOSA	Sensor, Observation, Sample, and Actuator
SPARQL	Protocol and RDF Query Language
SS	Sørensens Similarity
SSN	Semantic Sensor Network
SVM	Support Vector Machine
SWEET	Semantic Web for Earth and Environmental Technology
TPR	True Positive Rate
TWI	Topographic Wetness Index
URLLC	Ultra-Reliable and Low-Latency Communications
VIF	Variance Inflation Factor
VPD	Vapor Pressure Deficit
VPS	Vapor Pressure Saturation
WLAN	Wireless Local Area Network
XML	Extensible Markup Language
XMPP	Extensible Message Presence Protocol

1 INTRODUCTION

The rapid growth of urban populations has demanded studies that identify, prevent and act in situations of threatened Quality of Life (QoL) [6]. In Smart Cities (SC) [7], QoL is commonly dealt with by indicators that measure the effectiveness of services and sustainability of a city in domains/verticals, such as Environment, Healthcare, Security, Transport, Economy, Education, and Government.

SC technologies can act together towards improving such indicators through a more efficient use of resources. Its strong correlation with the sustainability principles can enhance QoL indicators, and data collected in the physical, digital, and biological processes can be used to promote the emergence of new business models for a more efficient management and more assertive decision-making.

Moreover, as a consequence of the progressive digitization and the high dissemination of access and sensor-based networks, the enormous amount of data collected from multiple sources indicates the need for adequate treatment of data. Big Data, for example, can enable analyses of large volumes of information gathered and, consequently, the development of solutions and advances with greater impact and benefit for the context of the city, thus improving the QoL indicators.

In this sense, SCs must provide interoperable tools that collect, store, and disseminate industry- and city-related data, and several sensors, frameworks, and SC platforms have emerged for such purposes. However, the lack of standards (e.g., their sharing and a common data format) has imposed several challenges, thus hampering the application/reuse of SC technologies. Geolocation concepts (e.g. address, buildings, local region or city) can be handled in different ways. The same is applicable for sensor measurements, according to which two different sensors can monitor the same parameter using different units of measure. Moreover, data may be out of date and/or defined as aggregated statistical data, which might hinder the application of real-time studies and development of time-restricted decision-support.

Another challenge refers to heterogeneity in networking and sensor technologies. The development of various connected physical world objects that form the Internet of Things (IoT) [8] has led to a heterogeneous environment of IoT devices/platforms that must be integrated into an interoperable one. Regarding SCs environmental data, a growth is expected in the number of IoT platforms that deploy sensors related to indicator data, and their integration must be considered in the backbone of environment-related city services.

The lack of a standardized definition of environmental indicators directly impacts interoperability among systems from different providers. ISO 37120 [9] has defined indicators that include fine particular matters (PM2.5 and PM10), emissions of greenhouse gases (such as nitrogen dioxide (NO2), sulfur dioxide (SO2), and ozone (O3)), acoustic contamination (AC), change in the percentage of native species, quality of water, and waste management services. However, the standard does not consider, for example, clear metrics for water quality (i.e. temperature, pH, turbidity, conductivity, and dissolved oxygen indicators) measured in IoT platforms, such as the ones proposed by Vijayakumar and Ramya [10] and Encinas et al. [11].

Current SC platforms have involved forest ecosystems, which has raised concerns from the scientific community. Uçar et al. [12] discussed the importance of urban forestry and urban greening and their relation to the QoL in SC. Urban forests improve air and water quality and act as a temperature stabilizer. However, the lack of tools for their monitoring has impacted the development of government policies and strategies.

Sharma et al. [13] highlighted the importance of agricultural land, open spaces, and extensive forest surveillance, since cities are scattered or surrounded by them. Any imbalance in such ecosystems can deteriorate the QoL, threatening the inhabitants of the cities, as in the Brazilian Federal District (FD) region studied in this research, which is inserted into the Cerrado savanna and is immersed between some conservation units (Figures 1.1 and 1.2).

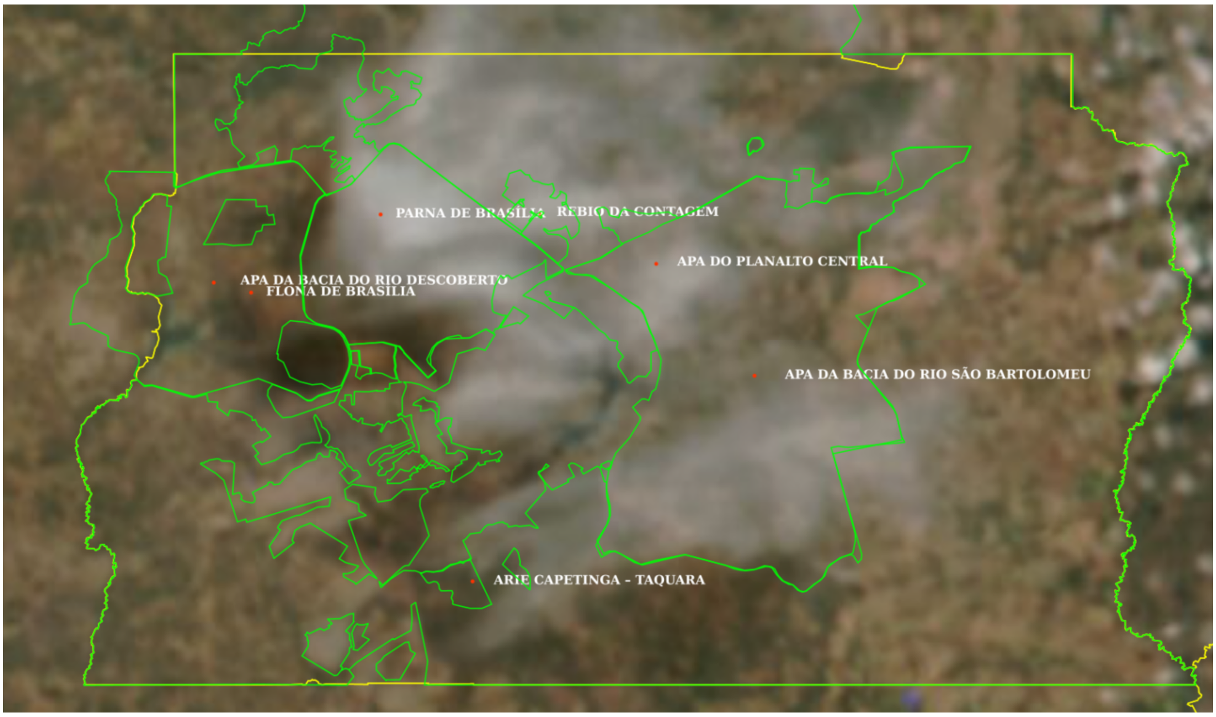


Figure 1.1: FD Conservation Units [1]

On the other hand, the Brazilian forest industry plays a key role in the economy of the country, representing 4% of total gross domestic product and generating approximately six million jobs [14]. In the coming years, forestry will face several issues moved by the increasing global competition in world markets [15].

The increase in the number of forest fires due to the unsustainable exploitation of forests and agricultural activities has caused the degradation of both Brazilian forests and the worldwide environment. Therefore, the assurance of environmental sustainability of production processes

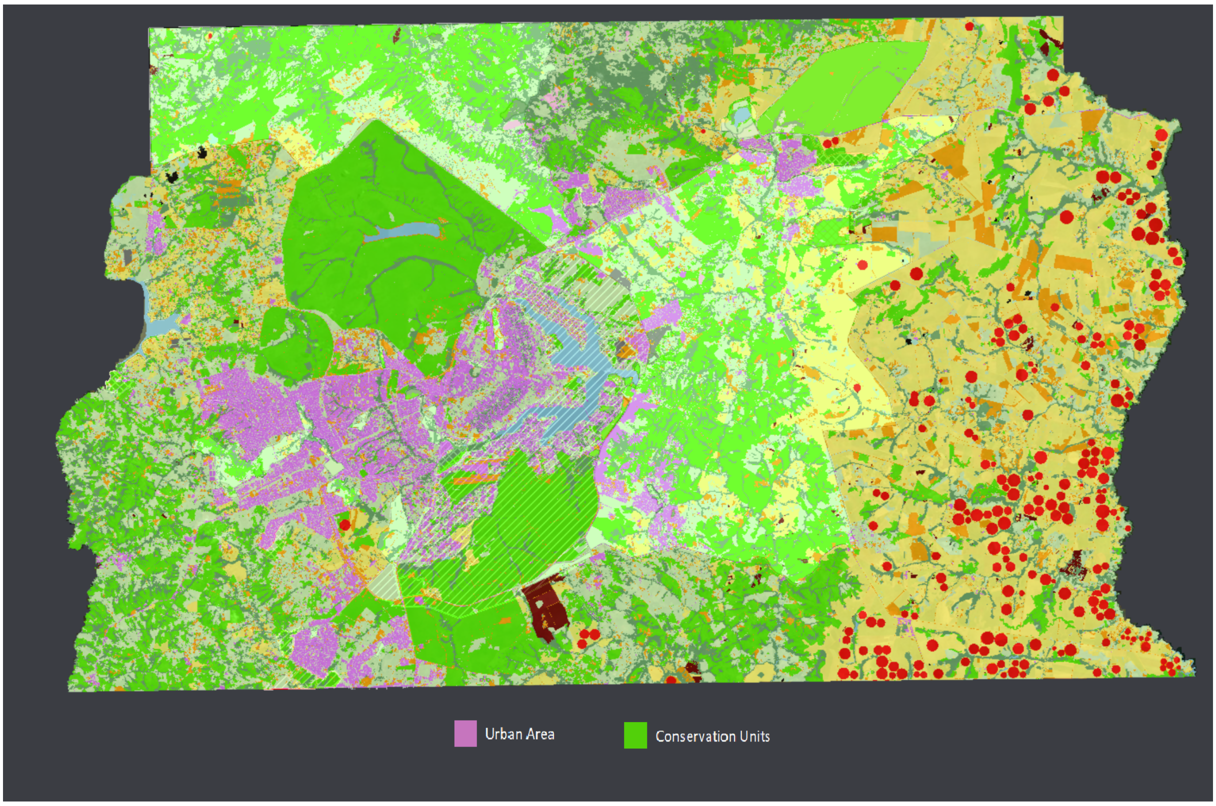


Figure 1.2: FD Conservation Units vs Urban Area [2]

and preservation of ecosystems and natural resources will be challenging. Forest fires, which have increased in Brazil due to climate changes and human impact, are one of the examples that require improvements in models for a better understanding of procedures for their prevention and combat [16].

An extension of forestry sensor networks, satellite images, and emerging technologies such as IoFT to the SC context will stimulate the development of studies of forest indicators towards a better understanding of their dynamics [17]. We believe Forestry 4.0, IoFT devices, and sensor networks, adopted for measurements of variables such as humidity, temperature, and carbon dioxide sensors [18] will be able to monitor environmental changes and prevent and combat forest fires.

IoFT solutions related to the environment domain also lack semantic interoperability. The available semantic tools (e.g., ontologies and resource-description frameworks, such as SSN and Resource Description Framework (RDF)) can contribute to the standardization of the semantic representation of sensors/platforms, unit of measures, locations, time, and other concepts frequently used in a sensing environment. However, no clear definition or consensus has been achieved by normative organizations on the use of semantic tools for providing semantic interoperability in a broad sense.

Environmental studies can comprehend many phenomena through the observation/analysis of different features (i.e., an earthquake can be predicted by vibrations (seismograph) or satellite image processing). Therefore, a platform for environmental studies must model and explore the semantic relation among heterogeneous data sets including those collected by Forestry 4.0 systems. Semantic tools (i.e. ontologies, linked data, reasoners, among others) enable the explicit modeling of such a relationship. Moreover, some issues on compatibilization of data formats must be solved.

1.1 MOTIVATION

The QoL of the inhabitants of the Federal District (FD) region has been negatively affected by the fire activities, which has led the government to spend resources on their fighting. The past decades have witnessed an increment in the number of fire spots, according to satellite data, thus highlighting the importance of fire-related studies on the region (see Figure 1.3).

The FD region is inserted in the Brazilian savanna (the Cerrado biome), comprised of 11,627 species of plants and which has been affected by a large number of fires. The dry climate together with the savanna vegetation create a favorable scenario for fire dissemination. Therefore, research into forest fires in the FD will both leverage the local firefighting decision-making and policies and probably decrease the number of fires in the Cerrado region.

No consensus on modelling methodologies of forest fire behaviour has been achieved. Although approaches involving complex mathematical models have been published ([19], [20], and [21]), their static characteristics hamper the representation of highly dynamic processes such as fire-line progress. Most of those empirical and semi-empirical models have been applied in laboratories and controlled field-scale experiments, which commonly consider two types of numerical approaches. The first is based on the complex modelling of physical and chemical processes ([21], [22]), whereas the other involves the rate of spread-correlating features, such as slope, wind, and vegetation type [23]. However, both have showed poor accuracy in real fire events and required high computational costs and simulation times, which are impractical for real-time decision support.

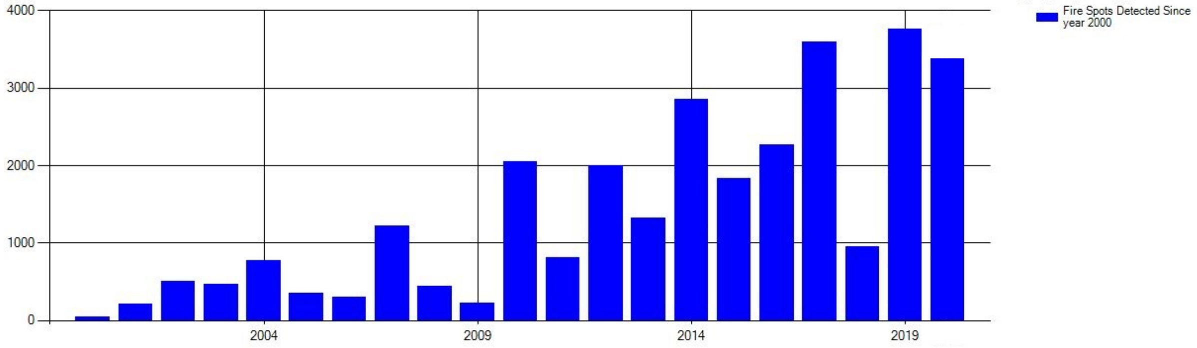


Figure 1.3: Number of fire spots per year (2000-2020)

The reproduction of prediction results is usually difficult due to the unavailability of data for training and their relationship with the quality of the results. Datasets available for each region show variations in number of input features (fire behaviour drivers), temporal and spatial resolutions, and sensors involved in acquisition. Other drivers such as anthropogenic impact can influence the behavior of fires in a populated region, but exert almost no effect on others.

Various approaches have compared the performance of ML models towards adjusting their hyper-parameters and finding one of best results. In this sense, a feature selection technique must be considered for the identification of features that effectively contribute to the accuracy of wildfire behaviour/spread prediction.

The literature lacks studies on the prediction of wildfire behaviour on lower temporal scales that aim at the avoidance of in-place inferences that consider short-term dynamics. Moreover, few studies have addressed the impact of climatic, anthropogenic, topographic, and vegetation conditioning factors on such a behaviour. [24], [25] and [26] treated them as isolated variables, and no joint study has identified whether, in fact, each variable by itself exerts a substantial impact on fire behavior.

The literature also lacks datasets for studies on short-term fire behaviors, specially for the FD region.

Despite the existence of several sources of government data related to the occurrence of forest fires, no platform for their extraction and collection in a simple way is available. Such an issue has motivated our proposal of a platform framed in the concepts of smart cities for the collection, monitoring, and processing of the parameters related to the city and the surrounding areas inserted in the Federal District, as a part of the Cerrado biome. In this sense, and since Brasilia is immersed between conservation units, the platform has been extended for the treatment of forestry-related parameters.

The aforementioned issues have motivated this research, whose aim is to show fire behavior can be predicted from both the coordinates (latitude and longitude) of an ignition point and the historical evolution of spatial and temporal data in the region in a short term, thus facilitating the determination of its most influential features. The way a standardized SC platform can help in the early processing of wildfire-related data is also addressed.

1.2 THESIS SCOPE

This thesis proposes a platform (Figure 1.4) for SC that allows the registration of sensors, the aggregation and transparent exchange of data, and the application of machine learning algorithms for decision support. The research considers the feasibility of semantic web-based data representation models and ontologies as a data schema. The aims are the identification of the best platform architecture, the construction and evaluation of a SC platform and the use of physical and application layer communication protocols for performance comparison and choice of best alternative.

On the other hand, the feasibility of predicting events and behavior of fires in the Brazilian FD region is studied through analyses of the ecology of fires in the region and the main originating factors. Based on these factors, heterogeneous data sources are identified and integrated into the SC platform through the proposal and use of ontologies for semantic representation, allowing the adoption and evaluation of machine learning models and identification of the model of best performance in wildfire risk and behavior predictions.

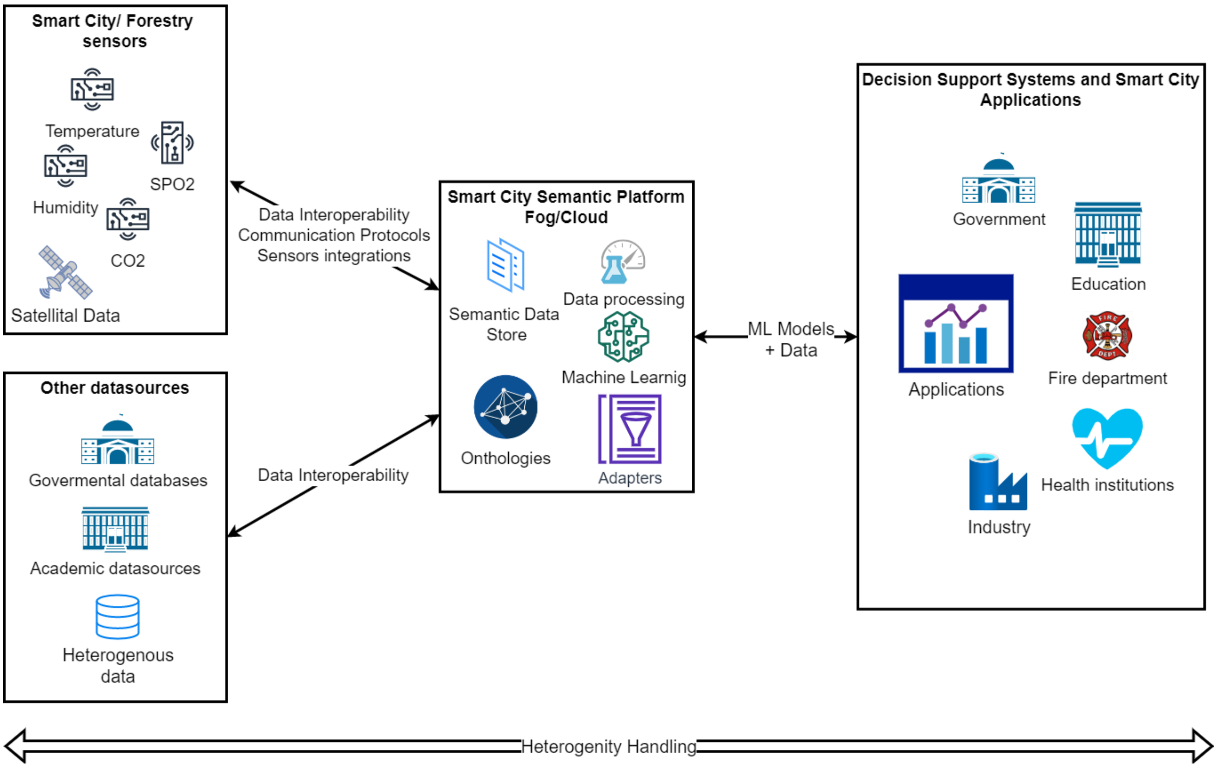


Figure 1.4: Scope of the thesis

1.3 OBJECTIVES

1.3.1 General Objectives

The general objectives of this research involve proposal of ontologies that facilitate the collection and exchange of data in the context of smart cities, with emphasis on the environmental vertical for the design and validation of a semantic platform for environmental monitoring and, as a case study, prediction of both risk and behavior of forest fires in the Federal District Region supported by the semantic integration of data and machine learning.

1.3.2 Specific Objectives

- Design of a general purpose ontology that represents the SC terminology and the technologies involved;

- Proposal of ontological extensions with the terminology of the environment vertical and related to forest fires;
- Proposal and evaluation an IoT-based semantic platform for data collection and exchange in smart cities;
- Integration of heterogeneous communication protocols and data representation formats involved in the data sharing process;
- Collection and analyses of data relevant to forest fires in the Federal District region and proposal of a bench-marking dataset;
- Application of machine learning models for wildfire risk and behaviour prediction in the Federal District ;and
- Ranking of the importance of each input feature for the identification of those that effectively increase the accuracy of wildfire behaviour/spread prediction.

1.4 RESEARCH METHODOLOGY

We followed the methodology described below for the meet the objectives presented in the previous section.

Regarding the proposal of an SC platform: the seamless integration of the heterogeneous IoT/IoFT devices and external datasources demanded the study, implementation and comparison of communication protocols at physical and application layers. Moreover, studies on SC platform architectures and sensor observation data representation format were also required for the definition of a data interchange format and sensor bindings. A review on ontologies related to SCs was made and some ontologies were proposed, leading to the definition of a dynamic data scheme and the implementation of an aggregation datastore.

Regarding the implementation of a wildfire risk and behaviour prediction application: fire ecology was studied to identify the main wildfire related variables. Particularly, for the FD region a set of explanatory features was considered based on a literature review of fire drivers. The most employed machine learning models were also identified and we compare them to validate their performance. To measure how the considered originating factors could describe the fire effects we made a variable importance study. Finally, algorithms for the prediction of point of ignition and for the progressive reproduction of fire scars were proposed.

1.5 CONTRIBUTIONS

The contributions of this research involve:

1. An ontology that represents indicators and environment SC terminology for extending and improving other ontologies previously published [3]. A wildfire ontology has been extended towards the implementation of wildfire behavior use-case in the Federal District Region.

2. Definition, implementation, and testing of a semantic IoT-based platform that can cover several SC verticals and enables semantically coherent data interchange and processing [3]. The platform exposes management Application Programming Interface (API) and promotes the integration of data through several communication protocols [3]. A performance study about physical and application layer protocols is also reported.

3. Extension of the platform towards covering Forestry and IoFT concepts in a use case of wildfire behaviour prediction [4].

4. A review of recent research on wildfire prediction for the identification of the main features and proposal of an open dataset for the FD region [4]. A set of short- term spatial/temporal data sequences is also provided and represents the behavior/spread of fire in the region originated from the history of fire scars and ignition points enriched with topographic, climatic, anthropogenic, vegetation, and environmental measurements.

5. A review of the main ML models employed for wildfire risk probability prediction and validation of their feasibility and performances compared according to different validation metrics [27].

6. Analysis of whether the neighboring conditions of fires can be the basis for the dynamic prediction of their spread direction by four machine learning models, namely Deep Artificial Neural Network, Support Vector Machine, Random Forest, and Adaptive Boosting (AdaBoost) [4]. Each model was subjected to a feature selection process that identified the most relevant features based on their importance (calculated by a permutation method). The models are ranked according to their performance considering Area Under the Curve Receiver Operating Characteristics (AUC-ROC), F1 score, accuracy, and recall metrics.

7. Construction of various fire scars according to data predicted and analysis of the precision of the predicted burned areas.

All contributions either have already been published, or are in a review phase in high-reputation journals, as shown in Table 1.1, where each indicated appendix contains the 1st page of the respective paper.

1.6 THESIS STATEMENT

Smart Cities and Forestry 4.0 trends have reached several areas, and their joint adoption has led to enhancements in citizens' quality of life. Data collected by them are crucial for improving solutions to several QoL-related problems and contribute to the implementation of strategies of

Table 1.1: Publications

Contribution	Title/Journal	Status	Appendix
1, 2	*IoT-based Platform for Environment Data Sharing in Smart Cities.* International Journal of Communication Systems.	Published	A
3, 4, 6	*Forestry 4.0 and Industry 4.0: Use Case on Wildfire Behaviour Predictions.* Computers and Electrical Engineering Journal.	Published	B
5	*A Performance Comparison of ML Models for Wildfire Risk Prediction in the Brazilian Federal District Region.* Environment Systems and Decisions Journal.	Submitted	D
6 (Extension), 7	*Application of machine learning models in the behavioral study of forest fires in the Brazilian Federal District region.* Engineering Applications of Artificial Intelligence Journal.	Published	C

decision-making systems. However, dealing with the heterogeneity introduced by such types of systems is challenging, and ontologies can be considered towards solving such an issue.

The development of a platform for the collection, storage, and processing of data from the heterogeneous technologies involved and that follow a consistent and standardized semantic data model improves and simplifies the application of big data and machine learning techniques.

One scenario that can take advantage of such type of platform is related to forest fire studies and fire fighting strategies in the Federal District region, which have been affected by the lack on data related to those events. Therefore, new methods for the collection, analysis, and exchange of fire-related data must be developed.

This thesis proposes a platform for the collection and ontology-based sharing of Forestry and Smart City data in the environmental vertical and analyzes wildfires in the Federal District region are analyzed for the prediction of fire risk and behavior based on ML theory, studying and identifying their main originating factors.

1.7 ORGANIZATION

The remainder of this study is organized as follows:

Chapter 2 presents the state-of-the-art of ontologies related to SC, environment indicators, wildfires, and forestry terminologies. Moreover, semantic web studies are discussed and a general purpose SC ontology is introduced towards seamless integration and representation of SC data.

Chapter 3 discusses studies on SC platforms with a focus on platform architecture, communication standards, and data representation and exchange formats and is also devoted to the proposal, development, and validation of a semantic platform based on IoT for data collection and exchange of Smart Cities and Forests. The platform was validated by performance metrics obtained in different experimental scenarios.

Chapter 4 addresses the contributions in the study on wildfires risk and behavior prediction in the Federal District Region. Initially, a review of the wildfire-related literature identified a common methodology for the assessment of wildfires predictions. The ecology of the Cerrado Biome is then characterized and the main fire originating factors are discussed. Several governmental datasources related to wildfire explanatory features were integrated to the platform and ML models were trained and validated regarding prediction of fire occurrence (fire-risk maps) and fire behaviour (fire scars).

Chapter 5 provides some conclusions about the proposed solutions and some of the still opened issues are listed and outline suggestions for the next steps of research.

The final part contains the bibliographic references that founded this research and four appendices that present the resulting studies in the form of articles, published or under review in JCR-ranked journals.

2 ONTOLOGIES AND SEMANTIC WEB

This chapter addresses a literature review of ontologies and semantic web use for the semantic integration of SC data and proposes a GPSCO extended with the Environmental Indicators SC Ontology (EISCO) and wildfire-related terminology. Moreover, the representation of specific entities related to Forestry is introduced.

2.1 BASIC CONCEPTS

Ontologies are some of the most suitable tools for the representation of concepts and knowledge and have been widely applied for the reuse of data.

The literature reports several definitions for the term ontology. According to Gruber [28], an ontology is a specification of a conceptualization, e.g., a description of concepts and relationships between those concepts, and Euzenat et al. [29] defined it as a vocabulary specific for a domain of interest and a specification of the meaning of terms in that vocabulary. However, both [28] and [29] stated ontologies are constructed for the sharing and reuse of data and knowledge. In computation, the concepts present in ontologies must have a formal specification and Guarino [30] suggested a way for the creation of formal definitions for concepts.

Let <D, W> be such that D represents the domain in question and W represents all existing concepts in D. This structure is called Domain Space. A conceptualization C is a structure <D, W, R> where R is the set of relations that represents the domain. A conceptualization defines an intended structure of the world, represented by C. From the computational point of view, a conceptualization must be specified in a particular language L and predicates must be logically consistent with an interpretation function.

Therefore, ontologies become partial specification mechanisms representative only in relation to a given domain, and not to the completeness of knowledge. Particularly, the SC context comprehends several specific subdomains (e.g., healthcare, transport, energy, education, among others) and many proposals are focused only on the definition of integration mechanisms and ontologies for that subdomain (Section 2.1 covers some related work that validate the previous statement). Subdomains segmentation affects interoperability in a global proposal for SC and different definitions of same concepts in each subdomain require a process of alignment among all ontologies.

Euzenat et al. [29] formalized the ontology matching process as the construction of a set of correspondence rules between concepts provided by two or more ontologies towards solving the semantic heterogeneity present in multiple definitions for the same concept in different domains. Operations that use the set of correspondence rules must be merged for the obtaining of a new ontology that contains the concepts defined by the input ontologies aligned in a domain that includes the input subdomains and the region where the logic assertions between the subdomains are fulfilled. If both ontologies C_1 and C_2 are described in a same computer language, the resulting ontology can be a C_3 structure, where $D_3 = \{D_1 \cup D_2\}$, $W_3 = \{W_1 \cup W_2\}$, and $R_3 = \{R_1 \cup R_2 \cup R_a\}$, where R_a corresponds to the rules for the matching process between D_1 and D_2 compatible with the logic-consistent associations inside C_1 and C_2.

2.2 RELATED WORK

Regarding environment, INFORMEA [31], in the Law and Environment Ontology provides environmental terms and links to related laws, as well as definitions for air pollution, water quality, and gas emission, which help the standardization of a model for environmental data. However, it considers the law aspect of environmental data collection, rather than a technical specification of environmental parameters. On the other hand, the Air Pollution Ontology [32] primarily focused on pollutants such as SO2, NO2, CO2, and PM2.5, but ignored other key definitions (e.g., sensors and indicators).

The Semantic Web for Earth and Environmental Technology (SWEET) [33] is a mature ontology that contains over 6000 concepts that cover all major definitions for SC environmental indicators. However, it lacks specifications about observation process, unit of measures, geolocation, and descriptions of the classes and their properties. The ENVO [34] covers all the classes defined in SWEET that are useful for an SC environment platform and has a more detailed conceptualization. Nevertheless, similarly to SWEET, ENVO does not cover topics related to observation of the process, unit of measure, and other important concepts for the definitions of indicators (e.g., provenance, validity, and time).

On the other hand, ISO 37120 [9] has defined indicators that enable evaluations of city services and quality of life and provides a reliable foundation of globally standardized data. Fox [35] proposed the GCIO, which covers the ISO 37120 specifications and adds definitions for the representation of supporting data that generate an indicator value. However, Fox [35] and Dahleh [36] demonstrated the standard does not cover all aspects relevant to an SC and, specifically for environmental indicators, considers only eight definitions. Although Fox [35] enhanced the indicator definitions including conceptualizations for place names, measurements, provenance, time, trust, and validity, those definitions do not consider other concepts, such as type of sensor used for the sampling, or manufacturer and observation parameters (e.g. sampling frequency).

SSN [37] is an ontology that describe sensors and their observations, procedures involved, features studied, samples of interest, and properties observed. It includes a self-contained core ontology called SOSA (Sensor, Observation, Sample, and Actuator) for its elementary classes and properties. SSN provides a technical conceptualization to represent sensors, platforms, and IoT devices and SOSA enables the representation of an indicator observation through classes, such as observation, observable property, measure, and result. However, SSN neither offers an indicator definition, nor considers geolocation, unit of measure, time, and other important concepts associated with an indicator.

Ganzha et al. [38] reviewed ontologies for IoT, e-Health and Transportation/Logistics and, regarding IoT, observed any fusion between IoT and semantic technologies would take advantage of SSN [37]. However, the research covered four independent ontologies and ten that extended SSN . Table 2.1 shows a summary of such ontologies.

Table 2.1: SC-related Ontologies

Ontology	Goal	Main Concepts Covered	Extends SSN
CSIRO Sensor Ontology [39]	Generic Sensor Ontology	Describes functional, physical, and measurement aspects of sensors through sensors, features, operations, results, processes, inputs and outputs, accuracy, resolution, abstract and physical properties, and metadata links classes	Maybe a Predecessor
MMI Device Ontology [40]	Ontology for marine devices	System, process, platform, device, sensor, and sampler	No
Extensible Observation ontology (OBOE) [41]	Ontology for the capture of the semantics of scientific observations and measurements	Observation, Measurement, Entity, Characteristics, Standard, Protocol	No
Sensor Cloud Ontology (SCO) [42]	Ontology that extends the sensor terminology with the Cloud Concepts	Extends SSN including The Observation and Measurement Ontology (OM) [43] and GEO Ontology (WGS84) [44] to provide terminologies and conceptualizations related to measures, unit of measures, and sensor geolocation	Yes
AEMET Ontology [45]	Ontology for Meteorological data representation	Extends ssn:sensor to cover specific sensors such as thermometers, barometers, among others. The authors use the OWL Time ontology [46] for time events and GEO Ontology (WGS84) [44] for sensor location, as in the previous ontology.	Yes
Sensor Web Resources Ontology for Atmospheric Observation (SWRO-AO) [47]	Ontology for Atmospheric Observation	Extends SSN concepts regarding sensors and adds specific domain concepts such as swroaao:weather_station	Yes
IoT Lite [48]	Lightweight ontology for the representation of Internet of Things (IoT) resources, entities, and services	Uses QU Ontology [49] to extend SSN providing quantity and unit of measure description and GEO Ontology (WGS84) [44] to associate geolocation data. It defines concepts such as iot-lite:coverage to determine the geographic area covered by the sensor.	Yes
Smart Appliances Reference Ontology (SAREF) [50]	Standard appliances used in a Home and Building environment	Handles concepts related to different domains (i.e. Construction Industry, Air Conditioning and Refrigeration, electrical systems, security, among others.) and extends SSN to represent the devices that belong to such domains	Yes

Several studies provide information on devices geolocation and, since SSN does not predefine the way to handle such concepts, a vocabulary must be aligned for dealing with IoT devices mobility and platforms deployment locations. Moreover, the way measures are represented is important, since, differently from geoinformation (many studies use align GEO Ontology (WGS84)), they can be represented in several forms (i.e., Observation and Measurement Ontology (OM) and QU Ontology).

Such studies help the identification of concepts to be presented in an SC scenario and covered in an SC ontology and clarify the definition of an SC ontology logically consistent with an application e.g., environmentally-related SC applications. AEMET extends SSN showing the scalability and interoperability provided by semantic technologies for the development of an SC.

However, those ontologies do not consider city-related concepts in a more comprehensive scope, which is expected, since they were defined specifically for a sensor context. Concepts such as origin of the data (provenance, i.e. Institution, Company, etc.) and way of taking advantage of the data in an SC scope (i.e. City Indicators) are not considered.

Other ontologies focused on SC do not consider common city definitions. Gupta et al. [51] proposed a platform for smart city data management based on semantic web and linked data. All available data on government (mainly Extensible Markup Language (XML) or Excel tabular data) were transformed into an RDF graph by Google Refine tool and an OWL ontology was defined over it and deployed in a Lena Apache Server that enables population of individuals and Protocol and RDF Query Language (SPARQL) queries through applications.According to the authors, a merging process of subontologies generated by Google Refine Tools must be conducted almost automatically, however, human help would align some mistakes of concepts. Since the authors did not follow a common base ontology, the merging (guided by humans) might return a new ontology that would not fit those previously generated by the same system.

Gaur et al. [52] followed a similar approach defining an SC architecture where the data gathered from the sensors is converted to an RDF graph enabling the definition of concepts using OWL ontologies too. Equally to the Gupta´s proposal [51], it is not clear about the ontologies used and what are the indicators used to describe the quality of life in the SC. Also, there is no a definition about how the data can be collected from sensors.

Abid et al. [53] defined a base ontology for SC focused on the reporting of faults in public services. They used Geonames Ontology for geographic data representation and included domain-specific concepts such as “ReportOfFault”, “Status” and the most important “person” concept defined by the Friend of a Friend Ontology [54].

Petrolo et al. [55] reviewed the VITAL platform based on linked data standards (e.g., RDF and JavaScript object notation (JSON)) to model and access data officially specified by OWL ontologies. The authors highlighted the importance of using ontologies, especially SSN , for representing the IoT environment inside SC. Although VITAL covers many topics defined in ISO 37120 (e.g., environmental indicators), to the best of our knowledge, it does not follow all the ISO 37120 definitions, which might affect the joint use of devices data between cities in a same country.

The proposal analyzed by DÁquin et al. [56] deals with the integration of different data providers in an SC. The HyperCat project [56, 57] describes a standard for exposing OWL-based Internet of Things data catalogs through conventional Web resources. The idea is to use distributed data repositories jointly through applications and query their catalogs in a uniform format. HyperCatś specification achieved those objectives by employing the same principles on which linked data and the Semantic Web are built, i.e., data accessible through standard formats and Web protocols (HTTPS, JSON, and others), identification of resources through URIs, and establishment of common, shared semantics for datasets descriptors.

From an SC perspective, although Hypercat can be useful for the dissemination of data providers, it does not cover city-related concepts, since it is a general purpose specification, and data providers must be aware to define the citie's data.

Lea et al. [58] studied the IoT interoperability problem in the SC context. The authors implemented a datahub for IoT integration and, rather than the previous approaches, they used Hypercat to disseminate each data provider. The sensors data (treated as real time data) are gathered using the WoTKIT platform [59] and the external static data (i.e. static files or web content) are integrated using the CKAN Dataset API [60].

Abreu et al. [61] proposed an ontology for the description of IoT infrastructure in the context of SC. It covers concepts such as network links, interfaces, and devices that compose it and considers metrics of its performance.

Particularly, those metrics and the whole ontology demonstrate the importance of defining indicators at city level. An extension of an SC ontology with such a vocabulary can lead to a standardization of, e..g.., the Telecommunications theme of an SC. However, the alignment of the ontology requires a base ontology that defines indicators over the metrics and another that aligns the devices inside the city in a more comprehensive way, as SSN .

Since no generic ontology for SC is available, we propose an ontology that considers both indicators and the data used in their estimation and acts as a general purpose ontology for SC covering the concepts analyzed to date.

The exclusive use of ENVO, GCIO or SSN ontologies is not enough to cover all the requirements of an ontology for SC Environment Indicators. Section 2.3 introduces our approach, which combines them into one ontology and will be the semantic base of our SC Environment Platform.

2.3 PROPOSAL OF A GENERAL PURPOSE SMART CITY ONTOLOGY

This section describes our General Purpose Smart City Ontology (GPSCO), which is based on SSN and GCIO. A matching technique that ensures logic consistency and enables the mapping and correlation in the new SC domain must be applied for the merging of both ontologies. The methodology considers i) a formal definition of both ontologies, ii) the application of a matching process for the obtaining of an alignment (mapping-rules), and iii) the merging of the ontologies

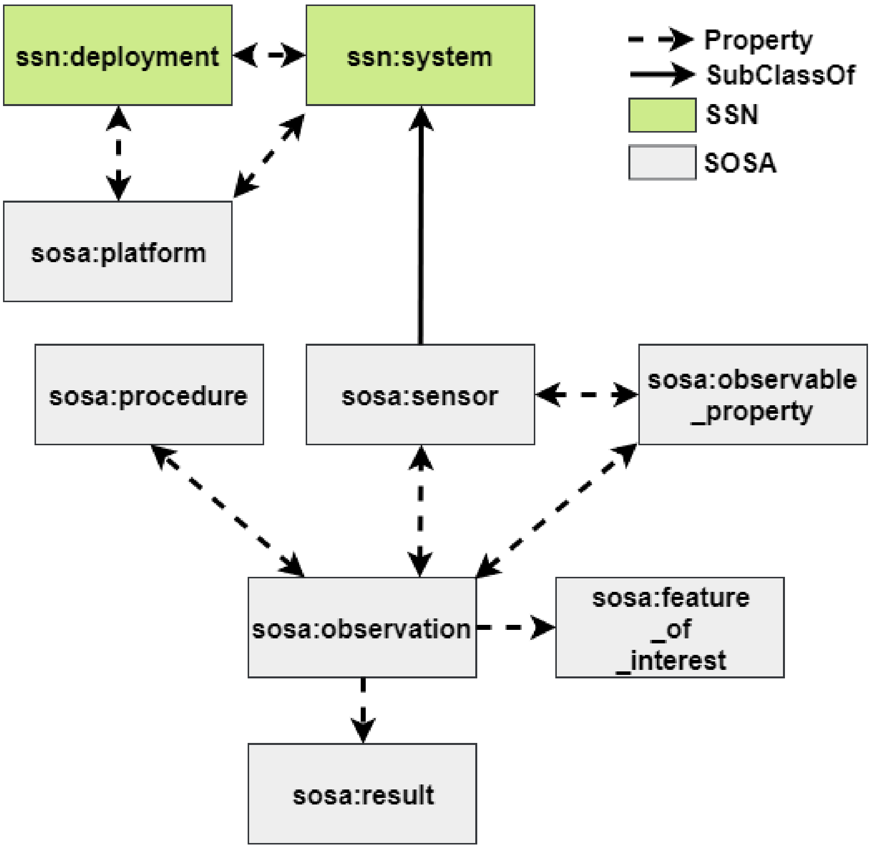


Figure 2.1: SSN Ontolgy (Observation Concepts) [3]

according to the alignment.

2.3.1 Formal Definition of SSN

As addressed elsewhere, SSN is a general purpose ontology that represents Sensor Networks and their data. It is defined according to a vertical architecture over a self-contained core ontology called SOSA (Sensor, Observation, Sample, and Actuator) (Table 2.2) for its elementary classes and properties.

Since this research is focused on the data interoperability problem for SC and cities must not intervene in the behaviour of citizens' devices, the SOSA concepts regarding Actuators/Actuation have been ignored. An SC will only collect the data and applications nourished by them can directly act on the devices.

Our proposal considers only the observation and sampling scenario defined by SSN . Towards simplifying the understanding of the concepts considered and the way of aligning SSN with GCIO, the SOSA definition was split according to the Observation and Sampling scopes (Figures 2.1 and 2.2, respectively).

Regarding observations (Figure 2.1), SSN has `sosa:observation` with associated object properties to represent the result of an observation (`sosa:result`), the thing observed

Table 2.2: SOSA+SSN definitions

Concepts (W_{SSN})	Description	Relations (R_{SSN})
sosa:ObservableProperty	An observable quality (property, characteristic) (i.e. height of a tree)	subclassof ssn:Property sosa:isObservedBy ONLY sosa:Sensor
sosa:Observation	An act of performing a (Observation) Procedure for the estimation or calculation of a property value.	sosa:madeBySensor EXACTLY 1 ONLY sosa:Sensor sosa:usedProcedure ONLY sosa:Procedure sosa:hasFeatureOfInterest EXACTLY 1 sosa:FeatureOfInterest sosa:observedProperty EXACTLY 1 sosa:ObservableProperty ssn:wasOriginatedBy EXACTLY 1 ssn:Stimulus sosa:phenomenonTime EXACTLY 1 sosa:hasResult MIN 1 sosa:Result
sosa:Sensor	Device, agent (including humans), or software (simulation) involved in Procedures that can respond to a Stimulus or prior Observations and generate a Result.	subclassof ssn:System sosa:observes ONLY sosa:ObservableProperty
sosa:Sample	Feature to be representative of a FeatureOfInterest on which Observations can be made	subclassof sosa:FeatureOfInterest subclassof sosa:Result sosa:isResultOf ONLY MIN 1 sosa:Sampling sosa:isSampleOf ONLY MIN 1 sosa:FeatureOfInterest
sosa:Sampling	An act of Sampling for the creation or transformation of one or more Samples	sosa:madeBySampler EXACTLY 1 ONLY sosa:Sampler sosa:usedProcedure ONLY sosa:Procedure sosa:hasFeatureOfInterest EXACTLY 1 ONLY sosa:FeatureOfInterest sosa:hasResult MIN 1 ONLY sosa:Sample sosa:resultTime EXACTLY 1
sosa:Sampler	A device used or that implements a (Sampling) Procedure to create or transform one or more samples.	subclassof ssn:System ssn:implements MIN 1 sosa:madeSampling ONLY sosa:Sampling
sosa:FeatureOfInterest	An object whose property is estimated or calculated in the course of an Observation towards a Result, or whose property is manipulated by an Actuator sampled or transformed into an act of Sampling	ssn:hasProperty MIN 1 ONLY ssn:Property sosa:hasSample ONLY sosa:Sample
sosa:Result	The Result of an Observation, Actuation, or act of Sampling. hasSimpleResult property to be used for the storage of an observation´s simple result.	sosa:isResultOf MIN 1
sosa:Procedure	A workflow, protocol, plan, algorithm, or computational method that specifies the way of making an Observation and creating a Sample	ssn:hasInput ONLY ssn:Input ssn:hasOutput ONLY ssn:Output ssn:implementedBy ONLY ssn:System
sosa:Platform	An entity that hosts other entities, particularly Sensors, Actuators, Samplers, and other Platforms.	sosa:hosts ONLY ssn:System ssn:inDeployment ONLY ssn:Deployment
ssn:System	A unit of abstraction for pieces of infrastructure that implement Procedures.It can have components, i.e., subsystems, which are other Systems.	sosa:isHostedBy ONLY sosa:Platform ssn:implements ONLY sosa:Procedure ssn:hasSubSystem ONLY ssn:System inverse Of ssn:hasSubSystem ONLY ssn:System ssn:hasDeployment ONLY ssn:Deployment

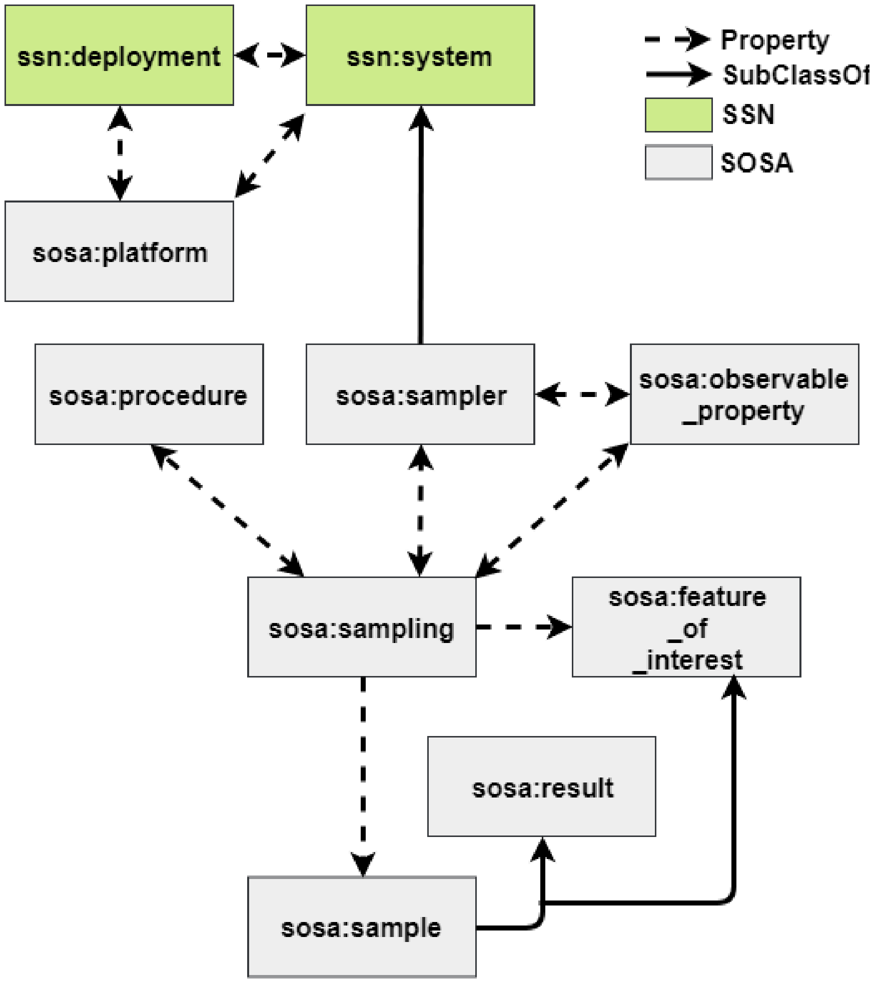


Figure 2.2: SSN Ontology (Sampling Concepts) [3]

(`sosa:feature_of_interest`), the system or sensor that gathered the data (`sosa:sensor`), (`sosa:system`), the observable parameter (`sosa:observable_property`), and the procedure followed (`sosa:procedure`).

Regarding sampling (Figure 2.2), SSN has `sosa:sampling` with associated object properties to represent the whole sampling process: (`sosa:sample`) represents the samples that belong to a sample sequence result of a sampling action, and (`sosa:sampler`) represents the device that gathered the samples. The remaining concepts and properties behave equally for the observation scenario.

The main difference between the contexts is one is devoted to a measure represented by only one value at a point in time, whereas the other represents a sequence of measures. Such a difference helps the representation of the moment at which an indicator value was provided as a single value precomputed by someone (observation) and data are available for supporting the computation of the indicator (Sampling).

We followed the formal definition of SSN described by Guarino et al. [30]:

$$C_{SSN} =< D_{SSN}, W_{SSN}, R_{SSN} >$$

where:

- D_{SSN} is the domain of sensors and their observations, the observations procedures, the features of interest studied, the samples used, and the properties and actuators observed;
- W_{SSN} are all the concepts defined in SSN, such as "sensor", "platform", etc. (Table 2.2); and
- R_{SSN} denotes the relations between the concepts (rules) in W_{SSN} that represent an inheritance or a relation property (Table 2.2).

Table 2.3: Ontologies used in the definition of GCIO

Ontology	Goal	Main Concepts Covered
Geonames	Identify the geographic area over which the indicator has been calculated	City, Country, State, GeoCoordinates, GeoShapes, Neighborhood, Building, among others.
OM	Represent the measurement theory	Quantity, Unit of Measure, Ratio, Measure
PROV	Define the provenance of an indicator	Entity that aims to specify its provenance, Activity (Procedure that creates or transforms the entity), Agent (The one who changes the entity)
OWL-Time	Define the time at which measurements are taken, computed or derived	DateTimeDescription, DateTimeInterval, DayOfWeek|Month|Year, Duration, Instant, Interval, MonthOfYear, etc.
Knowledge Provenance	Represent the validity (certainty) of a proposition	Validity (value between 0-10) associated with a period of time
The Trust	Represent the degree of trust in the provider of an indicator value	TrustValue ("low", "medium")

Below is the formal definition of GCIO for the application of a matching strategy.

2.3.2 Formal Definition of GCIO

As addressed elsewhere, GCIO extends the ISO 37120 definitions through the merging of the OWL ontologies summarized in Table 2.3 and defined as a conceptualization:

$$C_{COMP} =< D_{COMP}, W_{COMP}, R_{COMP} >$$

where:

- D_{COMP} is the multidomain that considers the concepts related to georeference, measurement theory, provenance, time definitions, validity, and trust;
- W_{COMP} are all the concepts defined by those ontologies; and
- R_{COMP} denotes the relations between the concepts (rules) in W_{COMP} that represent an inheritance or a relation property.

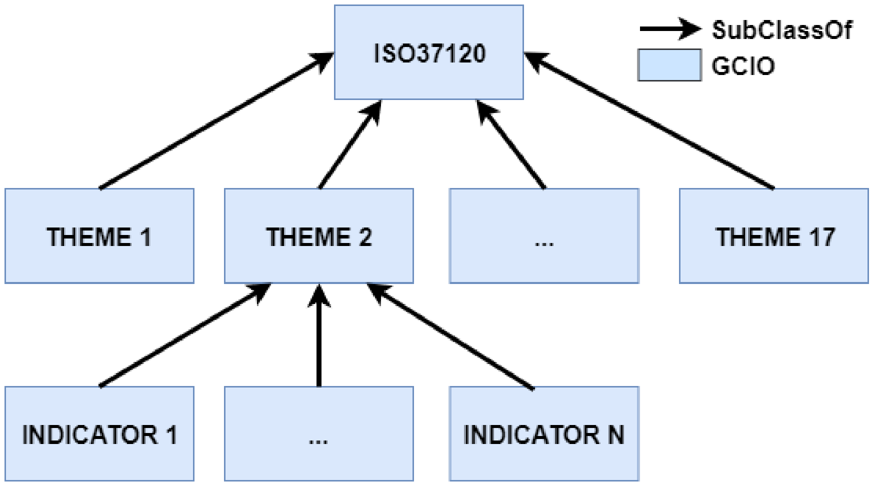


Figure 2.3: GCIO Base [3]

C_{COMP} is logically consistent.(Any pair of subdomains merged in C_{COMP} is disjoint regarding their semantic definition.)

Fox [35] defined the classes and attributes to represent the ISO 37120 indicators concepts in GCIO (Table 2.3). Therefore, GCIO can be defined as

$$C_{GCIO} =< D_{GCIO}, W_{GCIO}, R_{GCIO} >$$

where:

- D_{GCIO} is the domain composed of a D_{COMP} union of the domain of City Indicators Defined in ISO 37120;
- W_{GCIO} represents the set of concepts defined by the W_{COMP} union of the ISO 37120 ones; and
- R_{GCIO} denotes the set of relations between the concepts (rules) in the W_{COMP} union of the rules and associations defined by Fox [35].

Figure 2.3 shows GCIO defines groups of Indicators to represent the concept of theme handled in ISO 37120. Each indicator is defined as an inherited object of the theme that defined it (all indicators inherit one of the 17 themes defined in ISO). They also inherit other definitions and contain properties for the representation of their values. Figure 2.4 shows a summary of the way an indicator is defined in relation to the other ontologies that extend GCIO.

2.3.3 GPSCO definition

The integration of SSN and GCIO requires an alignment $A < R_a >$ between (C_{SSN}, C_{GCIO}), where the set of correspondence rules R_a is logically consistent and represents the relations between the concepts in both domains D_{SSN} and D_{GCIO}. A semantic analysis was manually performed and the definitions followed the SSN split introduced in the previous section (observation/sampling contexts).

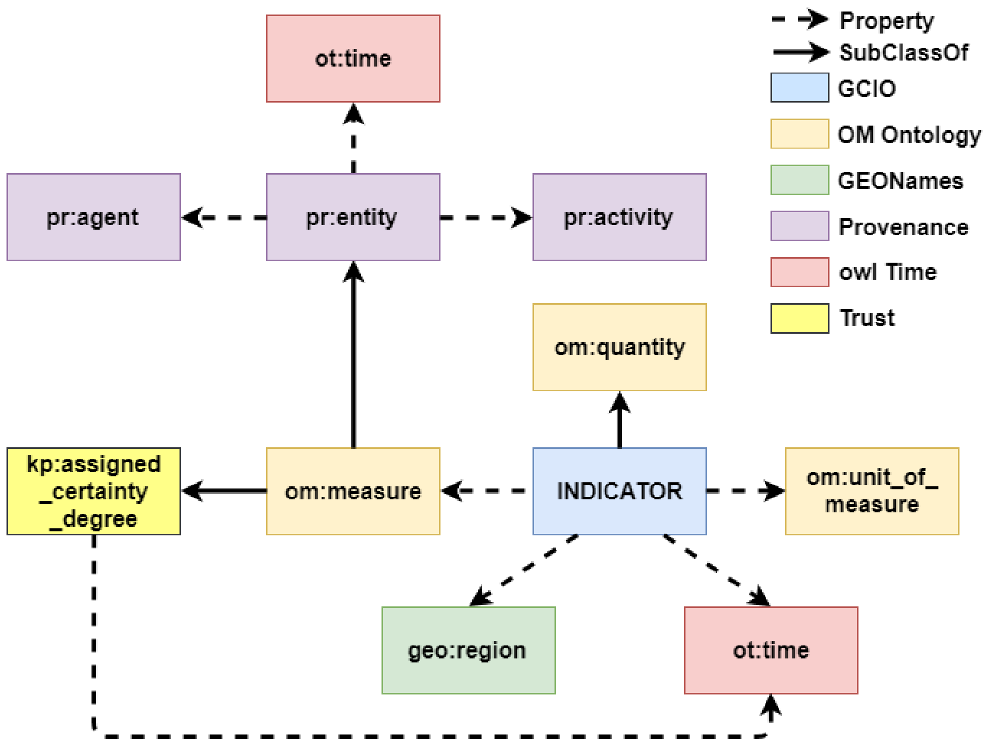


Figure 2.4: GCIO Indicators [3]

Figure 2.5 displays some of the rules added to the SSN observation context and the GCIO merging. The alignment enables the representation of supporting data on the entities involved in the synthesizing of an indicator value. For example, an indicator value (om:measure) could be a sosa:result obtained by an observation generated by a sensor or a system inside a platform. The observation is made over a sosa:feature_of_interest (it extends the om:quantity that defines the indicator) and the system/sensor that generates the value contains information on its provenance. This alignment enables the representation of scenarios whose indicators value provider is a person, a computer system, a governmental organization, an application or an agent, since sosa:sensor could be anything after the extension for a particular scenario.

Moreover, SC-related data that have established a low/no relationship with city indicators can be represented in this approach.

The use of a sosa:result for the description of an indicator value does not enable the representation of the historical data that generate it. Figure 2.6 shows how a sosa:sampler that has a sosa:feature_of_interest is associated with an indicator and correlates it with the samples that enable the estimation of the indicator.

Both sosa:feature_of_interest and om:quantity are definitions of something measured and sosa:feature_of_interest class was extended by om:quantity. This association implicitly establishes sosa:observation can be a gcio:indicator.

According to this approach, the ontology covers cases of pre-computed indicators values or in-

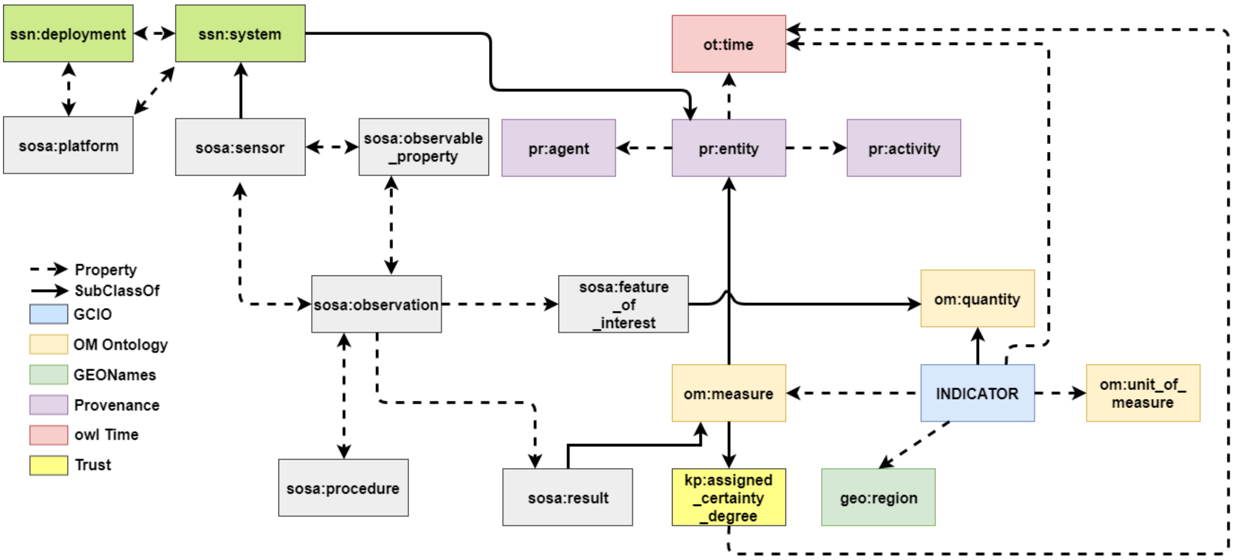


Figure 2.5: GPSCO For Precomputed Indicators Data [3]

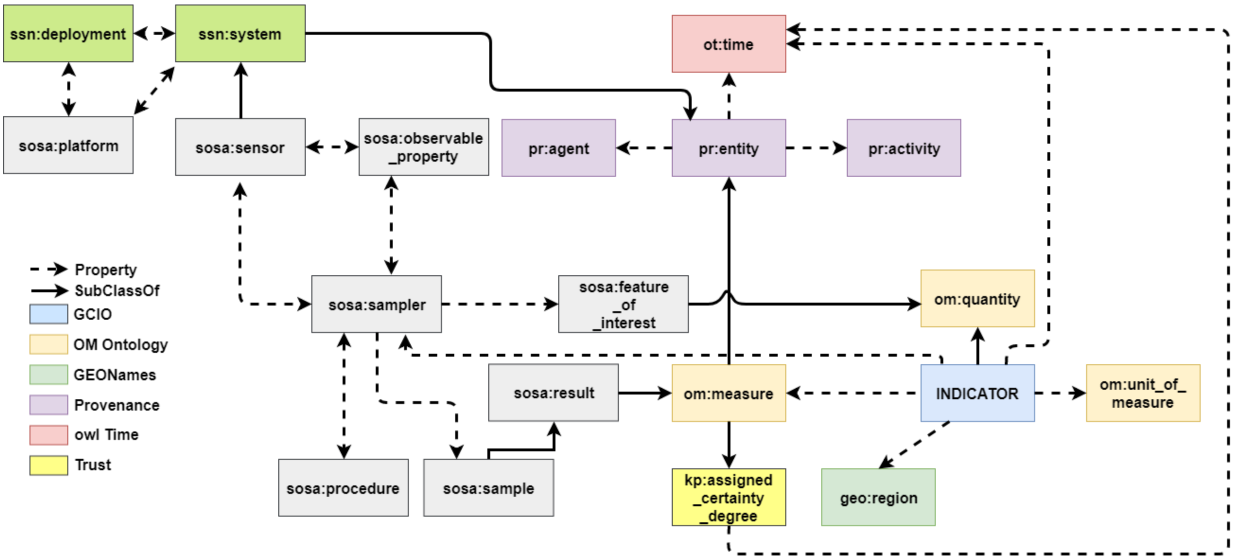


Figure 2.6: GPSCO For Supporting Indicators Data [3]

dicators resulting from a sampling procedure represented by ssn:procedure, which describes the generation of the result (i.e., an average over the sample).

Similarly to GCIO, GPSCO considers the origin of the data and their provider through PROVENANCE ontology, in which all classes that require provenance-related information are conceptualized as a subclass of prov:entity. The prov:activity class describes the set or sequences of actions that transform the entity. The agent class represents the person, organization, or system that performs the activity.

Our proposal aims at representing the indicators provenances and supporting data provenance following the alignment provided by Compton et al. [62], in which ssn:system derives from the prov:activity class and ssn:sensor and ssn:sampler derive from prov:agent.

Regarding Time representations, both scenarios require the date of origin of the indicator value. Similarly to GCIO OWL-Time ontology [46] will help the definition of intervals, temporal positions, temporal units, etc. Property ssn:phenomenomTime, as a property of an ssn:observation or an ssn:sampling, represents the time at which an indicator value was measured.

Any ssn:result or ssn:sample related to the supporting data time has an ssn:resultTime property that represents the time at which the indicator result was computed and the supporting data samples were collected.

In an SC environment, all indicators are associated with at least one geographic area (city area), which must be considered for the description of the region of collection of supporting data. The more granulated the geographic information, the better the development of the predictor or actuator. For example, low water pollution-related indicators of an entire city do not help the finding of the origin of the pollutant. On the other hand, supporting data labeled with more locally geographic information, as a geometry in a river, would provide such information.

In GPSCO, the geographic data are represented by GeoSPARQL Ontology [63]. In GeoSPARQL, a geo:feature denotes a geographic area or point of buildings, rivers, cities, countries, streets, and their geometries. It is set as an attribute over an ssn:observation and the ssn:sampling that represents the city area and denotes the specific location of a sensor or sampler (i.e., river, building, latitude, longitude) for the supporting data (ssn:sample). GeoSPARQL also enables queries over the supporting data (i.e. all ssn:samples from a specific geo:feature belonging to a period of time in which geo:feature can be an entire city, or a more specific location or geometry).

Validity in GCIO refers to the usability of an indicator over time. At the point of publication, the indicator is assumed valid; however, after a period, it (or its supporting data) may not be valid any longer and must be discarded in queries. In GPSCO, validity is represented as in GCIO, since it is a concept related to an indicator, and not to sensors.

On the other hand, trust measures the reliability of data and their creator. Less trusted data

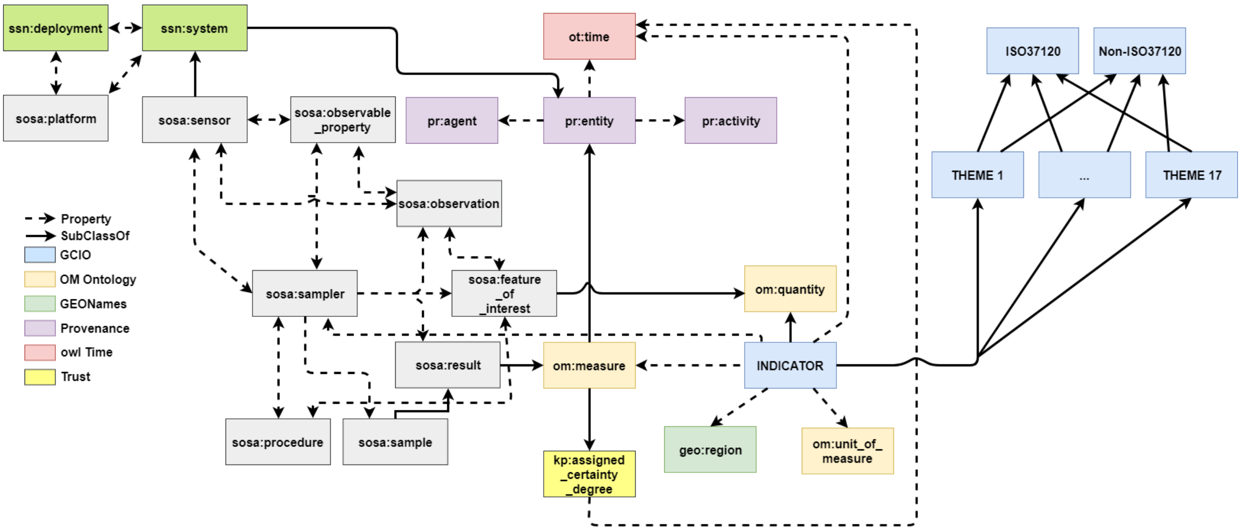


Figure 2.7: GPSCO Full Ontology [3]

producers will negatively affect the validity of an indicator or a dataset. In GPSCO, the behaviour of the trust concept is the same as that of GCIO.

Finally, GPSCO can be defined as:

$$C_{GPSCO} =< D_{GPSCO}, W_{GPSCO}, R_{GPSCO} >$$

where:

- D_{GPSCO} is the domain composed of D_{SSN} union D_{GCIO} ;
- W_{GPSCO} represents the set of concepts defined by G_{SSN} union G_{GCIO}; and
- R_{GPSCO} denotes the set of relations between the concepts (rules) in W_{SSN} union W_{GCIO} plus the rules defined above.

Figure 2.7 shows all the concepts considered their alignment. Since GPSCO is a generic SC ontology, concepts such as person can be represented extending the provenance ontology with those as Friend of a Friend (48) for citizens representation. Proposals, as the one analyzed in Section 3, which extends SSN , are also compatible. Therefore, proposals of more specific domains can take advantage of GPSCO to define indicators values based on ISO 37120 Indicators. Indeed, only 100 indicators were defined in ISO 37120 and some of the specific domains were not covered. In such cases, a new concept is added for representing non-ISO indicators, as shown in Figure 2.7.

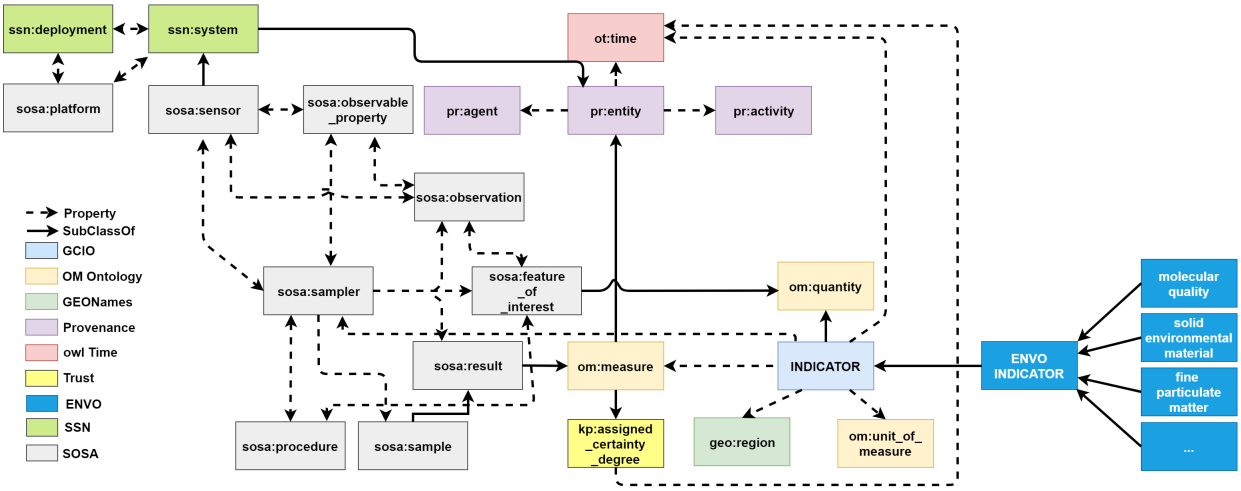


Figure 2.8: SSN , GCIO and ENVO Alignment [3]

2.4 DEFINITION OF ENVIRONMENT INDICATORS FOR SMART CITY ONTOLOGY

Below is the definition of an ontology for environment in SC called EISCO. The ontology is based on ENVO, GCIO, and SSN ontologies and covers the main definitions required for semantic data representation in the smart city environment domain. Since ENVO does not consider concepts related to indicators, we propose following the GCIO approach designed by Fox [35].

GCIO defines groups of indicators to represent the concept of themes handled in ISO 37120. Each indicator is defined as an inherited object of the theme that defined it (all indicators inherit one of the 17 themes defined in ISO). However, ISO 37120 and GCIO cover other themes rather than the environment and, specifically for the environment, they only cover eight indicators. In EISCO, the theme definitions are ignored and an indicator is represented by reusing the GCIO definition (Figure 2.8). An indicator inherits other definitions and contains properties for the representation values related to time, provenance, geolocation, and measurements.

However, ISO 37120 is poor in terms of definitions (only eight indicators) and ENVO ontology will help solve this issue with a broad set of environmental definitions. The “ENVO Indicator” class (Figure 2.8) inherits from “GCIO Indicator” and groups all concepts that can be quantified is here defined. The quantity that represents an ENVO indicator is associated with a unit of measure inherited from GCIO Indicator and can represent a scale (i.e. water pH), a concentration (i.e. PM2.5), or a well-defined unit (i.e. solid waste weight measured in tons).

Therefore, ENVO does not include the unit of measure related to the environment concept definition. In this sense, a review of all ENVO concepts that can be quantified was made and a mapping of the ENVO concept to unit of measure was proposed. The resulting mapping enabled a proper definition of indicators based on ENVO definitions.

2.5 ONTOLOGY TO SUPPORT THE SEMANTIC REPRESENTATION OF IOFT DATA

Figure 2.9 shows our ontology definition for IoFT. The ontology was proposed from an extension of GPSCO ontology for the representation of concepts related to Internet of Forestry Things observations.

An ioft_observation is here conceptualized as an extension of a sosa:observation, which also extends an om:quantity for the representation of a unit of measure. An ioft_observation has a time (ot:time) to represent the moment at which a sensor gathered data, and a region (geo:region) to denote the location of the observation.

An ioft_observation has a sosa:result that represents the numerical value of the observation and correlates it with the sampler strategy that made the observation (sosa:sampler). The sampler enables the identification of the IoFT sensor that captured raw data and associated them with the property observed, i.e., observable property, which is here considered any measurable IoFT -related parameter (e.g., temperature, humidity, CO2, among others).

The proposed ontology promotes the retrieval of IoFT aggregated data represented in RDF format. For example, all ioft observations of an observable property (i.e temperature) made in a region (i.e. Federal District) can be queried between two times (ot:time). Other virtual devices can be implemented in the cloud as micro-services and directly communicate with others in cloud platform components.

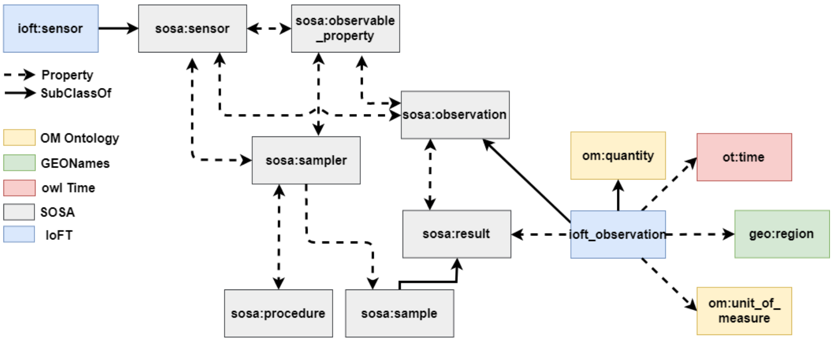


Figure 2.9: IoFT Ontology [4].

2.6 CONCLUSIONS

In this chapter, after a literature review of ontologies available for the treatment of sensors, IoT and environment and Smart City (SC) data, we proposed GPSCO ontology for the semantic integration of SC data, which was extended for the proposal of the EISCO, resulting in a broad set of environmental indicators making possible the definition of data semantics for a Linked Data Storage and enabling the application of ontology reasoners to extract knowledge from stored data. Besides indicators data, the ontology covered other observation-related concepts, such as sensors, platforms, data provenance, and trust. Moreover, the definition of IoFT was included to enable the representation of forestry data by a semantic approach.

3 SMART CITY PLATFORM

This chapter presents and discusses, initially, some relevant studies on Smart City (SC) platforms, with focus in the areas of Network and Application Layer Standards and Data Interchange Formats, as well as in layer-based architectures. In the sequence, functional and non-functional requirements for a smart city platform are elicited, and a SC platform is proposed. Such platform is then extended to forestry and wildfire predictions, and experiments are accomplished in order to evaluate the platform.

3.1 RELATED WORK

Regarding networking standards, IoT and SC are heterogeneous environments of devices and sensors with different network interfaces. Li et al. [64] reviewed those networking standards and concluded IEEE 802.11 (WLAN), IEEE 802.15.1(Bluetooth, Low-energy Bluetooth), IEEE 802.15.6 (wireless body area networks), and 3G/4G were the most used.

On the other hand, 5th generation mobile network (5G) is expected to improve latency, data rates, bandwidth, and energy consumption [65]. The improvements of the quality metrics over those parameters are very important since IoT devices are, in many cases, constrained devices. Moreover, a better utilization of the devices capacities are always welcome.

Particularly, 5G addresses three scenarios closely related to the SC context [66]. Enhanced Mobile Broadband (eMBB) covers the exchange of data between various user equipment including text, and multimedia, characterized by large bandwidth requirements [67]. Massive Machine Type Communications (mMTC) covers a large number of connected devices (e.g., sensors and wearable devices) through a dense deployment in a city [68]. These devices are used to provide different services (i.e., automatic monitoring of environment parameters). Ultra-Reliable and Low-Latency Communications (URLLC) covers those communications that are time-critical and-or require high delivery probability [69].

In terms of application layer standards, Rashed et al. [70] proposed the communication of IoT devices with the gateway through the Message Queue Telemetry Transport (MQTT) protocol [71] and Datta et al. [72] defined a cloud-based gateway using Representational State Transfer (REST) web services similar to the middleware proposed by Paganelli et al. [73]. Jan et al.[74] considered the Constrained Application Protocol (CoAP) [75] to communicate IoT devices with fog servers and Petrolo et al. [55] used CoAP for communication between IoT devices and cloud resources. Desai et al. [76] introduced a gateway that communicates with IoT devices through the MQTT protocol, besides the Extensible Message Presence Protocol [77] (XMPP) and the Advanced Message Queuing Protocol (AMQP) [78].

Dizdarevic et al. [79] compared the aforementioned protocols and concluded MQTT and REST-ful HTTP are the most suitable for IoT, since they are the most mature and stable ones. On the other hand, CoAP should also be taken into consideration; it also rapidly evolves as an IoT messaging standard and is likely to reach a level of stability and maturity similar to that of MQTT and HTTP in a near future.

The format of data interchanged through the communication protocols must be standardized. Since IoT devices are resource-constrained, the extension of data formats (e.g. RDF) to this domain might be impossible for several applications. SenML [80] represents an important alternative for solving this issue, since it defines media types to represent simple sensor measurements and device parameters. The representations are provided in Java Script Object Notation (JSON) and XML, which shares the common SenML data model. Datta et al. [72] implemented it for data representation, where the gateway receives the sensed data in a sensor custom format and transcodes them following the SenML specification. The SenML-encoded data are sent to the upper layers; however, other aspects that affect interoperability are disregarded. Different units of measurement, sampling rates or numerical systems, in scenarios that involve more than one platform, can lead to a wrong interpretation of data.

Other related studies addressed complementary relevant issues. Zhao et al. [81] developed an incomplete multi-view clustering methodology that projects multi-view data with missing features for a complete and unified representation in a common semantic subspace. The authors used an affinity graph and a deep neural network to construct a multi-layer non-linear correlated set of complete views and proposed an objective function that updates the model from one dataset to another. Zhao et al. [82] proposed a transfer learning method for multimedia co-occurrence data based on deep semantic mapping. It integrates deep neural networks with canonical correlation analysis towards modeling a semantic subspace for associating data across source and target domains. Liu et al. [83] reviewed the application of deep learning for urban big data fusion and highlighted the accuracy and importance of such methods. However, the present study demonstrates the adoption of those methods in a general-purpose platform, as the one proposed, is almost impossible. Many of the proposals that consider a semantic integration that differs from the use of ontologies (at data level) are restricted to specific use cases to which semantic integration models and algorithms are adjusted.

As shown in Figure 3.1, IoT devices sense the city environment, collect data, and forward them to a fog or a cloud system through gateways. The fog system provides a local environment near IoT devices for data storage and deployment of applications on a lower scale for the processing of data on buildings or neighborhoods scale. The gateway functionalities enable the forwarding of the data to the cloud environment, which provides service on a city scale for their storage and sharing for applications.

In terms of interoperability-based treatment of data, the entire data life-cycle initially considers a new environment ontology for promoting semantic compatibility. Moreover, standardized communication protocols and data formats are considered jointly with a standardized query mech-

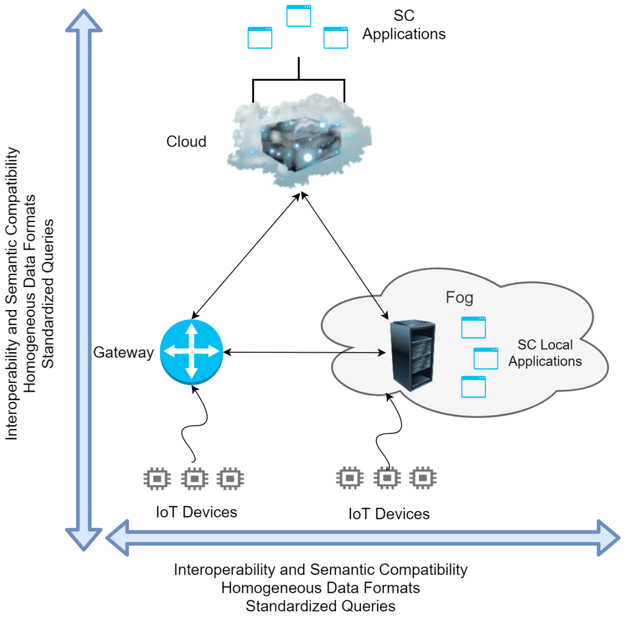


Figure 3.1: Platform overview

anism. According to a common environmental terminology, data are linked to their semantic definition at any stage, thus enabling the applications to extract the relation between measurements of related concepts and the automatic generation of indicators values.

The proposal of an interoperable platform demands the study of architectural and platform features. Irfan and Ahmad [84] considered a three-layered structure ideal for a logical fragmentation and abstraction of complexities of IoT architectures composed of Things and Intermediate and Integrated Application layers. The Things layer is comprised of heterogeneous IoT objects that communicate through various communication protocols and networks. The Intermediate layer is represented by a middleware or gateway that handles the IoT devices, processes the data at a local stage and is implemented by Multi-agent, Service Oriented, RESTful or Publish-Subscribe technologies. The Integrated Application layer stores huge amounts of data and processes them through several applications.

The proposals of [72, 85, 86, 87, 88, 89] involved architectures that follow the aforementioned three-layered approach. At the lowest layer, all studies considered IoT devices that deploy sensors and/or devices that control a sensor network. Such devices connect to the platform through different access networks (i.e. WiFi, Bluetooth, Long Term Evolution (LTE)), sense the data and forward them to the platform that finally makes them available for applications with diverse objectives, including treatment of big data.

Santos et al. [90] followed the same three layered architecture but considered 5G as the access network. The authors also proposed a device to device communication between sensors and gateways. On the other hand [91] considered 5G-enabled sensors that connects directly to cloud servers and 5G-enabled fog servers that act as gateways with local processing capabilities.

Rahmani et al. [92] proposed an IoT-based health monitoring system composed by the layers:

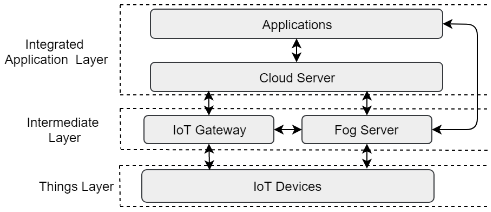


Figure 3.2: Structure of a three-layer IoT architecture [3]

Smart Devices, which include the Sensor Networks and IoT Devices; Edge/Fog, that contains the Gateways, which its primary function is to forward the health data to the Cloud and to provide services for the discovery and control of the IoT devices; and Cloud, where the data is processed, stored and consumed by applications. In this study, the gateway plays the key role enabling Fog Services for local data preprocessing and storage improving mobility and accessibility issues. However, the communications protocols, the data format used and the integration with external applications is not clear.

Figure 3.2 shows the three-layer architecture, which considers both fog-based and gateway-based approaches. As in Howell et al. [93], our proposal takes advantage of a fog server for local processing, while governmental services are processed globally in the cloud. Moreover, home gateways collect data in an aggregated and local way.

Regarding networking technologies, the platform considered the IEEE 802.11 (WLAN), IEEE 802.15.1 (Bluetooth, Low-energy Bluetooth), IEEE 802.15.6 (wireless body area networks), and 3G/4G/5G standards to grant communication among the three layers, and has developed two different implementations of the IoT Gateway. The first is based on an Android smartphone that considers Bluetooth and WiFi network connectivity, whereas the second is a micro-controller-based gateway (Arduino, Raspberry Pi) with Bluetooth, WiFi, and ZigBee. Both connect to the cloud through the Internet accessed by mobile data networks. The Fog Server is based on a common PC/Server with Bluetooth and WiFi interfaces.

Similarly to the gateway-based approach proposed by Desai et al. [76], each component (IoT Gateway, Fog Server, Cloud Server) in Figure 3.3 provides a set of adapters that implements CoAP, MQTT , and REST over HTTP. An abstract adapter enables instances defined by other application protocols, thus extending the gateway functionalities. The Settings Manager entity stores the settings and handles all adapter instances and the addressing among gateway, fog, and cloud adapters. Each adapter instance can act as a bridge with its counterpart in another entity. For example, a gateway CoAP adapter can make a bridge with a fog CoAP adapter, which enables data transfers between IoT devices and applications regardless of the entity that handles them.

All data among IoT devices, gateway, and fog server are represented in SenML format as defined below:

{ “n”: “IRI of the sensor on the Environment SC Ontology” + “/Uid”,

“t”: “time at which observation was received”, “u”: “IRI of the unit_of_measure on the Environment SC Ontology”,

“v”: “Numeric value obtained in the observation”,

“vs”: “String encoded value, if it exists”}

Each IoT sensor is identified by a unique International Resource Identifier (IRI) that aggregates semantic information and is associated with ontology classes that define the type of sensor, the type of data observed, among other semantic characteristics represented in an ontology (next section). The SenML messages are transmitted through the adapters as the payload of the selected application protocol. Finally, the applications in the Integrated Application layer retrieve the data represented in both SenML and RDF formats.

On the other hand, IoT applied to forestry industry enables the collection of large quantities of data for supporting several decision-making processes. For example, the platform proposed by [17] shows the importance of IoFT data such as temperature, humidity, and carbon dioxide in the early detection of fires. However, it ignores other complementary data, thus hampering the development of other types of studies (e.g., behaviour prediction). Although the authors addressed the way ML can be considered, they clarified neither how to take advantage of it, nor how to handle the heterogeneity of IoFT platforms and definitions such as observation measures, process, and communication protocols involved.

Tsiropoulou et al.[94] discussed an energy and physical aware-based framework for coalition formation and resource distribution among wireless IoT applications. Numerical results validated the energy-efficient characteristic of the proposal; however, the framework considers a static data representation scheme difficult to extend. The research was devoted to the capture system rather than to an aggregation and interchange platform for IoFT data.

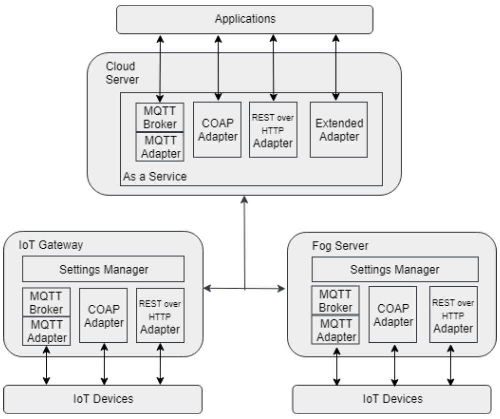


Figure 3.3: Communication Protocols [3]

3.2 REQUIREMENTS ELICITATION

Regardless of the use cases, systems for SC and IoFT share common functional and non-functional requirements. The former enable the definition of services or functions provided by the platform, whereas the latter are more focused on the quality of services, performance problems, and issues related to the implementation of the platform.

3.2.1 Functional and Non-Functional Requirements

In general terms, the platform must enable the collection and exchange of SC/IoT/IoFT devices data in a simple way and reduce integration efforts between data producers (sensors) and data consumers (applications). The following functional requirements were defined for the meeting of such objectives:

1. Sensors integration: The platform must promote the identification and integration of several sensors and registration functionalities such as sampling process, unit of measurement of the observations, and type of feature observed must be provided for the identification of sensor- collected data and definition of sensor metadata.
2. Devices Heterogeneity: The platform must deal with the heterogeneity promoted by sensing devices. Tools should be proposed towards simplifying the collection of data from SC/IoFT solutions and sensors and other complementary data producers should be considered for the sake of integration by the platform.
3. Collection of heterogeneous data: : Data are the core of the proposed platform and refer mainly to sensed observations. The platform must provide data management services from sensor to applications, including data acquisition, data processing, and storage in a standardized format.
4. Context information: Context is very important in SC/IoFT and its applications. A large number of sensors generates large amounts of data, which have no value unless they are jointly analyzed, interpreted, and understood. Since temporal and spatial context plays a vital role, the platform must provide mechanisms for a context information representation.
5. Resource Limitation: Sensors are commonly limited in terms of processing, memory, and communication capacity. Therefore, the platform should consider standards and tools aligned to such constraints.
6. Data-related services: The platform should provide tools for the definition of pre-processing strategies, which may include data filtering and aggregation strategies, and services for query and streaming data in a standardized manner must be made available for consumers´ applications. It must also enable and manage the semantic relationship between data, and the data schema must allow the extension and edition of semantic definitions.

7. Event Management: A large number of observation events is generated by broad SC and IoFT systems and must be managed as an integral part of the platform, which transforms observed events into meaningful ones and enables real-time analyses so that downstream applications are driven by accurate and real-time information and intelligence.
8. Inference Services for Decision Support: The platform must provide machine learning services that facilitate data analysis and extraction of insights by the final applications.

Non-Functional Requirements:

The following key non-functional requirements were considered for the IoFT platform:

1. Interconnectivity: The platform should support as many modes of connection and communication protocols as possible for the forwarding/production of observed data.
2. Extensibility: The platform must be extensible towards the integration of new devices and applications with no alterations and enable the semantic models to be updated so that the data schema can evolve dynamically in function of new application needs.
3. Real Time Treatment: Consuming applications (i.e. fire prediction) are highly sensitive to latency and should not experience data delays; therefore, the time between performing an observation and receiving data in the application should be minimized.
4. Scalability: The expansion of both devices and data collected leads to great concerns. The platform must guarantee a sufficient quality of service for supporting the expansion capacity of the network when more objects are added or when the volume of observed data increases.
5. Interoperability: IoFT must be usable by the applications and devices with slight changes performed by developers. Interoperability is improved when the platform provides APIs for developers and supports many protocols, such as MQTT (Message Queue Telemetry Transport), HTTP (Hyper Text Transfer Protocol), AMQP (Advanced Message Queuing Protocol), and CoAP (Constrained Application Protocol), widely used in sensor networks.

3.3 PLATFORM DESCRIPTION

Figure 3.4 shows the main platform components, namely applications, Cloud Server, IoT Gateway/Fog Server, and IoT devices. In the upper part are the applications that process the gathered environment data, apply data mining and other big data techniques, train complex machine learning models, or monitor the data towards supporting decisions on city strategies. The Cloud Server provides the platform with features required for data storage and processing for supporting the data used by applications. The data stored on the cloud are provided by the IoT Gateway and Fog Server, which primarily function as a data relay between IoT Devices and Cloud Server. Finally, IoT devices measure the environmental parameters and forward the measurements to either the gateway, or the fog server.

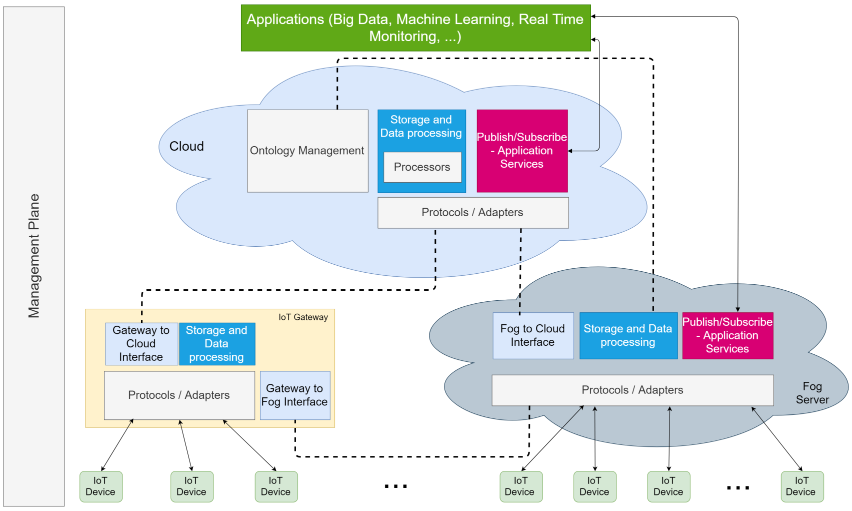

Management Plane
Applications (Big Data, Machine Learning, Real Time Monitoring, ...)
Cloud
Ontology Management
Storage and Data processing
Processors
Publish/Subscribe - Application Services
Protocols / Adapters
IoT Gateway
Gateway to Cloud Interface
Storage and Data processing
Protocols / Adapters
Gateway to Fog Interface
Fog to Cloud Interface
Storage and Data processing
Publish/Subscribe - Application Services
Protocols / Adapters
Fog Server
IoT Device
IoT Device
IoT Device
...
IoT Device
IoT Device
IoT Device
...
IoT Device

Figure 3.4: Proposed Smart Environment Platform [3]

All components are treated in a Management Plane that covers the main management areas, such as Configuration, Security, Performance, Faults, and Accounting. Regarding implementation, the platform uses several technologies and programming languages and follows the considerations detailed below.

3.3.1 Applications

Applications are considered clients that use the platform and can implement data mining techniques or near real time monitoring by either querying the data from the cloud and fog servers, or receiving them in a publish/subscribe approach. Machine learning algorithms (e.g., deep learning, clustering, neural networks, among others) can assess an indicator of interest. For example, an application can use dynamically collected indicator data to train a recurrent machine learning model (i.e. a recurrent neural network) and apply data mining techniques in pre-stored indicator data sets.

In the first case, the application subscribes to one or more topics in the MQTT broker of interest (in a fog server for localized processing, or in a cloud server for city scope) and whenever data of indicators are published in those topics, the machine learning model is trained. A user (or another system) can then make inferences using the trained regression model. In the second case, users can design big data algorithms (i.e. data mining techniques) that query data from cloud/fog servers using SPARQL sentences over RESTful web services, extract meaningful insights about the indicator value, and trigger actions over actuators (i.e. high temperature detected triggers anti-fire systems).

Supply of standardized services for the aggregation and distribution of data and management and automatic generation of indicators are the main objectives of the platform. The implementation or proposal of big data, data mining, or machine learning algorithms is not an objective of this study, although many applications can benefit from the platform. Since the data are standardized and public, applications with different objectives can be developed (e.g., continuous monitoring of indicators (e.g. PM2.5 and PM10) in a region (neighborhood or city) which is useful for citizens with, for example, respiratory diseases)). Different service providers such as health, transportation, and environmental care can benefit from our platform for establishing relationships among various environmental parameters. Regression models can be trained to estimate environmental conditions in a region and clustering algorithms can identify sources of contaminant generation. Another use case considers the early detection of fires, in which a continuous distribution of temperature data enables the implementation of alert systems. Therefore, since the platform promotes the distribution of environmental data following a Publish/Subscribe approach and acts as a Linked Data Source, numerous applications can be developed.

3.3.2 Cloud Server

The cloud server integrates different sub-components that can be instantiated according to a Platform and Infrastructure as a Service model and deployed on the Microsoft Azure cloud services provider. Its main functionalities are related to the maintenance of the semantic model and the storage and forwarding of data among gateways, fog servers, and applications. The following subsections describe the components that comprise the Cloud Server and their features.

The Ontology Management component provides a visual interface and functionalities for creation, update, deletion, and deployment of definitions on EISCO ontology. Its main functionalities include definition of sensors, platforms, geolocation areas, among other instances of the classes defined on EISCO. It was deployed as a Web application over an Azure App Service and use of Azure Blob Storage for the deployment of EISCO and generation of valid IRIs. Moreover, it enables the edition and publication of new ontology definitions provided by the Storage and Data processing component, which also supplies all functionalities for the storage and handling of indicator-related sensed data. When instantiated, the component starts a Virtuoso Storage for storing the RDF data using the linked data functionalities and according to EISCO provided by the Ontology Management component. It provides functionalities for the storage of the SenML data supplied by an adapter manager as RDF triples. The proposal of Su et al. Su et al. [95] supported the mapping of SenML in RDF . Once the messages have been translated, the component provides functionalities to store the data in Virtuoso storage and to forward messages to the Publish/Subscribe and Application Services.

Jena Apache server is started by the aforementioned component for the definition of a preprocessing pipeline and the development of reasoners for the inference of new RDF rules. The storage and data processing component provides functionalities to forward the new inferred data. Some filters applied prevent the repetition of data and check their both integrity and trust. Platform users define their own processing techniques deploying a data processor sequence, since the processors can accomplish complex tasks for data preparation, such as estimation of missed features of an observation, or for the calculation of an indicator value in a custom time period. A processor is provided by default for the management of indicators values originally defined in

EISCO.

The Publish/Subscribe and Application Service exposes a set of RESTful endpoints for the query of data using SPARQL query language and returns data in SenML and RDF format. It also enables the applications to subscribe to an MQTT broker where the data received by the Storage and Data processing component (forwarded by the adapters or inferred by Jena Apache reasoners) are published.

The Protocols/Adapters component handles the adapters that implement the application layer protocols discussed in Section 3.2. Configurations such as port numbers, topics for brokers, among other parameters necessary for the start of the adapters are provided by the management plane. The Protocols/Adapters component is composed of three main entities, namely Multi-adapter manager, MQTT adapter, and Non-MQTT adapters. The multi-adapter manager controls the data flow among the adapters and provides functionalities to start/stop them and configure data tunnels between them. If configured, it also enables a direct publication of data to the Publish/Subscribe Application Services. The MQTT adapters are instantiated to receive the data forwarded by different IoT gateways and Fog servers deployed in the architecture. Following a Publish/Subscribe approach, their main functionalities include the creation of topics on the MQTT broker for the forwarding of data from both gateways and fog servers. When a topic is created, the adapter notifies the adapter manager to subscribe to it. On the other hand, Non-MQTT adapters enable the communication of gateways and fog servers using the other application protocols dis- cussed in Section 3.2.

3.3.3 IoT Gateway

Gateway functionalities are provided in two frameworks. The first, in Java language and developed specifically for Android devices, offers a visual interface for the gateway parameter management, and the second is comprised of a set of C language libraries for a Raspberry PI platform and includes no visual interface.

The gateway Protocols/Adapters component follows the same definition as its counterpart on the cloud server. It contains a multi-adapter manager for data forwarding among IoT devices; however, it disregards the direct publication of data for applications, given the non-existence of a publish subscribe and application services component. Instead, it acts as a data relay for the storage and data processing component. The gateway storage is slightly different from its counterpart on the cloud server. Since gateways have the lower computing power and hardware resources, the storage is implemented in the form of a local cache that enables the verification of data integrity and processing of data for avoiding the forwarding of meaningless messages (i.e. duplicated observations). The fog/cloud interfaces are provided as Java Modules integrated in a Visual Interface Application and work as a SenML message relay to forward the IoT devices observations processed by the storage and data processing component to the cloud and fog servers.

3.3.4 Fog Server

The visual interface was developed in .NET as a web application that can be deployed on a Windows PC. The Storage and Data Processing, the Publish/Subscribe Applications Services, and the Protocols/Adapters components have the same functionalities and implementation of the Cloud Server, except that the data handled in the fog server are related to a localized region rather than to the entire city. The Fog to Cloud interface provides forwarding services to the cloud for gateways that cannot reach the cloud server adapters and IoT devices directly connected to the fog server.

3.4 PLATFORM EXTENSION TO FORESTRY AND WILDFIRES PREDICTIONS

Figure 3.5 displays an overview of the use-case of the previous platform, but adapted to a forestry- based wildfire prediction. IoFT devices collect observations of different environmental parameters and publish them in the IoFT data aggregation platform deployed in the cloud. Complementary data such as those collected by governmental sources are sent or mined to the platform. All space-temporal data aggregated in the platform are preprocessed and normalized in terms of data representation. Then, ML (ML) models are trained considering the fire-related data and published for prediction tasks. The training process involves the collection and preparation of fire scars and vegetation and climatic historical data.

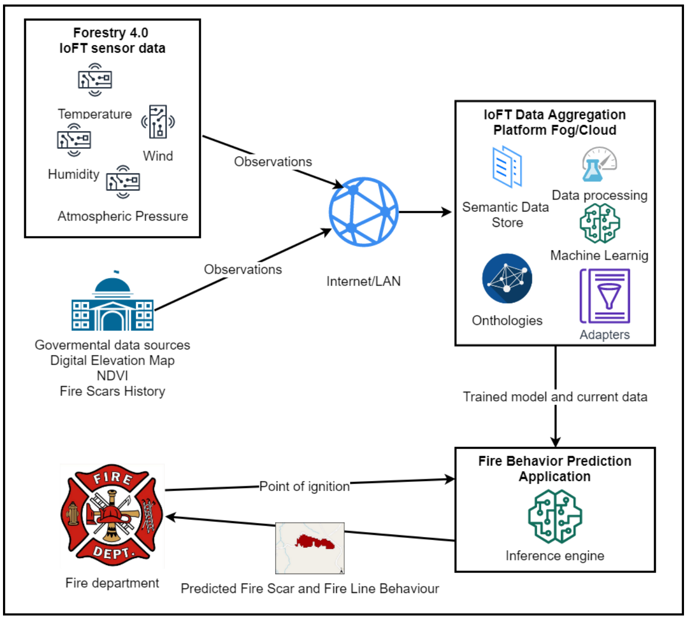


Figure 3.5: Overview of the proposed solution [4].

Governmental institutions such as fire departments can use the fire behaviour prediction application for predicting a fire evolution and the final scar. The application inference engine uses the models previously trained, the current environmental observations, and a fire ignition point for the prediction of fire behavior and fire scar.

Following the previously described three layer-based architecture and considering the elicited functional and non-functional requirements to be met, this sub-section presents aspects related to the platform adaptation.

An IoFT Virtual Sensors API (IoFTVSENS) implemented in .Net is proposed for the implementation of Virtual IoFT devices. It exposes APIs that provide functionalities for virtualization and promotes the integration of non compatible data-sources to the platform, offering high-level services for communication and service management (sensor registration, authorization, etc.). As an example, producers based on web crawlers, database streaming, or extract/transform/load processes can be integrated to the platform with few efforts. Section 4 provides examples of virtual IoFT implementations.

Gateway functionalities were supplied in a framework (IoFTGate) written in C for Arduino Platform, which acts as a data relay for the storage and data processing component. It considered several network interfaces such as Bluetooth, ZigBee, WiFi, and Ethernet, CoAP, AMQP, and MQTT as communication protocols with the IoFT devices. A storage and processing module works as a local cache for the verification of data integrity and the processing of data towards avoiding the forwarding of meaningless messages. Cloud interfaces were provided as another module and work as a SenML message relay to forward the IoFT devices observations processed by the storage and data processing component to the cloud. They were implemented as a set of MQTT publishers, one for each type of observation. The Settings Manager together with the adapter configurations handle the routing to the cloud.

The cloud server offered several Platform and Infrastructure Services and was deployed on the Microsoft Azure cloud services provider and Amazon web services. Its main functionalities are maintenance of the semantic model, storage of semantic observations, application of data processing, and knowledge extraction for ML models.

The Ontology Management component provides a visual interface and functionalities for creation, update, deletion, and deployment of definitions in the ontology. It was deployed as a Web application over an Azure App Service and use of Azure Blob Storage for storing ontology definitions and generating valid IRIs. It also enables the edition and publication of new ontology definitions provided by the Storage and Data processing component.

The MQTT Adapter component handles configurations such as port numbers, topics for brokers, among other parameters necessary for a direct publication of IoFT data from the gateway to the Publish/Subscribe Application Services. The Publish/Subscribe and Application Services exposes a set of RESTful endpoints for the query of data using SPARQL query language and returning data in SenML and RDF format. Moreover, it enables the applications to subscribe to an MQTT broker where the data received by the Storage and Data processing component published. The Storage and Data processing component provides all functionalities for the storage and handling of SenML observed data mapped to RDF format, according to the definitions in the Ontology Management component. Moreover, data transformation processes such as re-sampling and interpolation can be defined.

Applications are considered clients that use the platform services for several purposes. They can implement real-time monitoring by querying data from the cloud and receiving them in a publish/subscribe approach. Moreover, they can take advantage of deployed ML models for decision support based on the last environment state stored. As an example, an application that uses current data to predict a wildfire behaviour is defined in what follows.

In a first case, the application subscribes to one or more topics in the MQTT broker of interest (e.g., a fog server for localized processing) or in a cloud server for city scope. Whenever data of indicators are published in those topics, the machine learning model is trained and a user (or another system) can make inferences using the trained regression model. In a second case, users can design big data algorithms (i.e. data mining techniques) that query data from cloud/fog servers using SPARQL sentences over RESTful web services, extract meaningful insights about the indicator value, and trigger actions over actuators (i.e. high temperature detected triggers anti-fire systems).

Figure 3.6 shows a summary of the wildfire behaviour prediction use case detailed in what follows.

3.4.1 Data Producer Layer

Several Virtual IoFT devices were developed for the collection of meteorological data on 11 climatic features related to Atmospheric Pressure (AP), Air Temperature, Relative Humidity, and Wind obtained from five automatic stations (Figure 3.7) spread across the district [96]. Observations were taken at one-hour intervals since 2000 and sent to the platform with their proper sampling date and time. One virtual sensor was considered for each feature and for each location, totalling 55 environment-related IoFT devices. The data observed were forwarded to the gateway in SenML format. Figure 3.8 shows an example of a SenML encoded temperature observation.

Point of ignition and scars IoFT devices obtained data on historic fire spots, and each observation was defined by its location (latitude, longitude) and date and time of the events. Fire spots data were based mainly on a Moderate Resolution Imaging Spectroradiometer (MODIS) from AQUA, TERRA, NOAAs-15, 16, 17, 18, and 19, METEOSAT-02, and GOES-12 satellites. The second dataset contains data on fire scars identified in the same period with 30-meter space resolution and 16-day temporal resolution. Scars were generated by the processing of Landsat-8/OLI, CBERS-4/MUX and Resourcesat/LISS images.

NDVI data were obtained from MODIS /Terra Vegetation Indices available in [97]. They were sampled by the IoFT device in a 250-meter spatial resolution, measured at 16-day frequency, and identified as a contour observation in the ontology. Historical point of ignition, NDVI, and fire scars data were collected by several web crawlers which acted as IoFT virtual devices. Each crawler considered 1-hour verification interval and was implemented as .Net applications running on a local server.

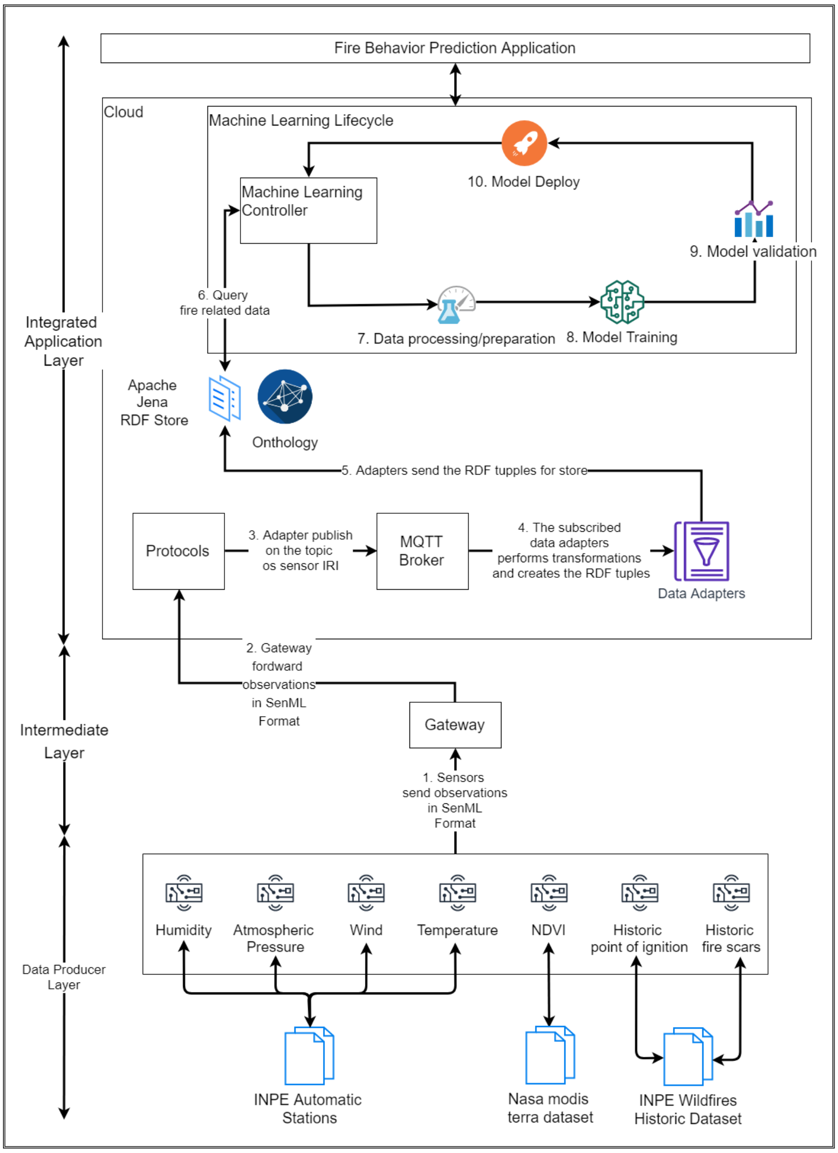

Fire Behavior Prediction Application
Cloud
Machine Learning Lifecycle
10. Model Deploy
Machine Learning Controller
9. Model validation
6. Query fire related data
7. Data processing/preparation
8. Model Training
Integrated Application Layer
Apache Jena RDF Store
Onthology
5. Adapters send the RDF tupples for store
Protocols
3. Adapter publish on the topic os sensor IRI
MQTT Broker
4. The subscribed data adapters performs transformations and creates the RDF tuples
Data Adapters
2. Gateway fordward observations in SenML Format
Gateway
Intermediate Layer
1. Sensors send observations in SenML Format
Humidity
Atmospheric Pressure
Wind
Temperature
NDVI
Historic point of ignition
Historic fire scars
Data Producer Layer
INPE Automatic Stations
Nasa modis terra dataset
INPE Wildfires Historic Dataset

Figure 3.6: Platform Applied to Wildfire Use Case [4].

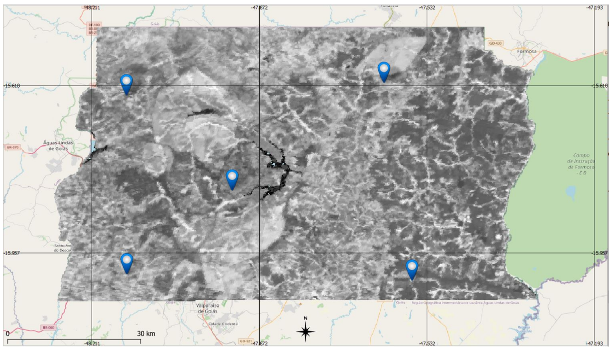
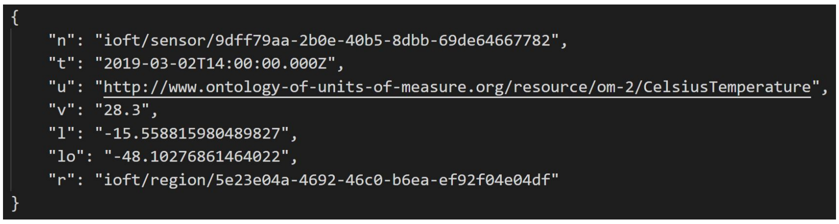


Figure 3.7: Federal District Automatic Stations and NDVI [5]



```
{
    "n": "ioft/sensor/9dff79aa-2b0e-40b5-8dbb-69de64667782",
    "t": "2019-03-02T14:00:00.000Z",
    "u": "http://www.ontology-of-units-of-measure.org/resource/om-2/CelsiusTemperature",
    "v": "28.3",
    "l": "-15.558815980489827",
    "lo": "-48.10276861464022",
    "r": "ioft/region/5e23e04a-4692-46c0-b6ea-ef92f04e04df"
}
```

Figure 3.8: SenML example of a temperature observation [5].

3.4.2 Intermediate Layer

Environment monitoring devices connected to gateways via WiFi LAN (Local Area Network) used CoAP protocol. On the other hand, Virtual IoFT devices were connected to the gateway via Ethernet and REST. All gateways connected to the cloud through WiFi and Data Over Cable System Interface Specification (DOCSIS) modem following an event-oriented approach considering MQTT a communication protocol, and were configured towards acting as SenML message relays.

3.4.3 Integrated Application Layer

SenML data received by the adapters/protocols are published in a topic of the MQTT Broker according to the sensor IRI. A data adapter converts them to RDF tuples and store them in the Apache Jena RDF Store. At this point, all historically collected data are stored and available for queries with SPARQL .

Amazon Sagemaker managed services promoted ML processing, since it enables the definition of a complete lifecycle of a machine learning solution taking advantage of cloud computing resources. Initially, data are retrieved a to be used for the training of the models. In our application, data are queried from Apache Jena RDF store and mapped to tabular records. In a second step, the data are processed and prepared for model training. Since they were available in different spatial and temporal resolutions, a re-sampling process was required.

Supervised learning was applied to solve the wildfire prediction problem. Given a fire situation at time t_1, the models estimate the sectors that will be affected by the fire in the future t_2. Since the data on fire scars are in a 16-day temporal resolution, we do not have all the frames (only the final one) to find out the way the fire progressed. Towards simulating the frames, we assumed all fires (starting at an ignition point) would progress no more than one sector distance from another burned sector in each step. Moreover, our dataset is composed of samples of 3x3 sectors generated from two consecutive frames (t_1, t_2). For each sector not affected by the fire and with at least one burnt neighbor, the ”burn” class represents it if burnt in t_2. Otherwise, the unburnt class is assigned to it.

Finally, at the top of the IoFT wildfire use case is the fire behavior prediction application, which estimates the scar resulting from the spreading/advancing of the fire given a point of ignition.

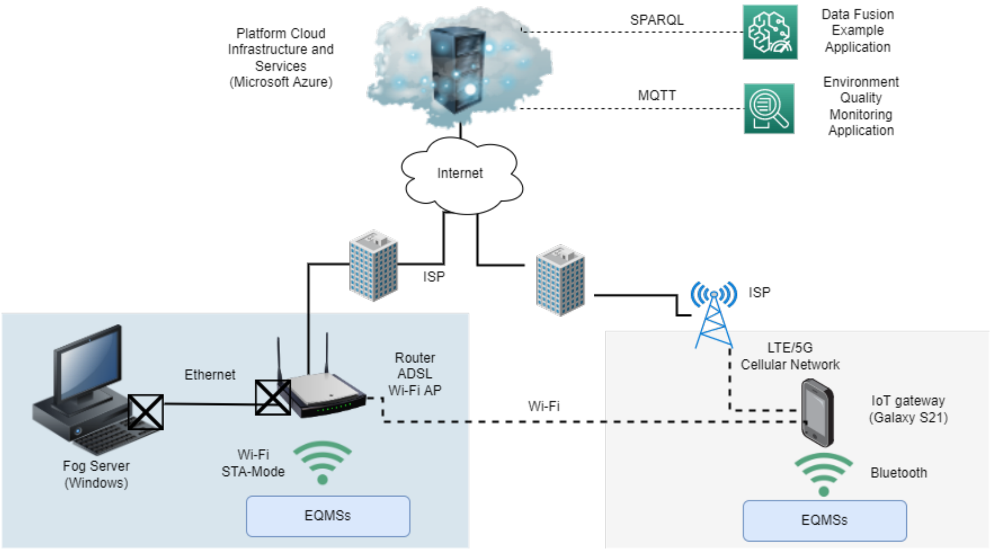


Figure 3.9: Experimental network topology [3]

3.5 EXPERIMENTS AND PERFORMANCE EVALUATION

Metrics such as latency and resources (CPU and Memory) consumption were analyzed for the performance evaluation of the platform. The tests considered a real use case of environment quality monitoring where PM2.5, CO2, O3, relative humidity, latitude, longitude, temperature, and noise data are monitored. A set of IoT devices, called Environment Quality Monitoring Station (EQMS), was developed by the combination of Arduino MEGA development platforms, ESP8266 Wi-Fi, and HC-06 Bluetooth modules for providing network access, GY-NEO6MV2 modules for geo-positioning, microphones for noise measurements, and DSM501, MG811, MQ131, and DHT11 sensors for the other indicators.

Figure 3.9 displays the network topology where EQMSs connect through a Wi-Fi Access Point hosted in a router connected to the fog server via Ethernet. The router also enables communication with the cloud services. On the other hand, EQMSs connect to the IoT Gateway through a Bluetooth connection using the HC-06 module. The connection between IoT gateway and fog server is provided through a Wi-Fi local area network, and a 4G LTE network adopted connected the IoT gateway and cloud services. An application for environmental quality monitoring is subscribed to the cloud publisher on each observation topic and uses the received observations for the detection and alerts on adverse environmental situations.

Several values of observations were transmitted from EQMSs to the fog server, thus enabling a comparison of latency for MQTT , CoAP, and REST adapters (see Figure 3.10). Table 3.1 shows the maximum and average latencies for each adapter. CoAP showed the lowest latency, i.e., 242 ms, whereas that of REST and MQTT was slightly higher, i.e., near 500 ms. The performance of the three adapters was stable during the experiment in which the REST-based one showed 801ms maximum latency. Table 3.2 shows a similar behaviour of the RAM memory consumption. Since CoAP is the most lightweight protocol, it consumed 2751kb, whereas REST and MQTT

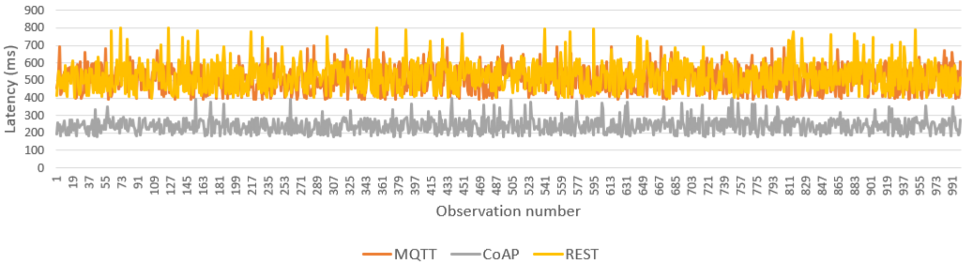


Figure 3.10: Transmission latencies of different adapters over Wi-Fi (for the path EQMSs to Fog Server).

Table 3.1: Maximum and average latencies for each Adapter/protocol.

Adapter	**MQTT**	**CoAP**	**REST**
Max. latency	699 ms	399 ms	801 ms
Avg. latency	509.83 ms	241.472 ms	523.55 ms

consumed 2965kb and 3259kb, respectively.

Figure 3.11 displays the overall latency between the EQMSs and the cloud subscriber applications passing through an LTE IoT Gateway. In general, the overall latency increased in relation to the fog server setup. The overall behaviour was similar, and CoAP was the most performative adapter. However, both REST and MQTT showed acceptable latencies, with average values below 630ms.

Figure 3.12 displays the overall latency between the EQMSs and the cloud subscriber applications but now through an enabled Gateway. The average latency was decreased nearby 34ms in relation to the LTE access network, in a preliminar evaluation.

Both experiments, LTE and 5G where performed within the same Internet Service Provider, however, these results are not conclusive since no data about networks loads were known at the time of the experiments.

Table 3.3 shows a summary of the latency results obtained. The CoAP protocol showed the best performance and, regarding celular network over the gateway, 5G presented better latencies than LTE, but additional experiments are required for a better evaluation.

A performance comparison with similar platforms would be interesting for a better evaluation of our proposal in terms of state-of-the-art. However, a fair comparison seems difficult, or even impracticable, given the difference in platform objectives, network architectures, topologies, and other features. Moreover, since no open-source environmental platform was identified, efforts for the construction of other platforms would be necessary for an effective comparison.

Table 3.2: EQMSs Maximum RAM memory used by each Adapter/protocol.

Adapter	**MQTT**	**CoAP**	**REST**
Memory used	3259kb	2751kb	2965kb

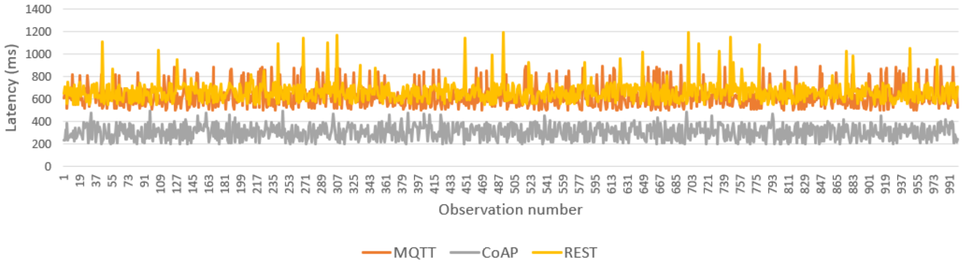


Figure 3.11: Transmission latencies considering different adapters over LTE (for the path EQMSs - IoT Gateway - Cloud - Subscribed Application).

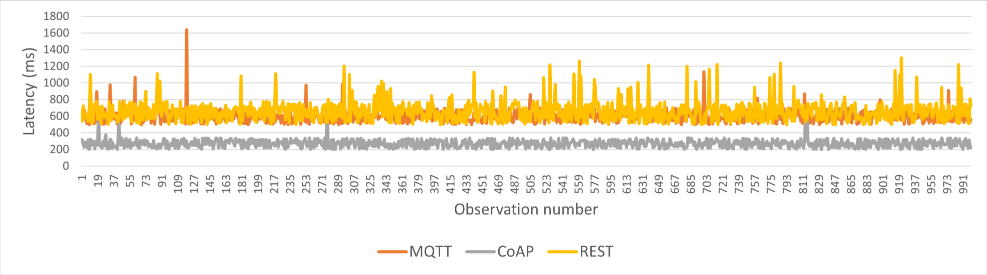


Figure 3.12: Transmission latencies considering different adapters over 5G (for the path EQMSs - IoT Gateway - Cloud - Subscribed Application).

Table 3.3: Performance Summary.

Adapter	MQTT	CoAP	REST
Max. latency LTE	899 ms	495 ms	1192 ms
Avg. latency LTE	627.52 ms	306.67 ms	661.21 ms
Max. latency 5G	1640 ms	692 ms	1302 ms
Avg. latency 5G	604.45 ms	272.46 ms	657.96 ms

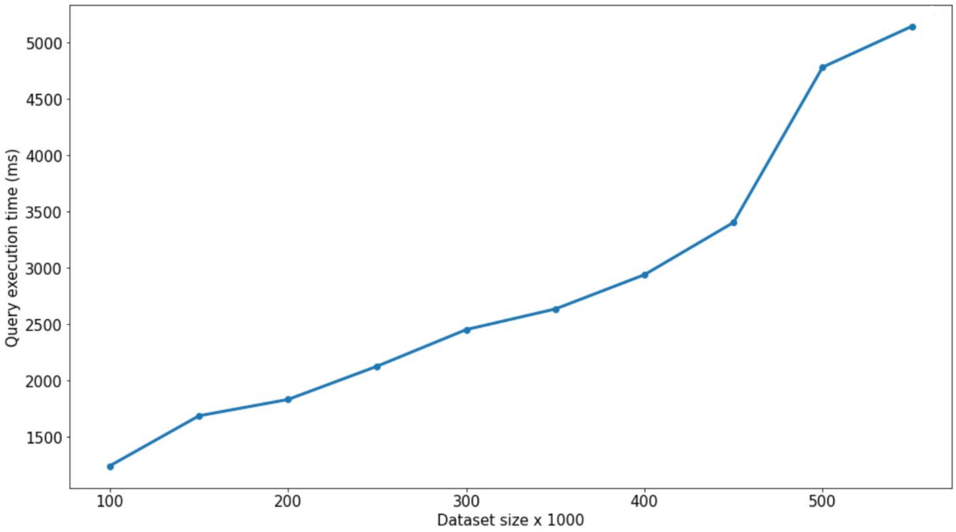


Figure 3.13: Query delay in respect to dataset size

Among the applications developed for the testing of the platform, one considered spatio-temporal data related to Weather Madrid (WM) [98] and Air Quality Madrid (AQM) [99] datasets and retrieved by SPARQL queries. AQM data were sub-sampled and a model similar to that proposed by Zhao et al. [82]) was trained considering categorical feature "Events" (Rain, Fog, Hail and Thunderstorm), available in WM dataset, for the classification in target dataset AQM. The results showed its high accuracy and the importance of data fusion methods. However, the modelling of complexities of such approaches and their adequacy to the problem characteristics showed data fusion models must be developed specifically for a particular use case for the obtaining of better quality results.

Measurements of the execution of SPARQL queries for RDF database (Apache Jena) were also collected. Figure 3.13 shows queries are performed as the data collected increase. The test dataset was defined in such a way the query always returned half of the data of the set. A monotonically crescent behavior can be observed in the increase of the execution time; queries with returns of more than 500 thousand samples required approximately five seconds.

3.6 REQUIREMENTS FULFILLMENT

All the requirements elicited in section 3.1 were met, as shown in the next two sections.

3.6.1 Treatment of Functional requirements.

Sensors integration: IoFTSENS API provides functionalities for sensor registration and identification enabling communication with the settings manager hosted at gateway and cloud services. Sensor meta-data, such as sampling process, unit of measurement of observations, and type of feature observed are represented in the data schema by SOSA ontology.

Devices Heterogeneity: The platform does not bind the type of sensor to be used. The dynamic and semantic approach implemented enables any data formatted in SenML . In case of non-standardized data, the platform accepts the implementation of virtual IoFT devices.

Heterogeneous IoFT data: The platform deals with heterogeneous data following a semantic representation rather than a fixed scheme. Data are formatted in SenML format in devices, thus enabling interoperability. If a device is not SenML compatible, the IoFTSENS API provides functionalities to simplify the data transformation. The schema proposed is flexible enough for handling all data types, unit of measures, and metadata previously defined in the ontology, and the application can consume the data following a widely used semantic standard such as RDF .

Context information: Observed data consider both time and location that relate to a geo:region ontological definition. Other types of observational context (i.e. type of sensor and sampling process) are managed by the platform through the ontology and other context information can be inserted through an ontological extension.

Resource Limitation: The platform considers standards such as ZigBee, CoAP, and SenML designed for devices with limited resources.

Data-Related Services: The platform enables the definition of customized pre-processing strategies that can be deployed in the gateway, using the IoFTGate framework, and in the cloud, considering the processors service functionalities. Moreover, it provides data services for query and event-based streaming through Apache Jena RDF SPARQL and Publish/Subscribe application services.

Event Management: Large volumes of observed data are handled by the platforms according to an event-driven solution. The MQTT server deployed on the cloud handles the topics defined by the ontology when observations arrive in an adapter, and the data converted to RDF are published in those topics. Consumer applications receive such data in almost real-time. The MQTT Server is managed by a cloud service provider with auto-scale capabilities.

Inference Services for Decision Support: With machine learning as a service, the platform enables both training and inference by several models provided by AWS SageMaker.

3.6.2 Treatment of Non-functional requirements.

Interconnectivity: IoFT gateway and cloud services implement several application protocols (e.g., CoAP, HTTP, MQTT , and AMQP). Several network protocols regulated by normative organizations such as Blue- tooth, Wifi, ZigBee, 6LoWPAN, and Ethernet were considered.

Extensibility: The flexibility introduced by the semantic scheme enables the definitions of any type of sensors, sampling procedure, and units of measure for the extension of the platform to almost any scenario and device.

Real Time Treatment: Since the platform considered an event-based approach, all data can be transmitted from devices to consumer applications in almost real time.

Scalability: The platform enables the accommodation of a large number of sensors, given the layered architecture followed. As the number of devices increases, the number of deployed gateways can be increased and the cloud service can be reached through the use of the MQTT adapter, thus avoiding bottlenecks. The supply of platform services in the cloud also grants several ways for scalability.

Interoperability: The platform supports communication protocols widely used in the IoT context.

3.7 CONCLUSIONS

This chapter, after a literature review of SC-related platforms, communication protocols and data interchange and representation strategies, presented an IoT-based platform for the environment SC domain, following a three-layered IoT architecture. The platform enables the collection, storage and processing of data from the city environment, in a local region (i.e., neighborhood or building) using Fog resources (for the local processing), and at the city levels through services and resources dynamically deployed in the cloud.

Some key points related to the platform must be highlighted:

i) Use of the adapter concept for a seamless integration of heterogeneous sensors;

ii) Fog and Cloud Interfaces functionalities, which improved the flexibility, since the architecture of the platform can be adapted to almost all use cases;

iii) Adoption of SenML towards compatibility in terms of data representation; however, the mapping of the IRI of the sensor (in both the gateway and the fog server) is mandatory for reductions in message overload and the consumption of resources in constrained devices;

iv) Possibility of application of big data techniques and machine learning on the cloud over city environmental data through a coherent semantic and standardized data model that defines the interrelation of the data in a robust way and extraction of implicit information through ontological reasoning; and

v) Presentation of queries made by client applications to the platform are presented in a standardized way through the use of SPARQL and RESTFul for query endpoints and SenML and RDF as the format of the response object.

The platform was extended to the IoFT context and a set of functional and non- functional requirements was proposed and validated.

4 WILDFIRE RISK AND BEHAVIOUR PREDICTION

Discussions on global climate changes and forest ecosystems risk have become more prominent. In 2019, wildfires in Australia, South Africa, and Brazil gained worldwide attention and were one of the main drivers responsible for losses in forested areas and devastating biodiversity and human and economic damages. Therefore, interest in wildfire-related studies has increased over the past decade [100].

The Brazilian Federal District (FD) region has shown increasing fire activity since the year 2000 [1], thus motivating our interest in the study of its effects. Such fires have impacted the native species of the region, even in protected zones [101], decreased the quality of life of its inhabitants, and forced local governments to spend resources on their fighting. However, not much research on the prediction of wildfires in the FD region, inserted in the Brazilian savanna (the Cerrado biome), has been developed [25].

This chapter focuses on the study of wildfires in the Federal District region and presents a review of the related work, the characteristics of the region, and two approaches for fire risk and behaviour predictions based on ML.

4.1 RELATED WORK

4.1.1 Wildfires risk predictions

ML-based fire prediction tools generally follow the flow shown in Figure 4.1 [102][103], which takes an observation of the current parameters of a sector (sub-region) as input and returns an estimate of the fire risk. The repetition of this process for all sectors enables the generation of risk maps on information for governmental decisions. The flow shows several originating features and machine learning models can be considered for fire prediction and the identification of the most suitable model for the estimation of Wildfire-related events is a challenge.

Rodrigues and de la Riva [100] considered forested areas in peninsular Spain, and a binary classification problem (classes "High", related to at least two fire ignitions and "Low", in other cases) for the dependent feature "fire occurrence". In a 1-km resolution grid, they considered human presence, Wildland Urban Interface (WUI), changes in demographic potential (1991-2006), Wildland-Agricultural Interface (WAI), electric power lines, engines, and machines working in or close to forest areas, the density of agricultural machinery, presence of roads, railways, and tracks and their accessibility explanatory features.

Three ML algorithms, namely, Random Forest (RF), Boosting Regression Trees (BRT) , and Support Vector Machine (SVM) were implemented and compared with traditional methods (e.g., Logistic Regression (LR)). RF showed the best Area Under the ROC Curve (AUC) with 0.74

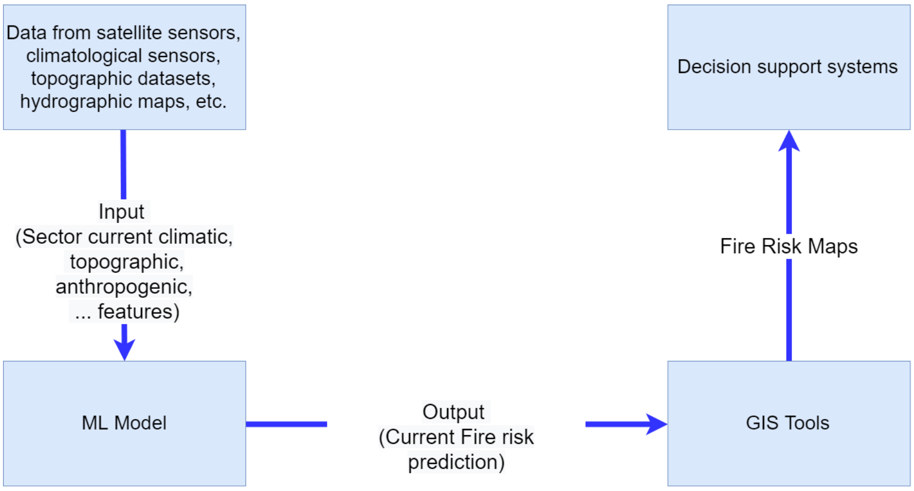

Data from satellite sensors, climatological sensors, topographic datasets, hydrographic maps, etc.
Input (Sector current climatic, topographic, anthropogenic, ... features)
ML Model
Output (Current Fire risk prediction)
GIS Tools
Fire Risk Maps
Decision support systems

Figure 4.1: Fire prediction use-case example [83][84]

accuracy and BRT and SVM showed 0.730 and 0.709, respectively. Regarding the importance of features, the authors followed two different approaches - one that considers the node purity of RF and BRT measured by the Gini criterion [104], and another based on an AUC procedure known as jackknife estimator of variable importance [105].

Ghorbanzadeh et al. [106] studied wildfire risk in the Mazandaran Province of Iran, considering 17 explanatory features (see Table 4.1) - four for anthropogenic factors and the others for meteorological, topographic, and hydrographic conditions. As in the previous approach, the study considered a 1-km sector to classify wildfire risk (the following five classes were considered: Very High, High, Medium, Low, Very Low). Three ML models (Artificial Neural Networks (Artificial Neural Network (ANN), SVM, and RF) were trained and validated through Cross-validation (4-fold), and RF showed the highest accuracy (0.88). Regarding feature importance, the slope aspect described the data for RF and SVM models better, whereas distance to road showed an impact for ANN model. The study also proposed a model that calculates the normalized feature importance based on Hong et al. [107].

Ghorbanzadeh et al. [102] focused on Northern Iran to predict wildfire susceptibility. As in the previous study, the authors considered 16 explanatory features and a k-fold cross-validation to validate the performance (0.801 AUC-ROC) of an ANN. They proposed a social/infrastructural vulnerability index using a geographic information system multi-criteria decision-making (GIS-MCDM). The Infrastructural Vulnerability Indicators (IVIs) are primarily based on land use types such as building, agriculture, and recreational areas, and those in conjunction with the risk of fire enable the generation of risk maps with information on fire occurrence, but also the damage (i.e. economic losses) caused. On the other hand, social vulnerability indicators consider factors such as population, age, gender, housing, education, health services, occupation, and facilities, which describe social inequities among people that presumably increase a society's vulnerability to natural hazards.

Miller and Ager [108] reviewed several studies and identified three main components of risk, namely likelihood, intensity, and effects. The former is related to ignition or burning; intensity is associated with fire behaviour, and effects refer to ecological, social, and economic vulnerabilities (as in the previous study). Since our research focuses on the estimations of the fire occurrence risk, only the likelihood component were considered - intensity and effects will be treated in future research.

Jaafari et al. [109] focused on the Hyrcanian Iranian region and designed a solution according to 11 explanatory features (see Table 4.1),of which 10 are similar to those used by Ghorbanzadeh et al. [106].The authors considered a 30-m sector resolution and combined the Adaptive Neuro-fuzzy Inference System (ANFIS) ML model with Metaheuristic Optimization Algorithms (MOAs). They also applied four different MOAs, namely Genetic Algorithms (GA), Particle Swarm Optimization (PSO), Imperialist Competitive Algorithm (ICA), and Shuffled Frog Leaping Algorithm (SFLA), and several metrics, such as RMSE, accuracy, sensitivity, specificity, false alarm ratio, Kappa, success rate, and prediction rate for validation. ICA-based ANFIS showed the best performance in the validation dataset for those metrics, with 0.99 accuracy and 99.09% prediction rate. Similarly, Jaafari et al. [103] studied other two hybrid models that rely on ANFIS and firefly MOA. The same region and conditioning factors were considered, and firefly MOA showed a 0.89 AUC prediction rate.

In a 1-km spatial resolution, Kim et al. [110] analyzed the influence of human activity using environmentally dependent features, such as precipitation, elevation, topographic wetness index, Fire Weather Index (FWI), and forest type, and anthropogenic ones (e.g., population density and distance from urban area). Maximum Entropy (Maxent) and Random Forest models predicted the spatial distribution of forest fire, and AUC metric validated their performance. The analysis revealed a strong correlation of fire probability with variables related to human activity and accessibility. The AUC values were higher in Random Forest in comparison to Maxent. The study considered the South Korean region.

Nami et al. [111] applied the quantitative data-driven Evidential Belief Function (EBF) model theoretically supported by the Dempster–Shafer uncertainty theory, and the results produced a distribution map of wildfire probability constructed for the Hyrcanian region (Iran). The dependent feature was considered for a classification problem with moderate, high, and very high probability classes. The authors used some of the explanatory features introduced in this section (see Table 1), and a 30 m spatial resolution. The AUC validation showed an 84% prediction rate.

Rihan et al. [112] studied fires in the Mongolian plateau to predict probabilities and identify their main factors. They considered a 50-meter per sector spatial resolution and used only the RF model, with a 0.951 AUC-ROC. Differently from other studies, they did not include anthropogenic features (see Table 4.1) to explain the fire occurrence. The most important feature was Fraction of Vegetation Coverage (FVC), defined as the percentage of the vertical area of vegetation projected on the ground as a percentage of the total area, which also reflects the vegetation growth.

Sayad et al. [113] studied an area formed by several zones in the center of Canada and built a set of environmental data related to NDVI, Land Surface Temperature (LST), and fire indicator (thermal anomalies) - the fourth column represents the corresponding class (fire or no fire). Two ML models (ANNs and SVM) were trained and the experimental results showed an above 95% prediction accuracy, validated through metrics, such as cross-validation and regularization.

Tonini et al. [114] analyzed the Italian region of Liguria and elaborated a wildfire susceptibility map by applying Random Forest. Susceptibility was assessed according to the probability of an area burning in the future in regions of past wildfires occurrence and which were the geo-environmental factors that favored their spread. The explanatory features considered were DEM, Slope, Aspect, Distance to Urban Area, Road, Pathways and Crops, Protected Area, Vegetation Type, and Neighboring Vegetation. The Root Mean Square Error was computed for validating the model at an approximately 91% success rate.

Gholamnia et al. [115] compared 11 ML models in the Mazandaran Province of Iran. The wildfire inventory data were collected at 1-km resolution and considered topographic, hydrographic, meteorological, and vegetation features. The ML methods applied were ANN, dmine regression (DR), DM neural, least angle regression (LARS), multi-layer perceptron (Multi Layer Perceptron (MLP)), RF, Radial Basis Function (RBF) , self-organizing maps (SOM), SVM, decision tree (DT), and logistic regression (LR). The authors considered 3-fold cross-validation for accuracy assessment and AUC-ROC assessed the accuracy of the ML approaches. RF showed the highest accuracy (88%).

Kaur and Sood [116] proposed a framework for real-time detection and prediction of forest fires. Initially, the system employed a Bayesian Belief Network (BBN) for real time fire detection and a fuzzy system to compute the wildfire susceptibility index. For the training of the BBN, the authors considered a dataset composed of records labeled as fire event or non fire event, and climatic features, such as temperature, precipitation, relative humidity, wind speed, atmospheric oxygen, carbon dioxide and monoxide levels. BBN showed an AUC-ROC value of 0.93. On the other hand, the fuzzy model was trained for the 5-classification considering temperature, relative humidity, precipitation and wind speed, for a 91% prediction accuracy.

The aforementioned studies considered several worldwide regions and were analyzed towards the identification of the main originating factors (features) and ML models applied. On the other hand, the studies below focused on the Brazilian Cerrado biome.

Galizia and Rodrigues [117] predicted wildfire occurrence through Random Forest and cluster analysis focusing on the influence of eucalypt plantation on wildfire occurrence. The dependent variable was modeled as presence or absence of fires, and the explanatory features were distance to several land cover regions, elevation, aspect, temperature, wind speed, relative humidity, population density, distance to roads, and electric and train lines. RF performed with a 0.75 AUC-ROC.

De Bem et al. [25] considered the Federal District region and used ANN and LR to predict the dependent feature fire occurrence; explanatory features slope, aspect, elevation, water supply, distance to road and urban areas, land use, population density, and NDVI were considered for the model training. Both models' performances were similar; however, ANN showed better AUC-

ROC (0.77) and accuracy. The authors compared the significance of each variable to the models and concluded the main driving aspects of the burned area distribution were land-use type and elevation.

This research ignored climatic variables and considered the study area with small spatial variations due to its size and local characteristics. However, studies in other regions [106, 103, 109, 112] showed climatic data are highly significant for wildfire prediction. Our study was motivated by those of De Bem et al. [25], who claimed "Fire risk prediction studies in the Brazilian savannas are still scarce", confirmed by Gomes, Miranda and Bustamante [118], citing Pereira et al. [119], who consider "fire modeling studies on the Cerrado are scarce and must be improved for the development of a more systemic approach".

Other wildfire-related studies about the Cerrado biome (Santana et al. [120], Guedes et al. [121], Greison [122], da Silva et al. [123], dos Santos et al. [124], Pereira et al. [125]) addressed other wildfire-related problems from a non-predictive perspective and focused on issues, such as detection of burned areas and statistical analyses.

Table 4.1 shows the conditioning factors identified in recent studies for fire risk prediction, and many of them can be used for the construction of decision systems. Related studies that included analyses of feature importance showed all of them can contribute to the refinement of the risk-prediction model with a stronger or weaker impact. Moreover, anthropogenic factors, i.e., activities of local people, tourism, or any human intervention can be considered wildfire conditioning factors. In this sense, apart from natural conditions, the human influence in the area of study must be analyzed and taken into account in the ML model training.

Historical data on fire events can be available in multiple forms/resolutions. Regarding fire prediction, the identification of the ignition point is crucial for the achievement of higher accuracy. Data must be cleaned for the identification of the fire ignition location and date time as exactly as possible. However, a specific ignition point does not represent the reality, since conditions for the start of fire are related to an area of same characteristics rather than specific coordinates. The previous studies have shown a consensus over a 1-km spatial resolution being sufficient for wildfire risk prediction.

Table 4.1: Main originating factors (features) considered for wildfire risk prediction

Proposal/Features		Topographic	Hydrological	Meteorological	Anthropogenic	Others
[100]	Fire Occurrence (High, Low)	-	-	-	Presence (human, electric line, power line, roads, railways and tracks), WUI, Demographic Changes, WAI and Density of agricultural machinery.	-
[106]	Susceptibility to Wildfires (Very High, High, Medium, Low, Very Low)	Slope aspect, Slope(%), Altitude, Topographic Wetness Index (TWI) , Landform and Plan Curvature.	Distance to stream and Annual rainfall	Potential solar radiation, Annual temperature and Wind effect (speed and direction)	Land use (Forest, Non-Forest, Farm, Settlements), Distance to Village, Distance to Road and Recreation Area.	NDVI
[109]	Fire Occurrence (Fire, No-Fire)	Slope aspect, Slope(%) and Altitude	Proximity to rivers and Annual rainfall.	Temperature and Wind effect (speed and direction)	Land use (Forest, Non-Forest, Farm, Settlements), Distance to Settlements and Roads	NDVI
[103]	Fire Occurrence (Very High, High, Moderate, Low, Very Low)	Slope aspect, Slope(%), Altitude and Soil Type	Annual rainfall.	Temperature and Wind effect (speed and direction).	Land use (Forest, Non-Forest, Farm, Settlements), Distance to Settlements and Distance to Road	-
[110]	Fire Probability (Regression)	Altitude, Forest Type and TWI .	Precipitation	-	Population Density, Number of National Park Visitors and Distance to Urban Area.	FWI
[111]	Wildfire Risk (Very High, High, Medium, Low, Very Low)	Slope aspect, Slope(%), Altitude, Plan Curvature, TWI , TRI and Soil Type	Rainfall, Evapotranspiration and Distance to rivers	-	Distance to Settlements and Distance to Road.	-
[112]	Fire Probability (Regression)	Slope and Aspect, Land Cover: Land Use Degree, Diurnal Temperature Range, Frost Day Frequency and Potential Evapotranspiration	Precipitation	Mean Temperature, Average, Minimum and Maximum Temperature, Vapor Pressure and Wet Day Frequency.	-	FVC
[113]	Fire Occurrence (Fire, No-Fire)	-	-	Land Surface Temperature (LST)	-	NDVI
[114]	Fire Probability (Regression)	Slope and Aspect.	Land Type.	-	Distance to Urban Area, Road, Pathways and Crops	Vegetation type
[115]	Susceptibility to Wildfires (Very High, High, Medium, Low, Very Low)	Slope aspect, Slope(%), Altitude, TWI, Landform and Plan Curvature.	Distance to stream and Annual rainfall	Potential solar radiation, Annual temperature and Wind effect (speed and direction)	Land use (Forest, Non-Forest, Farm, Settlements), Distance to Village, Distance to Road and Recreation Area.	NDVI
[117]	Fire Events (Fire	Distance to [illegible] and land		Temperature and Wind	Population Density, Distance to	

Table 4.2: Main ML models considered for wildfire risk prediction

Proposal	Models	Metrics	Winner ML Model-Accuracy
[100]	Random Forest Boosting Regression Trees Support Vector Machines	Area Under the ROC Curve (AUC)	Random Forest - 0.746
[106]	Artificial Neural Networks Random Forest Support Vector Machines	Cross Validation (4-fold) and AUC (false-positive)	Random Forest - 0.88
[109]	ANFIS ANFIS-GA ANFIS-PSO ANFIS-SFLA ANFIS-ICA	RMSE Accuracy Sensitivity Specificity False Alarm Ratio Kappa Success Rate Prediction Rate	ANFIS-ICA - 0.99
[103]	ANFIS-FA	AUC prediction rate	ANFIS-FA - of 0.89
[110]	Maxent Random Forest	AUC	Random Forest - Omitted
[111]	Evidential Belief Function(EBF)	AUC	EBF - 0.8
[112]	Random Forest	AUC	RF - 0.95
[113]	ANN and SVM	Several Metrics	ANN - 0.98 (Prediction Accuracy)
[114]	Random Forest	RMS	RF - 0.91
[115]	NN, dmine regression (DR),DM neural, LARS, MLP, RF, RBF , SOM, SVM, decision tree and LR.	AUC-ROC	RF - 88%
[117]	Random Forest	AUC-ROC	RF - 0.75
[25]	ANN and LogR	AUC-ROC	ANN - 0.77%

Regarding the ML models identified, Table 4.2 shows many have been considered for wildfire-related predictions. Most studies focused on the implementation of different models and their comparison for the selection of the one that performs best for the available dataset. RF was one of the widely used models and that showed higher accuracy for wildfire risk prediction, followed by ANN and SVM.

4.1.2 Wildfire Behaviour Prediction

[126] used Reinforcement learning (RL) to model an agent for fire spread direction (north, south, east, or west) considering temperature, wind speed and direction, land cover type, humidity, and intensity of rainfall. Satellite image data validated the approach, which proved accurate. Among five RL algorithms adjusted and compared, Asynchronous Advantage Actor-Critic (A3C)

showed the best, with 87.3% average accuracy and 0.92 Area under ROC curve (AUC).

[127] [104] considered environmental data related to vegetation, digital elevation, atmospheric pressure, temperature dew point, wind direction, wind speed, precipitation, and relative humidity. Convolutional ANN models were trained on FARSITE platform for fire spread predictions based on current neighboring conditions and the experiments showed 87% accuracy. Other validation metrics considered were Recall and F-Score. On the other hand, [128] used FARSITE to improve the modelling of fuel factors and fire perimeters. Monte Carlo-based RBF neural network was used for fuel adjustment estimations. The model was validated by the Otsuka-Ochiai metric, which indicates a 0.8 similarity between observed and predicted perimeters.

[129] employed deep learning methods for rating spread modelling on Corsica island. Given a fire situation, the model predicts where the fire will most probably occur in a 20km surrounding area. The inputs were wind speed, fuel moisture content, combustion heat, particle density, fuel height, and surface-volume ratio. The authors considered the Structural Mean Square Error (SMSE) as a validation metric and better SMSE (6%) results were obtained after 94 training epochs.

Similarly, [130] simulated the spread of forest fires by modelling the problem as a Markov Decision Process and considering wildfire an agent advancing over the region in response to wildfire-related parameters. The agent (fire) can choose to move in either the four cardinal, or four ordinal directions, or not to spread at all (same approach of [131]). The authors used LRCNN (Long-Term Recurrent Convolutional Neural Networks) to build a generative model for input to the Markov Reinforcement Learning Model based on analyses of the sequence of satellite-based data. The models showed a 70% burn boundary similarity (70% sectors of the boundary of the predicted area intersect with the sectors of the original fire scar boundary) and a near one Burn Area Ratio (ratio of predicted scar area and original scar area).

Perumal and Zyl [131] attempted to reproduce a fire behaviour with limited duration time of series data. Two recurrent neural networks (RNNs), namely Gated Recurrent Unit (GRU) and Long Short-Term Memory (LSTM) were used for the modelling of the behaviour. The authors aimed to determine whether a wildfire would continue to burn and the cardinal directions of the wildfire spread. GRU performance was worse than LSTM, however the models were not able to predict the continuation of the fire. The result was probably affected by the lack of auxiliary data, since only Fire Radioactive Power satellite data (FRP) and Elevation were considered.

[132] studied fire spread in the Tasmania region and defined a mathematical model of fire propagation over time considering environmental conditions (relative humidity, temperature, and wind speed) at ignition points. The unknown values in the mathematical model were determined by fitting the simulation data according to the results from a sensitivity analysis. The authors obtained a 0.897 Pearson's correlation coefficient for the model validation. The coefficient ranges between zero and one and higher values denotes a better correlation.

[133] focused on a wildfire spread dataset for the benchmarking of machine learning-based solutions. Several models (convolutional autoencoder, RF and logistic regression) were trained

considering the topographic, population density, weather, vegetation, drought index data. Better results were obtained by the convolutional autoencoder model.

The aforementioned studies reported several models yield accurate values in the prediction of fire behaviour, thus stimulating both identification and selection of the most appropriate ones for a non-studied area (in our case, the Federal District).

An extended review conducted by [134] showed the three most cited models in 37 studies on fire spread domain were ANN, SVM and RF, with ten, seven, and six citations, respectively (see Table 3 [134]). We decided to apply them based on their academic relevance and also adopted AdaBoost model, whose better performance in comparison with ANN was demonstrated by [135] in a similar task in the Mount Parnitha (Central Greece) region.

In relation to model evaluations, [136] validated a Monte Carlo-based model for fire propagation prediction. The authors considered historic data on 10 wildfires in the Wyoming region and analyzed several statistics in the observed and predicted fire perimeters.

However, [137] highlighted the evaluation of models´ performances should not be limited to perimeter observations, but include the predicted area. The authors analyzed metrics for the performance evaluation of models based on observed and predicted burned areas such as Sørensen similarity, Jaccard coefficient, Shape Deviation Index (SDI), and Area Difference Index (ADI).

Regarding feature definitions, several studies (such as those listed below) have reported different conditioning factors associated with fires and their behaviour; however, the lack of a consensus on such factors hampers the proposal of a fire behaviour model to be applied in any region. By considering different feature sets for training, a model (regression or classification) that shows good performance for a dataset / region may yield inefficient results in another. In this sense, the selection of characteristics (conditioning factors) and an appropriate ML model for a target region is challenging. Some of the features identified in the literature are presented below.

[106] and [109] considered topographic slope (percent change in that elevation over a certain distance) and aspect (indication of the directions the physical slopes face) for fire predictions. Precipitation data on different temporal scales (annually, monthly, weekly) have been widely considered ([106], [109], [110], [111], [112]) and distance to river/lake/steam has also been taken into consideration in fire-related studies ([106], [109], [111], [115]).

Regarding climatic data, temperature has been taken as a factor that increases the quality of the predicted data ([109], [103], [113], [115]) and relative humidity [117] and wind direction and speed [109] and [115] have significantly impacted fire behaviour.

Anthropogenic variations such as distance to urbanized areas (road, cities, or settlements) have been reported as other fire

Table 4.3: Comparison of the Related Works

Proposal	Objective	Geographic region	Features (quantity : specification)	Models	Best results
[126]	Fire spread direction	Northern Alberta, Canada	06: temperature, wind speed and direction, land cover type, humidity, and intensity of rainfall	RL Advantage Actor-Critic (A3C)	AUC-ROC:0.93, Accuracy: 87.3%
[111]	Fire probability maps	Hyrcanian ecoregion, Northern Iran	02: rainfall and distance to road	Evidential belief function	AUC-ROC: 0.84
[106]	Susceptibility maps	Amol County, Northern Iran	10: altitude, aspect, slope, plan curvature, landform, topographic wetness index, radiation, wind speed and direction and NDVI	ANN, SVM, RF	AUC-ROC:0.88
[109]	Fire probability maps	Hyrcanian (Iran)	07: slope, aspect, temperature, precipitation, distance to river and wind direction and speed	Neuro-fuzzy inference system + genetic algorithm	Accuracy: 0.97, Sensitivity: 0.98
[110]	Fire probability maps	South Korea	05: land type, elevation, precipitation, population density, distance for urban area	Maxent, RF	AUC-ROC: 0.90
[115]	Susceptibility maps	Amol County, Northern Iran	10: altitude, slope, aspect, plan curvature, topographic wetness index, landform, radiation, wind speed and direction and NDVI	ANN, RF, SVM	AUC-ROC: 0.88
[117]	Fire probability	Brazil, Cerrado Biome	10: land cover, distance to road, electric lines and train lines, temperature, wind speed, relative humidity, elevation, aspect, and population density	RF	AUC-ROC: 0.72
[132]	Fire spread	Tasmania	03: relative humidity, temperature and wind speed	Surrogate model	Pearson's correlation coefficient: 0.90
[133]	Burned area prediction	United States	07: slope, aspect, vegetation, temperature, precipitations, humidity and population density	Convolutional autoencoders and RF	-
[127]	Fire direction	Rocky Mountains, United States	8: vegetation, digital elevation, atmospheric pressure, wind speed, wind direction, temperature dew point, relative humidity, and precipitation	2D CNN	Accuracy: 0.87, recall: 0.91
[128]	Fire perimeters and fuel adjustment factors	California, United States	2: fuel adjustment factors and fire perimeters positions	Radial basis ANN	Otsuka-Ochiai similarity: greater than 0.8
Our proposal	Fire spread direction prediction	Brazilian Cerrado Biome	16: slope, aspect, mean atmospheric pressure, maximum dry bulb air temperature, minimum dry bulb air temperature, maximum temperature dew point, minimum temperature dew point, maximum relative humidity, mean relative humidity, wind direction, maximum wind gust speed, mean wind gust speed, distance to road, distance to urban area, NDVI, vapor pressure deficit	ANN, SVM, RF, AdaBoost	AUC-ROC: 0.92, F1 Score: 0.87, accuracy: 0.88, precision: 0.88, recall:0.89, Sørensens Similarity: 0.83

behaviour conditioning factors ([106], [110], [111]). According to the authors, in many cases, fires are closely related to human behaviour and commercial practices. For example, humans exert a direct impact on vegetation's characteristic, hence, on the type of fuel. An index that helps the representation of those vegetation variations is the NDVI, considered by many authors, including [113], [106], and [25].

Table 4.3 shows a summary of the aforementioned studies and that several ML models have been used for fire prediction purposes and different features sets can be considered. However, their reproduction becomes a complex task due to the lack of a same kind of data for the FD region. The following sections address the way we deal with such challenges through the collection and pre-processing of a richest dataset (involving a higher number of features - 16), and through model selection, training, and tuning for the prediction of burned areas.

4.2 MATERIALS

The Federal District Region, with an area of 5,802 km^2, is located in the Center-West of Brazil, and its capital (Brasilia), is the fourth most populous city of the country. The Federal District is inserted in the Cerrado, the richest worldwide savanna and a large South American biome, which has gained special attention for having been affected by a large number of fires.

Several actors determine the behavior of forest fires in the region [118], and a review was conducted towards the understanding the ecology of fires. The variables were identified for the processing and assembly of a dataset for training ML models in fire behavior predictions. Only the features available in public online data sources were considered.

Topography and terrain characteristics (specifically slope and aspect) have shown relevant in the progress of fires by influencing vegetation [138], hence the combustible material. Slope (percent change in the elevation over a certain distance) and aspect (orientation of the earth's surface with respect to the sun) was considered by [25] and significantly impacted the fires occurrences in the FD region. Therefore, slope and aspect data provided by TOPODATA project from [139], which is based on the Shuttle Radar Topography Mission (SRTM), were taken into account in our study. SRTM followed the interferometric synthetic aperture radar method [140].

Regarding hydrography, the Cerrado has well-defined rainy (October-March) and dry (April-September) seasons. During the wet season, fires occur naturally through lightning and are less severe than those in the dry season, since their spread is inhibited by the moisture content of the soil [118]. During the dry season, they occur mainly due to human activity and are more severe. Soil moisture is highly correlated with rains and water sources. Distance to rivers was also considered as one of some soil moisture conditioning factors and calculated as the minimum geodesic distance from the center of each sector to any of the rivers of the hydrographic map (Figure 4.2) available in [2]. Sectors whose region is 100% covered by water (i.e. Paranoa Lake) were disregarded. Total precipitation (amount of rain for a sector in a 16-day resolution) was considered as a fire explanatory variable.

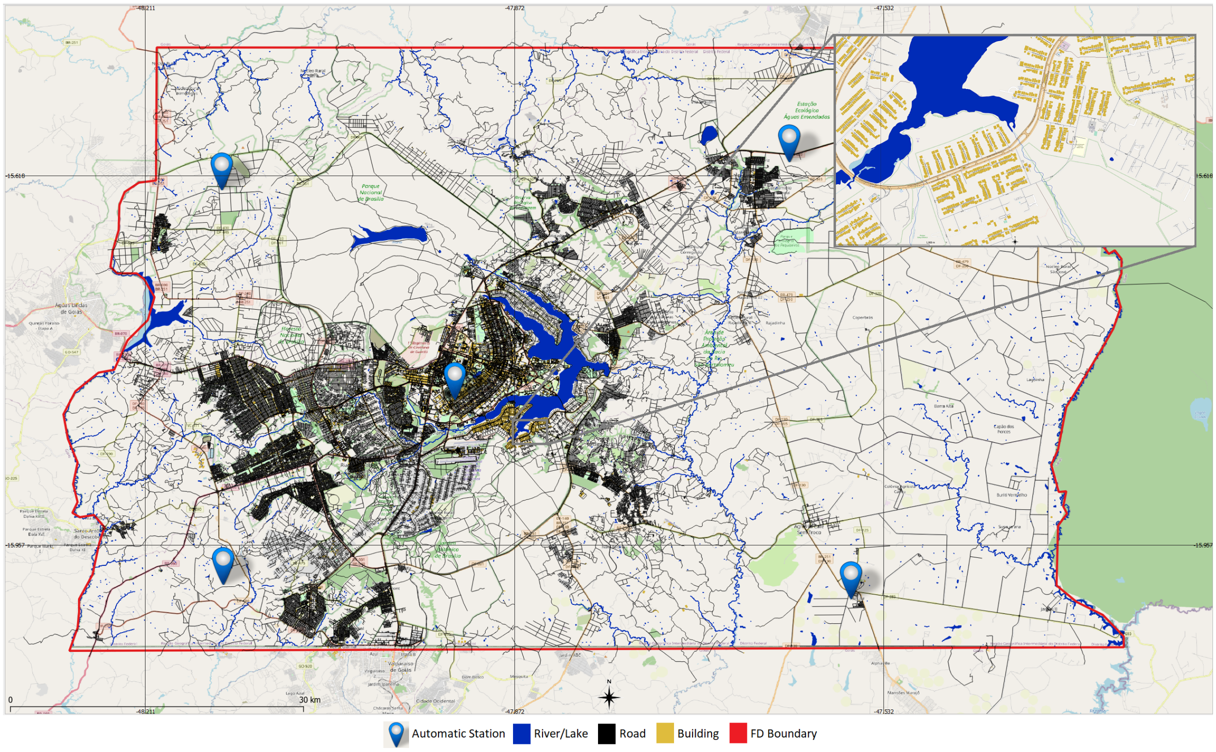


Figure 4.2: Federal District Roads, Rivers, areas with High Concentrations of Buildings, and Automatic Stations [5]

Climatic conditions also establish a close relationship with fire behavior in the Cerrado [141], and temperature, relative humidity, and wind direction and speed have showed some of the main factors of fire propagation. Data on 11 climatic features were obtained from five automatic stations (Figure 4.2) spread across the region [96]. Atmospheric Pressure, Dry Bulb Air Temperature, Temperature Dew Point, Relative Humidity, Wind Direction degrees, and Wind Gust Speed data were available at one-hour intervals.

Vapor Pressure Saturation (VPS) and Vapor Pressure Deficit (VPD) are other measure highly correlated to fire phenomena. [142] showed the relation of vapor pressure deficit with the occurrence of fires and highlighted a high VPD causes a fire to spread to a larger area. We followed the formal definition provided by [118] and showed below:

$$VPS = 610.7 * 10^{7.5T/(237.3+T)} \tag{4.1}$$

$$VPD = [1 - (RH/100)] * VPS \tag{4.2}$$

where T and RH denote, respectively, temperature and relative humidity.

The predominant savanna vegetation type of the Cerrado favors the spread of fire, since grasses increase flammability and produce large amounts of fine fuel during dry periods [118]. Therefore, we considered MODIS /Terra Vegetation Indices provided by [97] and particularly the NDVI, a

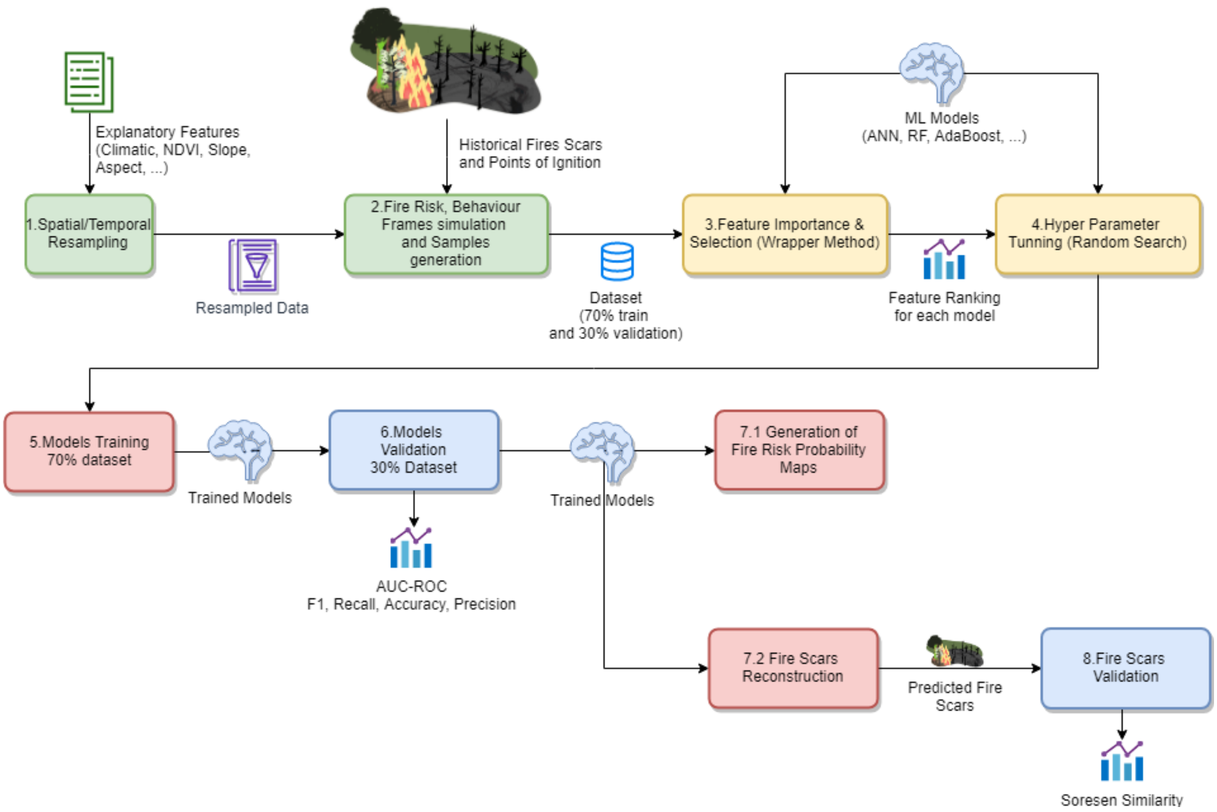

Explanatory Features
(Climatic, NDVI, Slope, Aspect, ...)
Historical Fires Scars and Points of Ignition
ML Models
(ANN, RF, AdaBoost, ...)
1.Spatial/Temporal Resampling
Resampled Data
2.Fire Risk, Behaviour Frames simulation and Samples generation
Dataset
(70% train and 30% validation)
3.Feature Importance & Selection (Wrapper Method)
Feature Ranking for each model
4.Hyper Parameter Tunning (Random Search)
5.Models Training 70% dataset
Trained Models
6.Models Validation 30% Dataset
AUC-ROC
F1, Recall, Accuracy, Precision
Trained Models
7.1 Generation of Fire Risk Probability Maps
7.2 Fire Scars Reconstruction
Predicted Fire Scars
8.Fire Scars Validation
Soresen Similarity

Figure 4.3: Methodology

reliable descriptor of fine fuel load calculated from spectral reflectance values as the ratio of the difference of near infrared and red reflectance to their sum.

Human presence also has a close relationship with fire occurrence in the Cerrado [143]. Fire occurrences and its behavior are affected by anthropogenic factors a considerable number of times. Distance to the nearest urban area, computed as the minimum distance to building concentration areas identified by the OpenStreet Maps API (Figure 4.2) represented the human presence. Moreover, distance to the nearest road was computed as the geodesic distance to the nearest road. Road geodata (Figure 4.2) were provided by [2].

4.2.1 Methods

Figure 4.3 shows the stages of the proposed methodology. Initially data are resampled for temporal and spatial normalization. Ignition points data are used to prepare a dataset for fire risk prediction and fire advance frames are then simulated according to historical data on ignition points and the burned area, once the fire has been mitigated for behaviour prediction. The datasets generated enabled the continuation of analyses of importance and selection of variables for mitigating multicorrelation problems and achieving better quality training of the models. The hyperparameters of the models are tuned according to the variables selected in each case. The next stages involve the training and validation of the models in the prediction of fire risk and behaviour. Finally, fire risk probability maps and fire scars are predicted and their similarity with

Table 4.4: Data considered

Type	**Feature**	**Spatial Resolution** original/re-sampled	**Temporal Resolution** original/re-sampled
Topographic	Slope and Aspect	30m/100m	constant
Hydrographic	Distance to River (DtRiver) Total Precipitation (TP) and	- coordinate/100m	constant Hourly/16 days
Climatic	Mean Atmospheric Pressure (MAP), Maximum Dry Bulb Air Temperature (MaxDBAT), Minimum Dry Bulb Air Temperature (MinDBAT), Maximum Temperature Dew Point (MaxTDP), Minimum Temperature Dew Point (MinTDP), Maximum Relative Humidity (MaxRH), Mean Relative Humidity (MeanRH), Wind Direction (WD), Maximum Wind Gust Speed (MaxWGS) and Mean Wind Gust Speed (MeanWGS)	coordinate/100m	Hourly/16 days
Anthropogenic	Distance to Road (DtRoad) and Distance to Urban Area (DtUA)	-	constant
Vegetation	NDVI	250m/100m	16 days/16 days
Other	VPD	100m/ 100m	Hourly/16 days
Fire	Spots (Points of ignition) and Scars	coordinate/100m 30m/100m	Hourly/16 days 16 days/16 days

the historical evidence is validated. The stages are detailed in what follows.

4.2.1.1 Spatial-temporal re-sampling

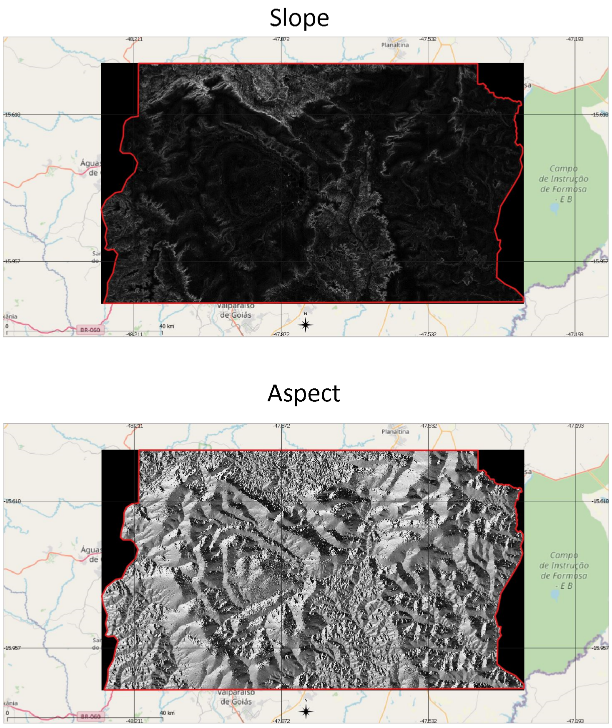
Table 4.4 shows the features considered and their respective sampling temporal and spatial resolution. Since data are in different spatial and temporal resolutions, a re-sampling process is required for their normalization and generation of the training dataset.

Topography data from a 30 m resolution were clipped to the Federal District Region (Figure 4.4) and re-sampled to our 100 m resolution grid. Given a 100-m sector x and an S_{30} set composed of slope and aspect sectors in a 30-m resolution that intersects x, the following equations computed $Slope(x)$ and $Aspect(x)$:

$$Slope_{100}(x) = \frac{\sum_{y \in S_{30}} Slope_{30}(y)}{|S_{30}|} \tag{4.3}$$

$$Aspect_{100}(x) = Mode(Aspect_{30}(S_{30})) \tag{4.4}$$

where $Mode(S)$ denotes the statistic mode of a set S (element with largest number of occurrences).


Slope

Aspect

Figure 4.4: Federal District Slope and Aspect

Regarding hydrographic data, distance to rivers was computed as the geodesic distance from the center of each sector to a lake or river. On the other hand, total precipitation data collected by the five automatic stations were interpolated as in the the climatic variables explained in what follows.

Climatic related data and VPD were spatially interpolated by Kriging Interpolation method [144],leading to a group of climatic features maps defined in a hourly basis (i.e. Figure 4.5, relative humidity for February 1, 2019 at 10:00 am). Kriging method assumes the distance and direction between the sample points reflect a spatial correlation that can explain the variation in each variable on the surface. It fits a mathematical function to a specific number of points (center of the sectors) and determines the output value for each location.

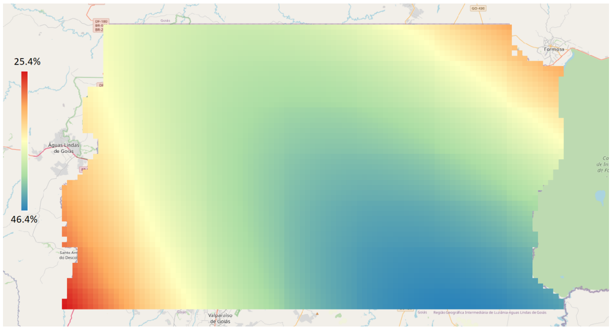


Figure 4.5: Example of relative humidity interpolated by Kriging method (February 1, 2019 at 10:00 am).

The calculation of distances to road and urban areas was similar to that of distance to rivers, which considered the geodesic distance from the center of each sector.

NDVI, sampled in 250-meter spatial resolution, was re-sampled according to the average value of intersected areas between sectors and samples (Figure **??**).

Temporal series were adopted for the modeling of explanatory features. Slope, aspect, or distance to road/rivers/urban areas, which can be considered invariant over time, were represented as constant time series, whereas temporal series of distinct sampling rates denoted the climatic ones. The lower format (16 days) was chosen for temporal resolution standardization, and three methods (minimum, maximum, and mean) were applied for the implementation. We followed a widely adopted approach [145], [146] and [147], according to which the inputs of the algorithm are a time series S, a windows size ws, and an initial date sd. Such data are then divided into windows of ws size (first windows start at sd) and the maximum, minimum, and mean aggregations are calculated for each window. Mean down-sampling (i.e. Temperature Figure 4.6) yielded better results than the other two metrics.

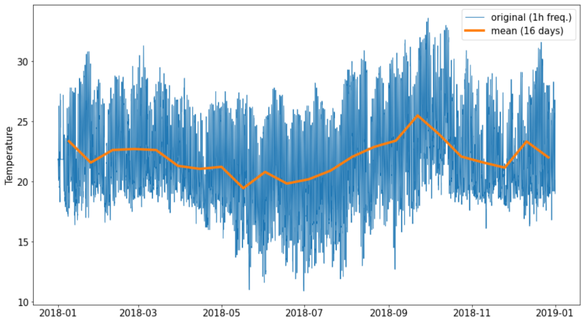


Figure 4.6: Down-sampling of temperature at an Automatic Station sector (2018).

4.2.1.2 Frames simulation and samples generation.

Two datasets from [1] provided fire historical data. The first set is devoted to historic fire spots; samples are defined by their latitude, longitude, and events (date and time), acquired mainly by a MODIS sensor from NOAAs-15, 16, 17, 18, and 19, AQUA, METEOSAT-02, TERRA and GOES-12 satellites.

The second dataset, is comprised of data on fire scars generated with 30-meter and 16-day spatial and temporal resolutions, respectively. Scars were generated by the processing of CBERS-4/MUX, Landsat-8/OLI, and Resourcesat/LISS images, and those up-sampled to a 100 m spatial resolution were considered. Therefore, given a 30m-sector fire scar set S_{30} and a 100m-sector Federal District set FD , the following rule calculates re-sampled scar S_{100}:

$$x \in S_{100} \leftrightarrow x \in FD \text{ and } \exists\, y \in S_{30} \;:\; x \cap y \neq 0 \tag{4.5}$$

where operator $\cap$ represents the common part between sectors x and y.

A 100m sector intersecting a 30m sector of the original scar belongs to the re-sampled scar,

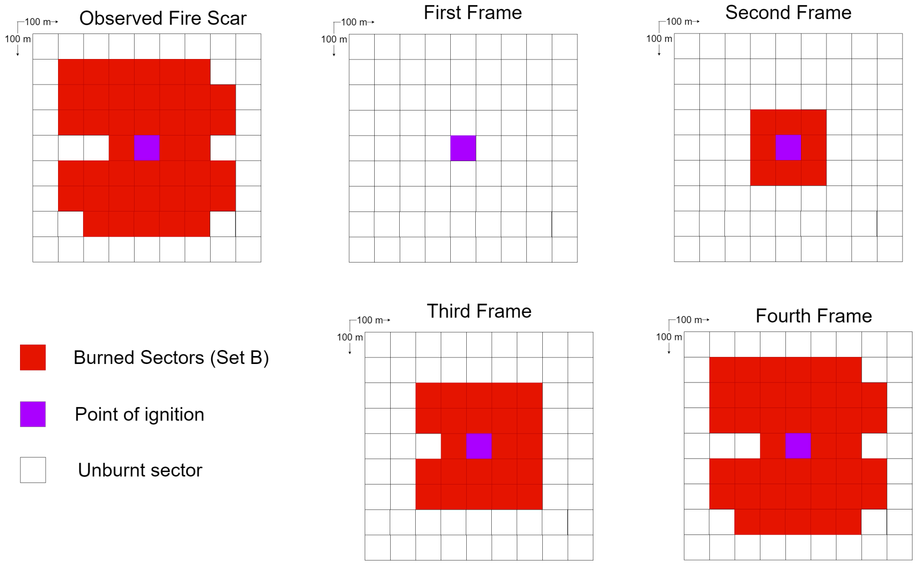


Figure 4.7: Fire progress simulation (Frames) [5]

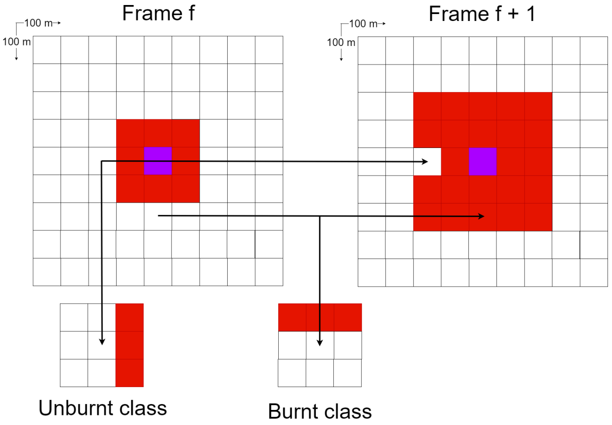


Figure 4.8: Sample Generation Process

and the ignition point was provided by the crossing between the two datasets.

Fire scars data were used in the simulation of fire progression. Scars were available in 16-day temporal resolution, thus hampering the obtaining of all frames. Only the final one was acquired through the simulation of fire progression. We follow an approach that considers fires to progress only for sectors neighboring a burned sector. Starting from the ignition sector (first frame), we iteratively generate frames where, in each iteration, neighbors burned in the scar are added to the next frame, and so on (see Figure 4.7).

Figure 4.8 displays our dataset comprised of samples of 3x3 sectors obtained from two consecutive frames (f, f+1). If a sector is not affected by fire at frame f and has at least one burnt neighbor and it is burnt in f+1, then the ”burn” class represents it; otherwise, the unburnt class is assigned.

Finally, 2500 samples of burnt class and 2500 samples of unburnt one were generated for training purposes.Figure 4.9 illustrates the format of a record in our dataset. Each neighboring

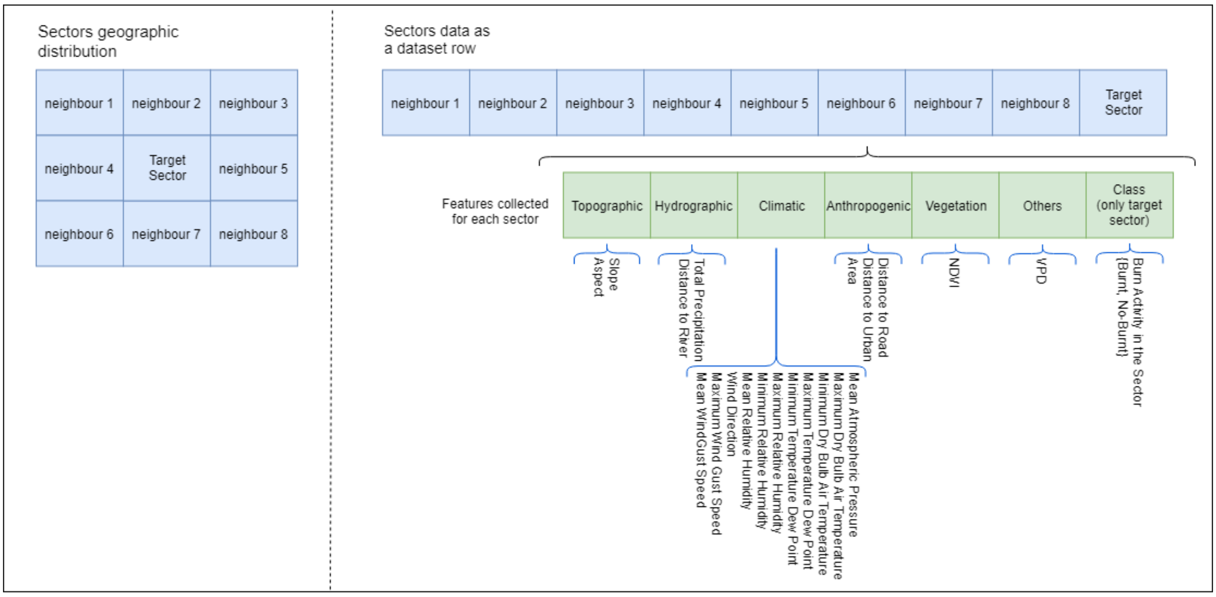


Figure 4.9: Record Composition [5]

sector was provided with topographic, hydrographic, climatic, anthropogenic, vegetation, and other features and the class that represents fire in the frame. Class for the target sector will be estimated; therefore, it has been omitted.

4.2.1.3 Feature Importance and Selection

The feature importance and selection process is a crucial phase to improve the training of machine learning models and to produce better predictors. Inconsistency, noise, missing data or even a small amount of data are handled in this phase by the identification of the features that better describe the variable (class) to be predicted [148].

Feature importance measure how each explanatory variable correlates with the fire behaviour. A higher score means that the specific feature has higher effect on the modelling for prediction of the dependent variable.

A set of z ML models and a set of n features from the dataset (in our case, z=4 and n=24) were considered. According to Figure 4.10, a permutation method evaluated the feature importance, taking into account the accuracy of the ML model. Feature importance calculation by permutation method [149] considers the training of the ML models in the entire dataset for computing a reference score (i.e. a validation metric such as accuracy). A single feature is then permuted multiple times and the model is trained again to producing a corrupted score. The difference between the reference score and the average corrupted scores becomes the importance of that feature. The process is applied to each feature, and a higher score means the specific feature exerts a stronger effect on the modelling for the prediction of the dependent variable. Scores are used for the establishment of a ranking to be used by a feature selection technique, as explained below.

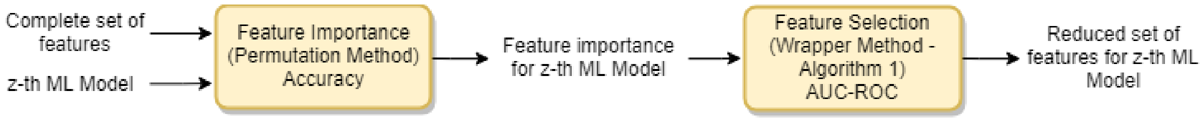

Complete set of features
z-th ML Model
Feature Importance (Permutation Method) Accuracy
Feature importance for z-th ML Model
Feature Selection (Wrapper Method - Algorithm 1) AUC-ROC
Reduced set of features for z-th ML Model

Figure 4.10: Feature importance and selection flow.

Feature selection techniques takes a dataset with n features as input and returns a reduced one with m features where m $<=$ n. Among the several approaches proposed for feature selection [150], filter and wrapper methods, described in what follows, were considered in this research.

Filter methods

The filtering methods [151] make the selection independently of any machine learning algorithms and features are selected on the basis of an score provided by statistical tests for their correlation with the outcome variable.

Chi-square test [152] showed the relation between each explanatory feature and the fire behaviour. It reveals whether a variable (i.e. temperature, relative humidity, NDVI, among others) describes the predicted phenomenon (fire behaviour) by interpreting the p-scores. A higher chi-square p-score denotes poor relation and the one near 0 means the explanatory feature can describe the wildfire behaviour. Therefore, variables with a p-score higher than a threshold can be dropped from the dataset. Since Chi-square can be applied only to categorical data, all continuous features were normalized and converted into five categories, namely "very-low", "low", "normal", "high", and "very-high".

Relief-F [153] was another filter method applied to the dataset. It produces a score for each feature enabling the selection of those with highest rank in the feature selection process. The score is based on the difference between the values of a feature in pairs of nearest neighbor samples. If a feature difference is observed in a pair of neighboring samples of a same class, the feature score decreases. Alternatively, if a feature difference is observed in a pair of neighboring samples of a different class, the resource score increases.

Wrapper methods are based on performance of machine learning models (wrapping it). It follows a search approach to evaluate combinations of features based on a evaluation criterion and with for each feature subset the model is trained and validated [150]. Once the search algorithm ends, the best feature subset is considered for the training of the model. They are computationally more expensive compared to filter methods due to the repeated learning steps and cross-validation. However, we considered the wrapper methods class due to the need for optimization of the models' performance and for solving some of the features dependency problems, as indicated by ([154], [150]) .

The search for a feature subset that maximizes the accuracy of each model is a non polynomial optimization problem that involves the testing of all subsets of explanatory features, i.e., $\mathcal{O}(2^n)$, where n denotes number of features. In this case, there are 2^{24} possible subsets and the testing of all cases is impracticable. This research followed the semi-greedy strategy described in Algorithm 1.

Algorithm 1 Feature selection for one model

```
procedure SELECT_FEATURES($model, features : \{< feature, importance >\}$)
    $features.sortByImportanceDesc()$
    $i \leftarrow 0$
    $result \leftarrow \{\}$
    $bestAUCROC \leftarrow 0$
    while i < length(features) do
        $temp\_features \leftarrow \{results\}$
        $tempFeatures.append(features.first())$
        $features.deleteFirst()$
        $tempAUCROC = model.train(tempFeatures).validate()$
        if $tempAUCROC > bestAUCROC$ then
            $bestAUCROC \leftarrow tempAUCROC$
            $result \leftarrow tempFeatures$
        end if
        $i \leftarrow i + 1$
    end while
    return $result$
end procedure
```

The algorithm takes a model and a set with all features and their importance for that model (computed by the permutation method) as input. The features are sorted according to their importance and a result set containing no features is created. The algorithm iterates over the sorted features and the model is trained in each step according to the result set plus the feature of highest importance in the feature set. If the AUC-ROC obtained is better than the best until that moment, the feature is included in the result set and discarded in the other case.

However, features with high importance for one model are not necessarily relevant for another, since there may be a high correlation between the variables.

Those are some of the most promising models in fire behavior prediction, according to the literature on fire behaviour/spread domain [134].

4.2.1.4 Hyper-parameter optimization.

Hyper-parameters are model parameters that must be defined prior to model training. Several different techniques aim to optimize them, resulting in better accuracy of their model. Random search optimization technique [155], which defines a bounded search space and randomly samples points with a specific value for each hyper-parameter, was considered in this study. The goal was to find a vector that results in the best performance of the model after learning (e.g., maximum accuracy or minimum error).

4.2.1.5 Models training and validation

The wildfire dataset was randomly shuffled and splitted in 70% for training and 30% for validation purposes. The performance of the machine learning models was evaluated by k-fold cross-validation, for the estimation of the model's adequacy regarding unknown data. Parameter k (k = 4) denotes number of groups of the dataset.

Both Receiver Operating Characteristic Curve and AUC-ROC were obtained for each model and fold. ROC displays the True Positive Rate (TPR) against the False Positive Rate (FPR)) at several threshold settings and has been used for comparisons of ML models in a same task. An AUC equal to or near one indicates very good or good performance of a model.

The models were ranked according to their performance and AUC-ROC, F1 score, accuracy, and recall validated them as defined in [156]. Data predicted by the ML models enabled the construction of various fire scars, and the precision of the predicted burned areas is analyzed.

4.2.1.6 Fire scars reconstruction and validation

This phase starts from a given point of ignition and estimates the resulting scar according to the spread/advance of the fire. Burnt area can be computed by Algorithm 2.

The algorithm takes a trained model and the ignition sector as input, and $previous\_cardinality$ and $current\_cardinality$ variables represent the cardinality of the $burned$ set in two consecutive iterations of the while loop (lines 7 to 24). Such variables are adopted for evaluating whether the $burned$ set size has changed from one iteration to another and are used as a stopping condition. Initially, $previous\_cardinality = 0$, since there is no previous burned set, and $current\_cardinality = 1$, since the actual burned set contains only the ignition sector; $cutoff$ is an internal variable for the determination of when the model classifies a sector as burned or not.

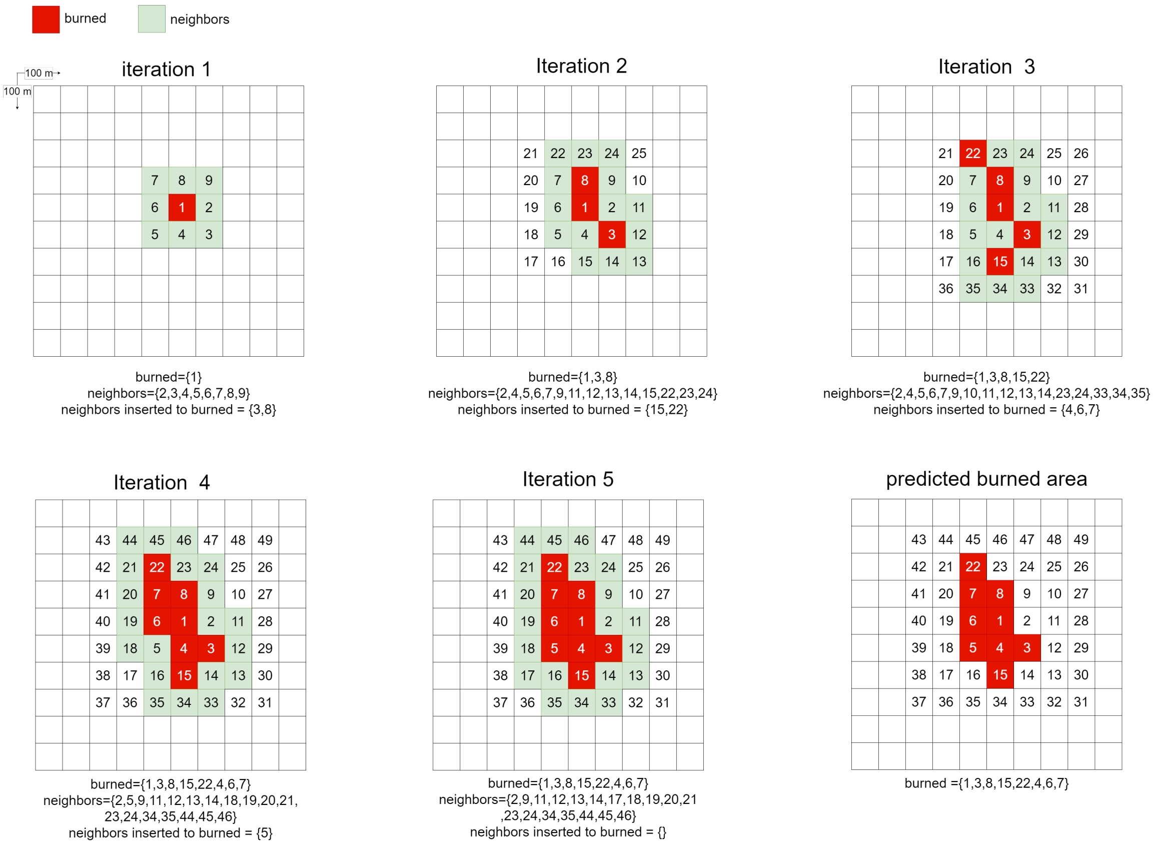
The algorithm starts from an empty neighbor set for each while iteration and, for each sector predicted as burnt, it progressively verifies among its neighbors with respect to additional sectors the possibility of its individual inclusion in the $burned$ set, naturally excluding neighbors previously included (lines 8-15). Such sectors are candidates that will probably burn due to the spread of the fire from the burned sectors, according to the evaluation of the model - if the result is "burn", the algorithm includes it in the $burned$ set (lines 16-20).

Cardinalities are updated (lines 21-22) and the next iteration starts. The while loop stops when previous and current cardinalities are equal, i.e., when no more neighboring sectors have been included in the $burned$ set, and returns the $burned$ set that represents the predicted burned area.

Figure 4.11 shows a running example where neighbors three and eight in iteration one are inserted in the burned set and no more neighbors are classified as burned in iteration five. The algorithm then stops.

Algorithm 2 Wildfire burned area predictor

```
procedure PREDICT_AREA($model, ignitionSector$)
    $burned \leftarrow \{\}$
    $previous\_cardinality \leftarrow \|burned\|$ //0 since burned is an empty set
    $burned \leftarrow \{ignitionSector\}$
    $current\_cardinality \leftarrow \|burned\|$ //1 since burned set only contains the ignition sector
    $cutoff \leftarrow 0.5$
    while $previous\_cardinality \neq current\_cardinality$ do
        $neighbors \leftarrow \{\}$
        for $b \in burned$ do
            for n ∈ b.neighbors() do
                if $n \notin neighbors$ && $n \notin burned$ then
                    $neighbors \leftarrow neighbors \cup \{n\}$
                end if
            end for
        end for
        for $n \in neighbors$ do
            if $model.Evaluate(n) > cutoff$ then
                $burned \leftarrow burned \cup \{n\}$ //According to the evaluation, the sector is added to burned set
            end if
        end for
        $previous\_cardinality \leftarrow current\_cardinality$ //Updating previous cardinality for the next iteration
        $current\_cardinality \leftarrow \|burned\|$ //Updating current cardinality for the next iteration
    end while
    return $burned$ //returns the predicted area
end procedure
```



burned={1}
neighbors={2,3,4,5,6,7,8,9}
neighbors inserted to burned = {3,8}

burned={1,3,8}
neighbors={2,4,5,6,7,9,11,12,13,14,15,22,23,24}
neighbors inserted to burned = {15,22}

burned={1,3,8,15,22}
neighbors={2,4,5,6,7,9,10,11,12,13,14,23,24,33,34,35}
neighbors inserted to burned = {4,6,7}

burned={1,3,8,15,22,4,6,7}
neighbors={2,5,9,11,12,13,14,18,19,20,21,
23,24,34,35,44,45,46}
neighbors inserted to burned = {5}

burned={1,3,8,15,22,4,6,7}
neighbors={2,9,11,12,13,14,17,18,19,20,21
,23,24,34,35,44,45,46}
neighbors inserted to burned = {}

burned ={1,3,8,15,22,4,6,7}

Figure 4.11: Process of construction of the predicted burned area.

Comparisons of observed and predicted burned areas must be made with the use of appropriate metrics for the topic [137]. Equation 6 shows the SS metric, where I is the area of intersection of predicted and known scars, P represents the predicted scar area, and K denotes the known scar area. Results close to one mean better precision, and those close to 0 denote lower quality predictions. Several investigations have considered SS ([157], [158] and [159]).

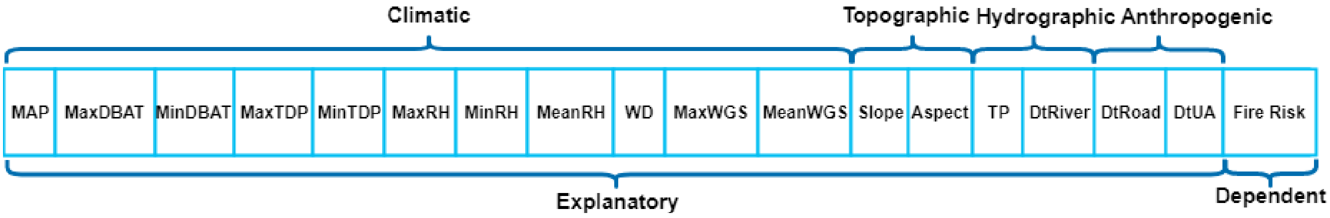
$$SS = 2 * I/(P + K) \tag{4.6}$$

4.3 EXPERIMENTS AND RESULTS ON WILDFIRE RISK PREDICTION

Initially, we generated the wildfire inventory dataset based on the features described in the previous section. Our inventory is composed of 6089 "high-risk" samples and 6089 "low-risk" ones. The dataset was randomly shuffled for avoiding bias in the classification process.

Figure 4.12 shows the format of our inventory records, in which the explanatory features values were calculated for a 1-km sector in a 1-hour time span and the fire risk categorical value was set to "high risk" for one or more fire occurrences and "low risk" for the other cases.

The best results for each model were obtained with the parameters described in Table 4.5,



MAP	MaxDBAT	MinDBAT	MaxTDP	MinTDP	MaxRH	MinRH	MeanRH	WD	MaxWGS	MeanWGS	Slope	Aspect	TP	DtRiver	DtRoad	DtUA	Fire Risk

Figure 4.12: Dataset register for a sector in a 1-hour time span [5]

which were adjusted by parameters optimization techniques.

Figure 4.13 shows the curves plotted for each fold by the ROC method and the trade-off between false positive and true positive rates on X and Y axes, respectively. AUC was used as a validation metric, according to which values close to one indicate better performance. Relevant results were achieved for the eight ML models; AdaBoost showed the best performance, with a 0.993 average AUC, and RF and ANN showed significant performance.

However, in terms of computational costs involved in the training phase, ANN, RF, and AdaBoost proved more efficient solutions. RF takes approximately two seconds for training, and AdaBoost and ANN require around 15 seconds and 13 minutes, respectively.

Figure 4.14 shows the values of AUC-ROC, accuracy, precision and recall metrics, according to the above-mentioned 70-30 rule. As expected (1-fold), AdaBoost showed the best AUC-ROC; the behaviour of the other models was quite similar. In terms of accuracy, AdaBoost and RF provided the best values, i.e., 96% and 95%, respectively, and the same results for precision and recall. Recall was very low for ANN; therefore, the classifier produced a high number of false negatives.

Figure 4.15 displays the Feature Importance (FI) for each model. For tree-based models (such as RF), it was obtained by Gini factor [160], which considered the relative importance of each decision tree node. The FI calculation for the other models was based on the feature permutation method [149], which extracts statistics on how much a mean absolute error will vary with respect to random changes in the feature value.

NDVI, atmospheric pressure, and relative humidity quantified the highest relation to a specific model. However, we cannot guarantee they are the main originating factor of wildfires in the Federal District. For example, RF was the second best model in terms of performance, and the FI values showed all features impacted the model.

The analysis of a single variable can be contradictory. For example, NDVI is of greater importance for models such as ADABoost and ANN, while its impact on others (e.g. LR and SVM) is relatively low or null. This fact corroborates the importance of considering several variables and models.

However, as shown in Figure 4.15, in general, the dependent features considered in this research positively contribute to the models, despite some exceptions such as the contribution of distance to road to LR and LogR models.

Table 4.5: Models Considered

Models	Settings
Deep Artificial Neural Network	Three hidden layers with 500, 100 and 50 neurons respectively, and Rectified Linear Unit (RELU) activation function. SoftMax output activation function. ADAM stochastic gradient descent method was chosen as optimizer. Categorical Cross Entropy for losses model. 0.001 Learning Rate Epochs number set to 1000. Batch Size set to 32
Support Vector Machine	RBF kernel. Kernel width (γ) of 0.001. Regularization parameter (C) of 1.0.
Random Forest	Sampling process trees set to 15. Number of variables for each split set to 4. Maximum number of trees set to 100. Voting threshold or cutoff set to 0.01.
Gaussian Naives Bayes	Prior probabilities of the classes adjusted from data.
K-Nearest Neighbors	K set to 20. Euclidean distance.
Linear Regression	—
Logistic Regression	Inverse of regularization strength $C = 1e^5$ Maximum number of iterations set to 500.
AdaBoost	Decision Trees as base estimators with max depth of 5. Number of estimators set to 1000. Learning Rate set to 1.

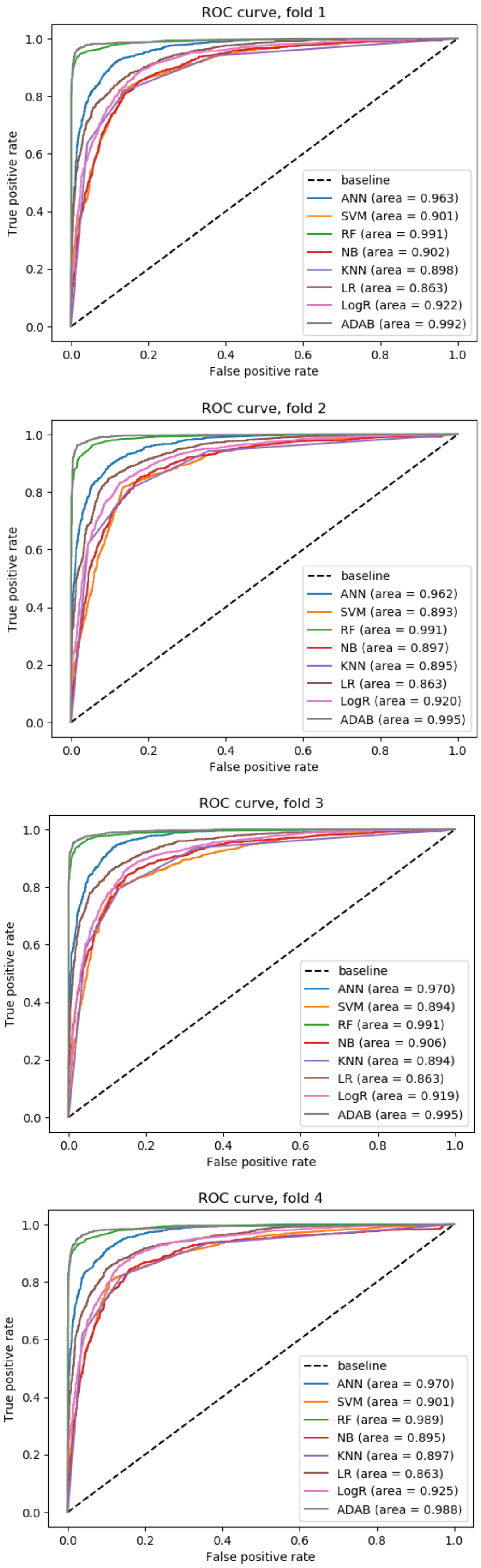


Figure 4.13: AUC ROC 4-fold for fire risk prediction

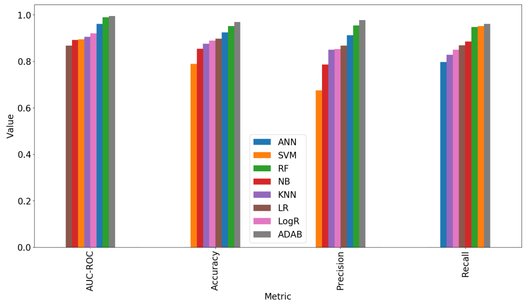


Figure 4.14: Validation metrics for fire risk prediction

In comparison to other studies involving the same area (De Bem et al. [25] and Galizia and Rodrigues et al. [117]), our dataset enabled the modelling and prediction of wildfires with significantly higher accuracy. AdaBoost (which was not considered in the previous studies) provided a higher than 0.2 AUC-ROC value, and, in contrast to De Bem et al. [25], we claim homogeneous and heterogeneous climatic features must be considered for a better wildfire risk prediction.

Finally, Figures 4.17 and 4.17 show the fire risk maps calculated for the 15th of January and 15th of August 2019. and generated according to ADA Boost model, which provided the best results. In both cases, sectors marked with red color were identified as those of high fire source risk, whereas those marked with green color denote low risk. High-risk sectors in urban areas are normally found in regions of parks or natural reserves.

Regarding number of sectors with high risk of fires, August 2019 shows a higher number of occurrences, as expected, if the frequency of fires displayed in Figure 2 is considered. Moreover, the climatic conditions of the month favor the occurrence of fires due to the low relative humidity, high temperatures, and existence of fuel such as dry leaves and grass.

Considering rainfall values, a relatively high number of high-risk sectors was reported in January 2019, which may have led to performance problems of prediction. From the total number of samples, approximately 28% occurred in the rainy season and 50% represented fire events, of which only 1272 samples out of 12000 were, in fact, ignition points.

The same validation metrics were applied for validating the effect of the fewer number of samples for the period, but considering only those from the validation set in the rainy period (October-April) (30% of the 1272 samples). The metrics revealed a poor performance in relation to the full dataset, as shown in Figure 4.18. Models were trained again, but only in the rainy period. As displayed in Figure 4.19, AdaBoost showed the best performance – near one - with all metrics and enabled the generation of other fire risk maps (see Figures 4.20 and 4.21).

Similarly to the training with data from the rainy period, we retrained the models for the dry season, with data related to only May-September period, for checking whether the models accuracy had increased. Figure 4.22 shows the AUC-ROC obtained. ADABoost performed better than the other models; however, the values of AUC-ROC validation were lower than the ones considering the entire dataset (figures 4.13 and 4.14).

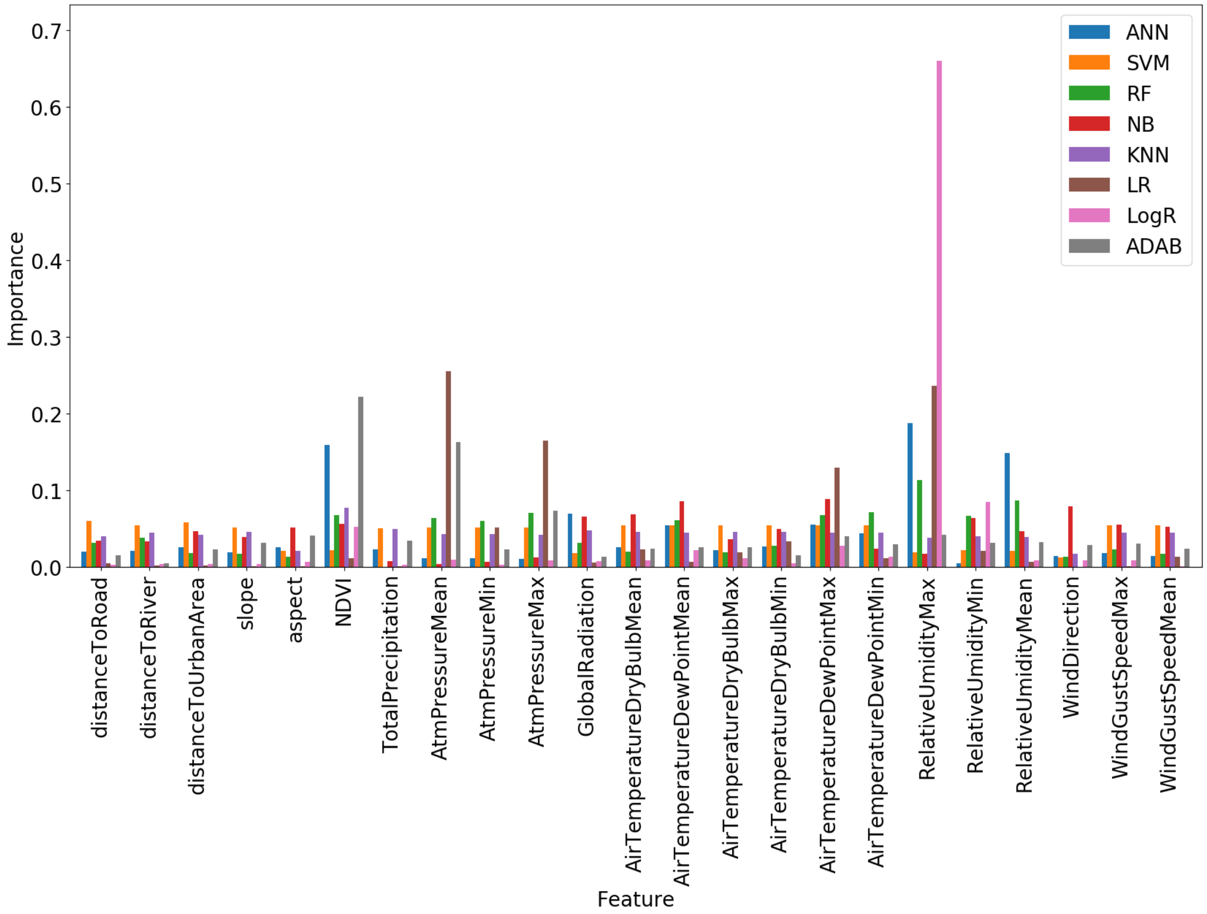


Figure 4.15: Feature importance for fire risk prediction

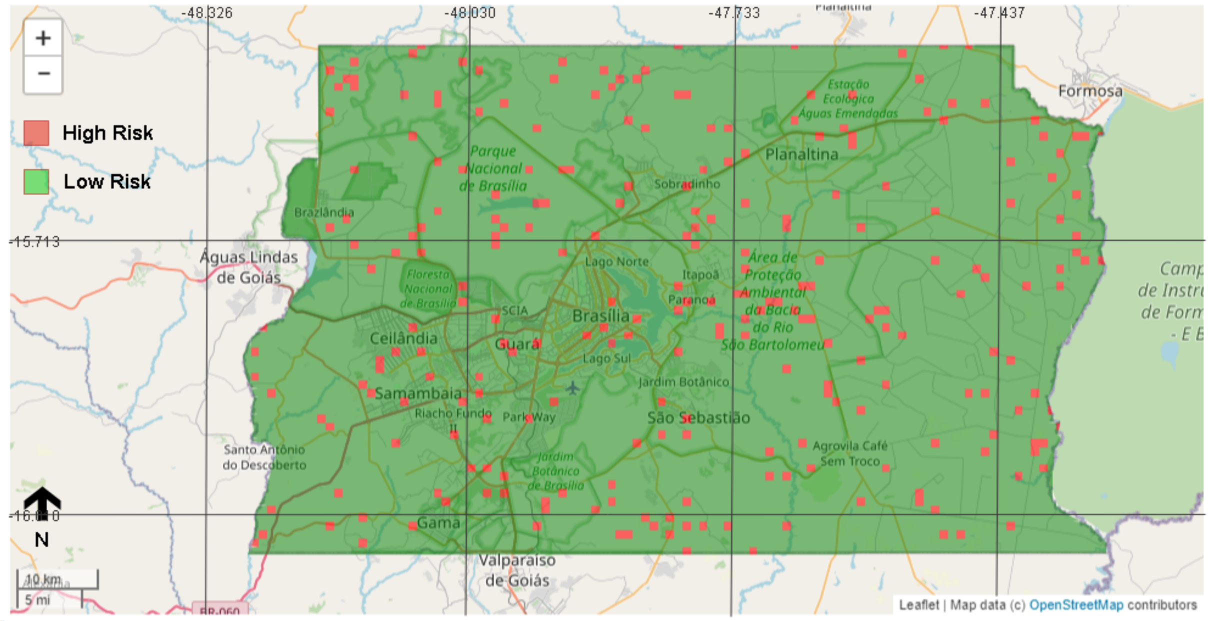


Figure 4.16: Fire risk prediction map January 15, 2019

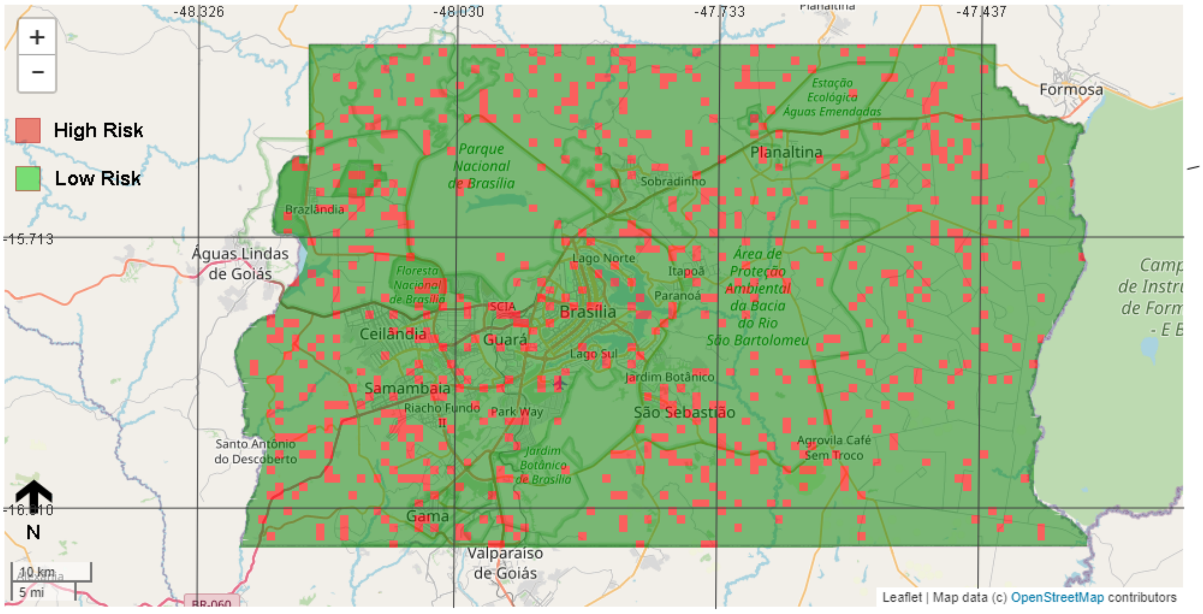


Figure 4.17: Fire risk prediction map August 15, 2019

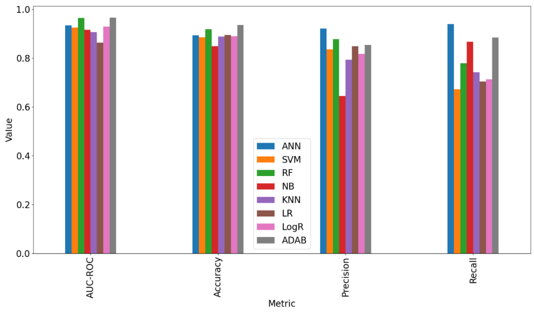


Figure 4.18: Validation metrics for rainy period

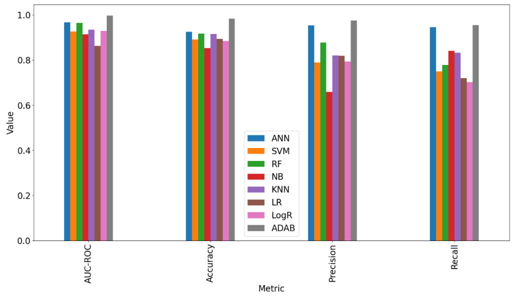


Figure 4.19: Validation metrics after training considering only the rainy period data.

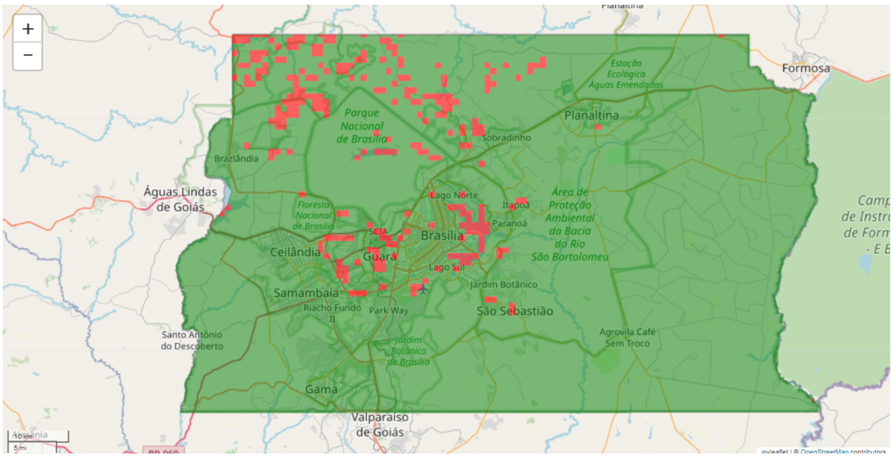

+
−
Planaltina
Formosa
Estação Ecológica Águas Emendadas
Parque Nacional de Brasília
Planaltina
Sobradinho
Brazlândia
Águas Lindas de Goiás
Floresta Nacional de Brasília
Lago Norte
Itapoã
Paranoá
Área de Proteção Ambiental da Bacia do Rio São Bartolomeu
SCIA
Brasília
Ceilândia
Guará
Lago Sul
Samambaia
Jardim Botânico
Riacho Fundo II
Park Way
São Sebastião
Santo Antônio do Descoberto
Jardim Botânico de Brasília
Agrovila Café Sem Troco
Gama
Valparaíso de Goiás
BR-060
ipyleaflet | © OpenStreetMap contributors

Figure 4.20: Risk map after retraining, January 15, 2019, 15:00

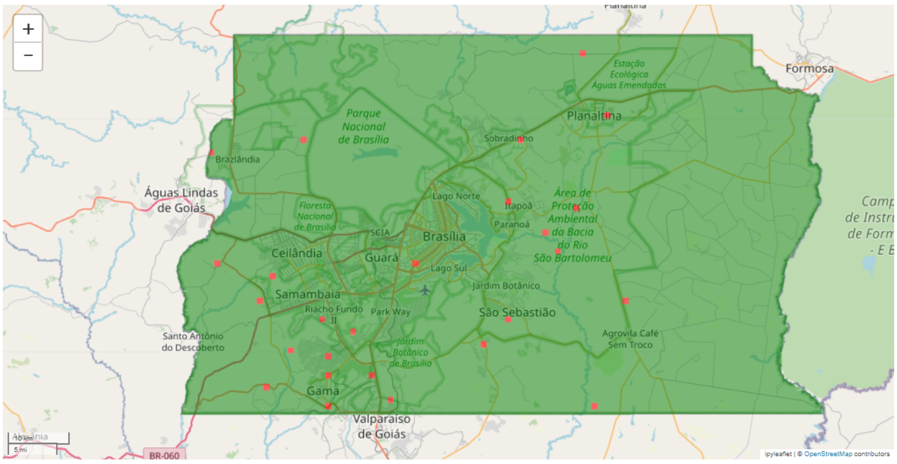

+
−
Planaltina
Formosa
Estação Ecológica Águas Emendadas
Parque Nacional de Brasília
Planaltina
Sobradinho
Brazlândia
Águas Lindas de Goiás
Floresta Nacional de Brasília
Lago Norte
Itapoã
Paranoá
Área de Proteção Ambiental da Bacia do Rio São Bartolomeu
SCIA
Brasília
Ceilândia
Guará
Lago Sul
Samambaia
Jardim Botânico
Riacho Fundo II
Park Way
São Sebastião
Santo Antônio do Descoberto
Jardim Botânico de Brasília
Agrovila Café Sem Troco
Gama
Valparaíso de Goiás
BR-060
ipyleaflet | © OpenStreetMap contributors

Figure 4.21: Risk map after retraining, January 15, 2019, 23:00

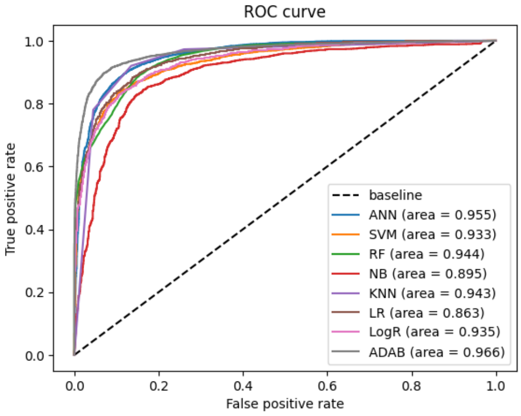


Figure 4.22: AUC-ROC considering only the dry period data.

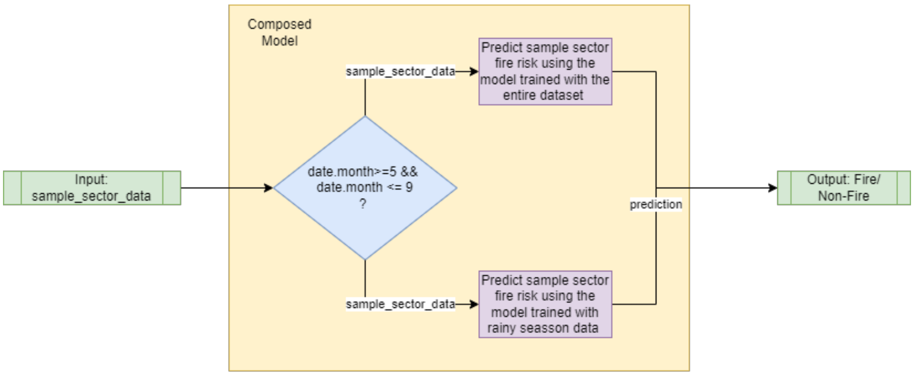

Composed Model
Input: sample_sector_data
date.month>=5 && date.month <= 9 ?
sample_sector_data
Predict sample sector fire risk using the model trained with the entire dataset
sample_sector_data
Predict sample sector fire risk using the model trained with rainy seasson data
prediction
Output: Fire/ Non-Fire

Figure 4.23: Composition of models.

Therefore, the two models were combined for predictions according to the flow in Figure 4.23. Given the data of a sector for which the risk of fire will be predicted, if the current month belongs to the rainy period, the model trained with data related to only that period is considered. Otherwise, the prediction is made with the model trained with the complete dataset, producing the results presented at figures 4.13 and 4.14.

4.4 EXPERIMENTS AND RESULTS ON WILDFIRE BEHAVIOUR PREDICTION

According to the defined methodology a dataset was defined and Deep Artificial Neural Network (ANN), SVM, RF, and AdaBoost were candidates for the prediction of the fire behaviour (treated as a binary classification problem). The models predict the time at which a sector will (or not) most probably burn, thus enabling the building of the area that will certainly burn.

4.4.1 Feature importance and selection study.

As explained in section 2.2.3, Chi-square test verified the dependency between features and fire behavior (Figure 4.24 shows the results - all p-values are below 0.2, thus reflecting the relation of all explanatory features and the behaviour of wildfires). Such a result enables the establishment of a ranking between the variables, according to which "Relative Humidity Mean" is the most relevant one and "Atmospheric Pressure Mean" is the first candidate to be discarded. However, since the p-scores were very homogeneous, no evidence was provided for the establishment of a feature selection.

On the other hand, Figure 4.25 displays the results of Relief-F, according to which higher values denote higher importance. The results have provided some insights into both occurrence and behavior of fires in the DF region. As an example, vegetation (NDVI) is the feature of highest influence, followed by variables related to wind and terrain topography.

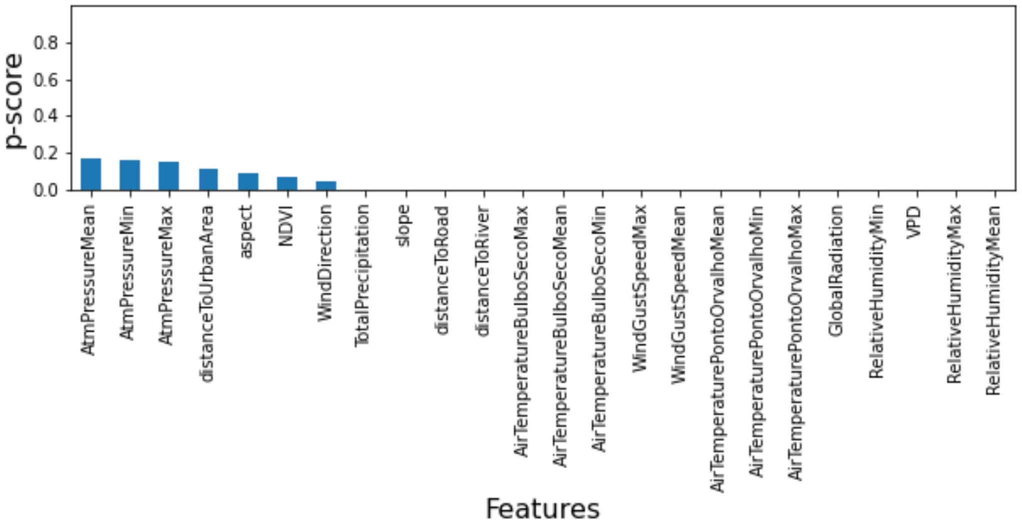


Figure 4.24: Chi-square test.

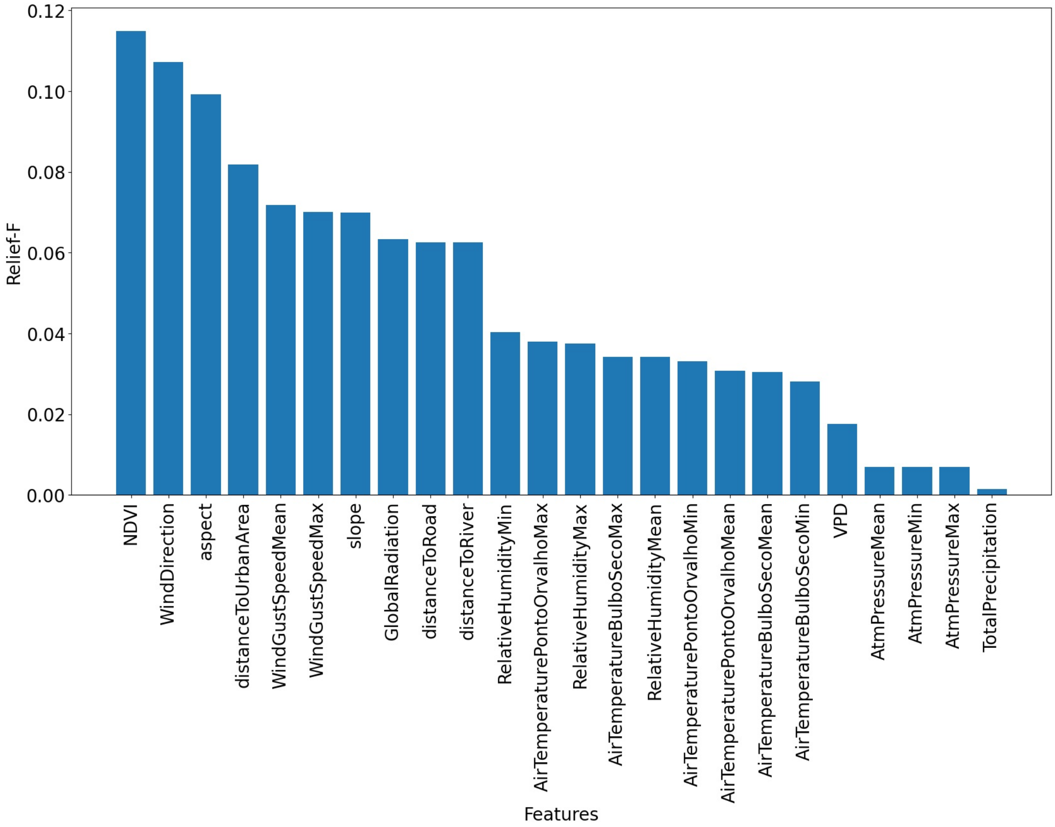


Figure 4.25: Relief-F.

The importance of the features was very similar (distributed in the [0.1, 0.2] interval) and all features showed some degree of importance. However, no cutoff importance value can be established for feature selection, since not enough information is provided. Furthermore, the joint analysis of Chi-square and Relief-F results is not conclusive. As an example, VPD showed fully correlated to the wildfire phenomenon in the Chi-square test, but values near 0.1 in Relief-F.

Therefore, both correlation and multicollinearity between the explanatory features were calculated to explain that effect.

Correlation describes the association between two variables and expresses a subject in terms of its relationship with the others [161]. Two correlated variables mean one of them can be predicted from the other, thus impacting the feature importance calculation. A proper correlation analysis can enable a better understanding of data. Pearson's correlation coefficient [162] was considered in this research – according to the coefficient, given a pair of sampled random variables (X,Y), the following equation calculates the correlation:

$$r_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{4.7}$$

where n is sample size, x_i and y_i are a sample point, and $\bar{x}$ and $\bar{y}$ are the sample mean of X and Y, respectively.

Figure 4.26 shows the results of Pearson's coefficient and a high correlation between the explanatory features. Several pairs (i.e. <NDVI, Relative humidity> and <Temperature, Global Radiation>) showed highly correlated, whereas variables such as distance to urban area presented a low correlation in respect the other variables. In addition to the pairwise correlations, a multi-correlation can be observed. As an example, pairs <NDVI, Relative humidity> and <Temperature, Relative Humidity> showed a high correlation; tupple <NDVI, Relative humidity, Temperature> may be multicorrelated or multicollinear.

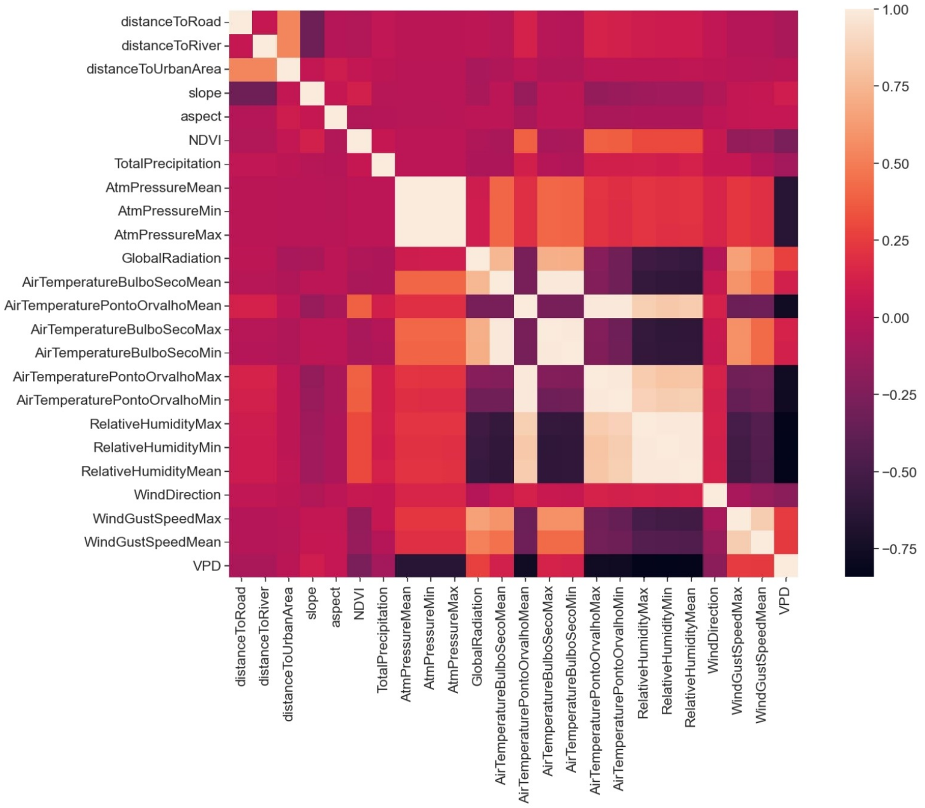


Figure 4.26: Features correlation

Multicollinearity occurs when a predictor variable can be linearly predicted from two or more variables with a high degree of accuracy, thus yielding misleading results. Variance Inflation Factor (VIF) is a metric commonly used to compute multicollinearity [163]. VIF was computed according to [161], who established values above 10 mean high multicollinearity between explanatory features. Its definition is shown in the equation

$$VIF(\hat{B}_j) = \frac{1}{1 - R^2_{X_j|X_{\_j}}} \tag{4.8}$$

where $\hat{B}_j$ denotes the coefficient of $j - th$ feature in a linear regression model and $R^2_{X_j|X_{\_j}}$ is the coefficient of determination of the regression equation of the first step, with X_j on the left side and the remaining predictive variables on the right one.

According to the VIF results (Table 4.6), multicollinearity is present in the data. Two or more correlated or multicollinear explanatory features describe the same phenomena and account for twice (or more) the computation of the feature importance; consequently, it is difficult to perceive which variable is really influencing the independent variable. Moreover, when features are correlated and collinear, the discard of one of them exerts a small effect on the models performance, since it can obtain the same information from a correlated feature. Although this is a common problem of filtering methods of have high computational efficiency, the selection produces incoherent validation metrics results in several scenarios when a same selected set of features is used

Table 4.6: Variance Inflation Factor

VPD	2.860730e+02
RelativeHumidityMin	7.937134e+01
RelativeHumidityMean	3.846107e+01
RelativeHumidityMax	3.063114e+01
AirTemperatureBulboSecoMax	2.890975e+01
AtmPressureMin	2.639599e+01
AirTemperatureBulboSecoMean	2.616578e+01
AtmPressureMean	2.414639e+01
AtmPressureMax	1.822316e+01
AirTemperatureBulboSecoMin	1.807455e+01
AirTemperaturePontoOrvalhoMax	1.494208e+01
AirTemperaturePontoOrvalhoMin	1.476668e+01
AirTemperaturePontoOrvalhoMean	1.252187e+01
GlobalRadiation	5.910211e+00
WindGustSpeedMax	5.447082e+00
WindGustSpeedMean	4.297431e+00
slope	1.828753e+00
distanceToRiver	1.738201e+00
distanceToRoad	1.720456e+00
distanceToUrbanArea	1.582141e+00
NDVI	1.282130e+00
TotalPrecipitation	1.129459e+00
WindDirection	1.084976e+00
aspect	1.023021e+00

to train different machine learning models [150].

Towards facing filtering method issues, we considered the wrapper method and feature importance was computed for ANN, SVM, RF, and AdaBoost by a permutation method [149]. ANN was structured as a multilayer perceptron (MLP), and Back Propagation Algorithm (BPA) was considered for the model training. The number of neurons at the input layer was the same of that of explanatory features plus the eight values that represent the state of the neighboring sectors (1=burning and 0 for the other case). Five hidden dense layers with 500, 400, 250, 100, and 50 neurons were taken and Rectified Linear Unit (ReLU) was adopted as an activation function. The output layer was comprised of two neurons - one activated for ”burn” predictions and the other activated for ”no-burn” ones. Stochastic gradient descent algorithm (Adam) with 0.005 learning rate, β_1 = 0.9, and β_2 = 0.8 was the optimizer. The optimization function was the categorical cross entropy; the learning rate was set to 0.1 and the number of epochs was 1000. SVM yielded better results with RBF kernel with $\gamma = 0.01$ and regularization parameter $C = 1.0$. RF achieved its best result with 30 sampling process trees and four as the number of variables for each split. 300 was the maximum number of trees and the voting threshold (cutoff) was 0.05. The aforementioned parameters were empirically defined after several tests.

According to Figure 4.27, all features contributed to the fit of the models on higher or lower scales. VPD, NDVI, and relative humidity highly impacted the models; human factors such as distance to road can be also related to wildfires in the Federal District. However, all permutation importance values were low (under 0.25), and features with a higher importance for one model were not important in the same degree for others.

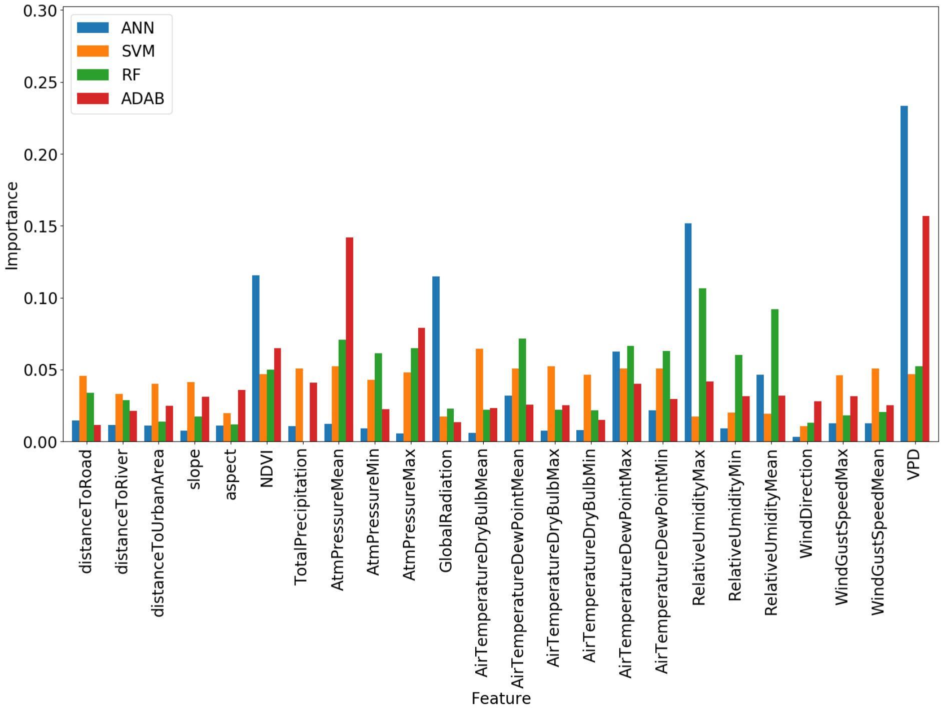


Figure 4.27: Importance of explanatory features

Table 4.7 shows the resultant features for each model, which will be used in the next sub-phase (training).

4.4.2 Hyper-parameter optimization.

The random search technique designed by [155] was adopted for the hyper-parameter optimization. Table 4.8 shows the distributions for each model and the best parameters found.

4.4.3 Validation of the models' performances

ROC curve method, applied to each fold, validated the models. Figure 4.28 displays the plotted curves and the trade-off between FPR and TPR , respectively on X and Y axes. According to the area under the curve (AUC), employed as a validation metric, values close to one denote

Table 4.7: Features considered for each model

Model	Features
ANN	VPD, Maximum Relative Humidity, NDVI, Maximum Air Temperature Dew Point, Mean Relative Humidity, Mean Air Temperature Dew Point, Maximum Atmospheric Pressure
SVM	Mean Air Temperature Dry Bulb, Mean Atmospheric Pressure, Maximum Bulb Air Temperature, Minimum Bulb Air Temperature, Maximum Wind Gust Speed
AdaBoost	VPD, Mean Atmospheric Pressure, Maximum Atmospheric Pressure, NDVI, Maximum Air Temperature Dew Point, Maximum Relative Humidity, Aspect, Total Precipitation and Slope
RF	Maximum Relative Humidity, VPD, NDVI, Global Radiation, Maximum Air Temperature Dew Point, Mean Relative Humidity, Mean Air Temperature Dew Point and Minimum Air Temperature Dew Point

Table 4.8: Distributions considered by random search optimization

Model	Distributions	Best value
ANN	activation : [relu, linear, softmax]	relu (hidden) and softmax(output)
	epochs: [10:10:300]	160
	batch size: [20:2:100]	38
	loss: [categorical cross entropy]	categorical cross entropy
	optimizer: [Adam, sgd, adadelta]	Adam
RF	criterion : [gini, entropy]	gini
	max features: [0.3:0.1:0.9]	0.5
	min samples leaf: [1,2,3,5,7,10,15]	2
	min samples split: [2,5,10]	2
	n estimators: [50:50:600]	450
SVM	kernel: [linear, sigmoid, rbf]	rbf
	gamma: [0.0001:0.0001:0.003]	0.0013
AdaBoost	estimators: [100: 100: 2000]	1100
	algorithm: [SAMME, SAMME.R]	SAMME
	learning rate:[0.1:0.1:1]	1

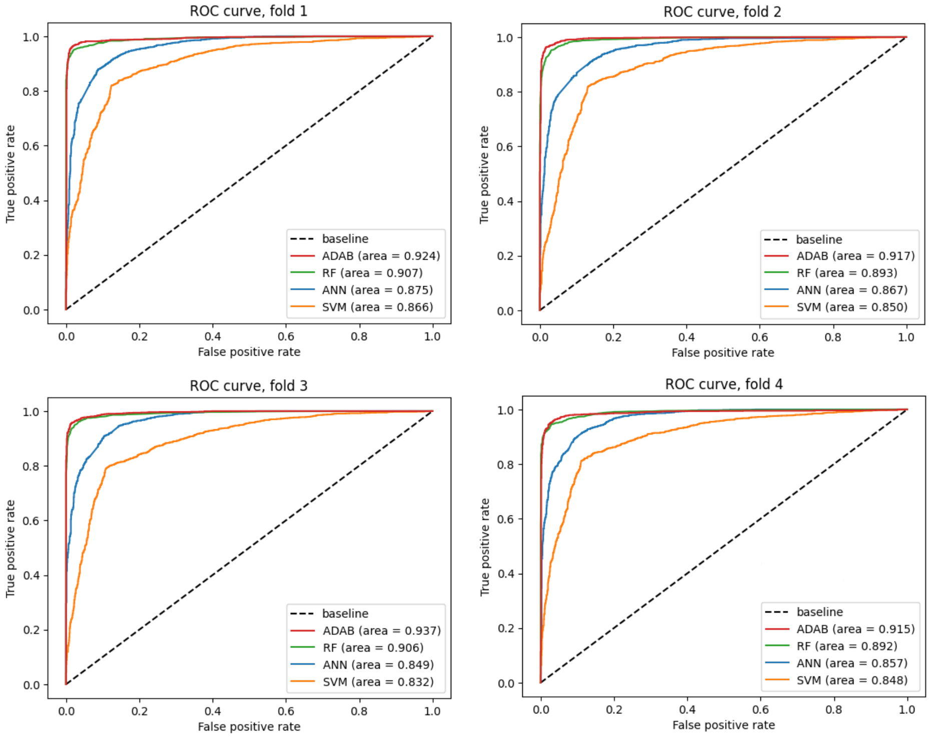


Figure 4.28: AUC ROC 4-fold for fire behavior prediction

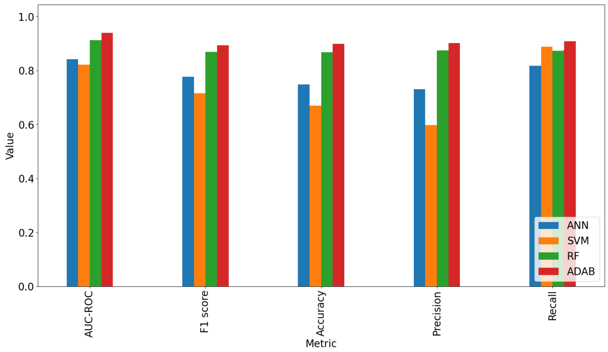


Figure 4.29: Validation metrics for fire behaviour prediction

high accuracy. The models yielded relevant results, and AdaBoost, with a 0.92 AUC, showed the best performance.

Figure 4.29 shows F1 score, Accuracy, Precision, and Recall metrics. AdaBoost reached values above 0.86, indicating the best performance, whereas the values reached by RF and ANN ranged between 0.81 and 0.85. SVM achieved the worst performance.

In relation to other proposals applied in different regions and with different datasets, the results are encouraging. In terms of AUC-ROC, the average value obtained (0.92) was equal to the 0.92 from [13]. When compared to the values of accuracy obtained by [13] (0.87) and [11] (0.93), the 0.89 accuracy can be considered satisfactory. The results are good regardless of the lack of research on the region and the predicted burnt area, as discussed below.

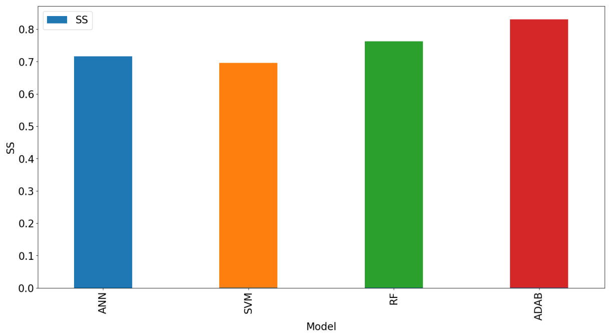


Figure 4.30: 500 samples Sørensen Similarity Mean

Table 4.9: SS agreement levels by ranges

SS range	Agreement level
(0, 0.2]	slight
(0.2, 0.4]	fair
(0.4, 0.6]	moderate
(0.6, 0.8]	substantial
(0.8, 1]	near-perfect

4.4.4 Prediction and validation of burned area

Once the models had been trained, Algorithm 1 generated the final burned areas, and SS metric, defined by Equation 6, measured the quality of the predicted burned areas.

SS was calculated for each model from 500 random samples belonging to the sets devoted to training and validation. Figure 4.30 shows all models exceeded 0.69 mean SS. AdaBoost provided the best results (0.83) and was considered the winner model.

As addressed elsewhere, the fires in the Federal District have not been extensively studied and no research on the final predictions of a burned area has been found. Consequently, the quality of the predicted fire scars cannot be easily compared and validated from a regional perspective. However, as defined by [164] (see Table 4.9), 0.83 SS value indicates near-perfect agreements, thus highlighting the relevance of the results of this research.

4.5 CONCLUSIONS

In this chapter, after a literature review of wildfires risk and behaviour predictions models and explanatory features, we introduced a fire-related dataset in the Brazilian Federal District, and series of short-term spatial/temporal records of wildfires for the prediction of wildfires behaviour. Four ML approaches, namely ANN, SVM, RF, and AdaBoost were trained according to previous wildfire events and vegetation, climatic, hydrographic, and anthropogenic factors to predict if a sector will burn based on its neighboring conditions. Their performances were compared and AdaBoost achieved the best results.

An algorithm for fire-scars reconstruction was provided and validated by Sørensen similarity metric. The computation of each explanatory feature importance revealed all features considered impacted the fire behaviour.

The workflow produced can be extended and adapted to other Brazilian regions (e.g., Brazilian Amazonia) and most probably to other countries. ML theory enable prediction of wildfire behaviours according to topographic, climate, hydrographic, and anthropogenic data, such as those of our inventory, leading to a precise evaluation of a burnt area.

5 CONCLUSIONS AND FUTURE WORK

Advances of Smart Cities and Forestry 4.0, supported by Internet of Things, Wireless Sensor Networks, Wireless Communications, and ML have imposed challenges regarding the standardization of data collection and sharing processes, towards the development of data mining applications in several city-related domains. New research and product development must focus on reducing data collection and interchange complexity for simplifying data mining and machine learning applications.

The lack of a common data format and sharing standard and the various networking and sensor technologies have led to a heterogeneous environment of IoT devices/platforms that must be integrated into an interoperable one. Regarding SCs environmental data, a growth is expected in the number of IoT platforms that deploy sensors related to indicators data, and their integration must be considered the backbone of environment-related city services.

This thesis presented a new platform for Smart Cities applications based on semantics technologies that enables a seamless collection and interchange of data, and the application of machine learning models over the collected data. The platform was validated through a use case on wildfire predictions.

To enable a dynamic data scheme for the representation of data, sensors, SC and forestry systems, ontologies aligned with IoT conceptualization were proposed, allowing a formal definition of SC indicators (based on ISO 37120) and forestry observational environment. The proposed ontologies promoted the retrieval of aggregated data represented in RDF format.

The state of the art about SC platforms was discussed, leading to the conclusion the use of semantic technologies for data representation and sharing between Smart City stakeholders can solve some of the heterogeneity challenges and the ontology proposed acts as a single definition of heterogeneous data sources and data collection technologies.

The IoT-based platform for the environment SC domain followed a three-layered IoT architecture and enabled the collection, storage, and processing of environmental data from city and forestry considering Fog resources for a local processing of data and through services and resources dynamically deployed in the cloud at city levels.

A set of adapters for several communication protocols was implemented for dealing with networks and sensors heterogeneity. Such adapters can be instantiated on the Fog and Cloud Interfaces. Moreover, SenML was considered for the sake of compatibility of data representation and for reducing the overhead of controlling data. A mapping of the sensor International Resource Identifier – IRI (in both gateway and fog server) promoted reductions in message overload and consumption of resources in constrained devices.

The research considered the collection and proposal of a dataset of fire occurrence in the

Brazilian Federal District which include climatic, vegetation, hydrographic, and anthropogenic related data.

Machine learning functionalities were provided on the cloud and enabled the preparation of training datasets in a standardized way, considering the use of SPARQL and RESTFul for query end- points and SenML and RDF as the format of the response object. The platform capabilities promoted the development of some wildfire studies in the Federal District Region Initially, the prediction of fire risk and its main wildfire conditioning factors were identified.

Eight ML models (ANN, SVM, RF, Naive Bayes, KNN, Linear Reg., Log. Reg, and AdaBoost) were developed and trained according to previous wildfire events and their performances were evaluated through 4-fold cross validation. AdaBoost achieved the highest AUC-ROC (0.993).

The importance of each explanatory feature was computed towards the identification of the main originating factors of wildfires in the region, and many models proved more sensitive to NDVI, relative humidity, and atmospheric pressure.

A set of short-term spatio-temporal series was also proposed for representing the fire behavior. ANN and RF were trained according to the fire sequences for predictions on how the fire-line moves after a fire has been initiated. An algorithm developed for the construction of fire scars showed the predicted scars can be classified as good with a 0.77 Sørensen similarity.

5.1 LIMITATIONS

The proposed platform meets several of the heterogeneity IoT-related challenges. However, it shows some limitations, as discussed below.

The deployment of the sensor is not seamless. Since the platform provides no discovery and pairing protocol, the sensor must be manually registered on fog or cloud servers to obtain an identification (id) to be configured prior its implantation/deployment towards the identification of the observations obtained. A sensor pairing strategy should be implemented by platform users for the inclusion of that feature.

Although some data may be sensible and the high volume of data collected by the IoT platform demands the maintenance of a certain degree of privacy, however, the platform offers no tools for data anonimization and considers no definitions for access control.

5.2 FUTURE RESEARCH

Regarding security, since the platform does not authenticate/authorize sensors/users, security protocols must be adopted or proposed. Almost all IoT/IoFT sensors are battery-sourced devices and topics such as battery consumption must be considered in future work. Moreover, extended functions for sensor identity handling must be provided in scenarios of a sensor network with a cluster relay.

As a continuation of this research, the platform will be extended towards enriching fire terminology on EISCO ontology and proposing a new set of fire-related QoL indicators. Besides environment, the conceptualization of other SC city verticals such as health, safety and security, and transportation will be studied and the GPSO ontology will be extended towards a fully set of SC IT-based tools.

BIBLIOGRAPHIC REFERENCES

1 Instituto Nacional de Pesquisa Espacial (INPE). *Banco de dados de queimadas*. 2020. Http://queimadas.dgi.inpe.br/queimadas/bdqueimadas/. Accessed 2020-01-06.

2 Governo do Distrito Federal (GDF). *GEOPortal*. 2020. Https://www.geoportal.seduh.df.gov.br/mapa/. Accessed 2020-01-06.

3 RUBÍ, J. N. S.; GONDIM, P. R. de L. Iot-based platform for environment data sharing in smart cities. *International Journal of Communication Systems*, Wiley Online Library, v. 34, n. 2, p. e4515, 2021.

4 RUBÍ, J. N.; CARVALHO, P. H. de; GONDIM, P. R. Forestry 4.0 and industry 4.0: Use case on wildfire behavior predictions. *Computers and Electrical Engineering*, Elsevier, v. 102, p. 108200, 2022.

5 RUBÍ, J. N.; CARVALHO, P. H. de; GONDIM, P. R. Application of machine learning models in the behavioral study of forest fires in the brazilian federal district region. *Engineering Applications of Artificial Intelligence*, Elsevier, v. 118, p. 105649, 2023.

6 FELCE, D.; PERRY, J. Quality of life: Its definition and measurement. *Research in Developmental Disabilities*, Elsevier, v. 16, n. 1, p. 51–74, 1995.

7 LIU, X.; HELLER, A.; NIELSEN, P. S. CITIESData: a smart city data management framework. *Knowledge and Information Systems*, Springer London, v. 53, n. 3, p. 699–722, 2017. ISSN 02193116.

8 BOTTA, A.; DONATO, W. D.; PERSICO, V.; PESCAPÉ, A. Integration of cloud computing and internet of things: a survey. *Future Generation Computer Systems*, Elsevier, v. 56, p. 684–700, 2016.

9 ISO. *ISO 37120:2014 - Sustainable development of communities – Indicators for city services and quality of life*. Https://www.iso.org/standard/62436.html. Accessed 2018-04-14.

10 VIJAYAKUMAR, N.; RAMYA, R. The real time monitoring of water quality in IoT environment. In: IEEE. *2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*. [S.l.], 2015. p. 1–5.

11 ENCINAS, C.; RUIZ, E.; CORTEZ, J.; ESPINOZA, A. Design and implementation of a distributed IoT system for the monitoring of water quality in aquaculture. In: IEEE. *2017 Wireless Telecommunications Symposium (WTS)*. [S.l.], 2017. p. 1–7.

12 UçAR, Z.; AKAY, A. E.; BILICI, E. Towards green smart cities: Importance of urban forestry and urban vegetation. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, International Society for Photogrammetry and Remote Sensing, 2020.

13 SHARMA, A.; SINGH, P. K.; KUMAR, Y. An integrated fire detection system using iot and image processing technique for smart cities. *Sustainable Cities and Society*, Elsevier, v. 61, p. 102332, 2020.

14 Sistema Nacional de Informações Florestais. *Bens e Serviços que a Floresta Fornece*. 2019. Https://snif.florestal.gov.br/pt-br/conhecendo-sobre-florestas/169-bens-e-servicos-que-a-floresta-fornece. Accessed 2021-10-23.

15 CHIRICO, G. B.; BONAVOLONTà, F. Metrology for agriculture and forestry 2019. *Sensors*, v. 20, n. 12, 2020. ISSN 1424-8220. Disponível em: <https://www.mdpi.com/1424-8220/20/12/3498>.

16 MONEDERO, S.; RAMIREZ, J.; CARDIL, A. Predicting fire spread and behaviour on the fireline. wildfire analyst pocket: A mobile app for wildland fire prediction. *Ecological Modelling*, Elsevier, v. 392, p. 103–107, 2019.

17 SAHAL, R.; ALSAMHI, S. H.; BRESLIN, J. G.; ALI, M. I. Industry 4.0 towards forestry 4.0: Fire detection use case. *Sensors*, Multidisciplinary Digital Publishing Institute, v. 21, n. 3, p. 694, 2021.

18 SALAM, A. Internet of things for sustainable forestry. In: *Internet of Things for Sustainable Community Development*. [S.l.]: Springer, 2020. p. 147–181.

19 JIANG, W.; WANG, F.; FANG, L.; ZHENG, X.; QIAO, X.; LI, Z.; MENG, Q. Modelling of wildland-urban interface fire spread with the heterogeneous cellular automata model. *Environmental Modelling & Software*, Elsevier, v. 135, p. 104895, 2021.

20 ROSSA, C. G.; FERNANDES, P. M. Empirical modeling of fire spread rate in no-wind and no-slope conditions. *Forest Science*, Oxford University Press US, v. 64, n. 4, p. 358–370, 2018.

21 CHEN, T. B. Y.; YUEN, A.; YEOH, G.; TIMCHENKO, V.; CHEUNG, S. C.; CHAN, Q.; YANG, W.; LU, H. Numerical study of fire spread using the level-set method with large eddy simulation incorporating detailed chemical kinetics gas-phase combustion model. *Journal of computational science*, Elsevier, v. 24, p. 8–23, 2018.

22 MUELLER, E.; MELL, W.; SIMEONI, A. Large eddy simulation of forest canopy flow for wildland fire modeling. *Canadian Journal of Forest Research*, NRC Research Press, v. 44, n. 12, p. 1534–1544, 2014.

23 ZHAI, C.; ZHANG, S.; CAO, Z.; WANG, X. Learning-based prediction of wildfire spread with real-time rate of spread measurement. *Combustion and Flame*, Elsevier, v. 215, p. 333–341, 2020.

24 ZUPO, T. M. Estratégias de persistência e regeneração em campo sujo de cerrado após o fogo. Masters's degree dissertation. Universidade Estadual Paulista (UNESP). 2017. Universidade Estadual Paulista (UNESP), 2017.

25 BEM, P. P. de; JÚNIOR, O. A. de C.; MATRICARDI, E. A. T.; GUIMARÃES, R. F.; GOMES, R. A. T. Predicting wildfire vulnerability using logistic regression and artificial neural networks: a case study in brazil's federal district. *International journal of wildland fire*, CSIRO, v. 28, n. 1, p. 35–45, 2019.

26 GORGONE-BARBOSA, E.; PIVELLO, V. R.; BAUTISTA, S.; ZUPO, T.; RISSI, M. N.; FIDELIS, A. How can an invasive grass affect fire behavior in a tropical savanna? a community and individual plant level approach. *Biological Invasions*, Springer, v. 17, n. 1, p. 423–431, 2015.

27 RUBÍ, J. N.; GONDIM, P. R. A performance comparison of ML models for wildfire risk prediction in the brazilian federal district region. *Environment Systems and Decisions Journal*, Springer, tbd, p. tbd, tbd.

28 GRUBER, T. *Definition of Ontology 07*. 2018. Http://tomgruber.org/writing/ontology-definition-2007. Visited 2018-10-07.

29 EUZENAT, J.; SHVAIKO, P. *Ontology matching*. [S.l.]: Springer-Verlag, Berlin Heidelberg, 2007. v. 18.

30 GUARINO, N. *Formal ontology in information systems: Proceedings of the First International Conference (FOIS'98), June 6-8, Trento, Italy*. [S.l.]: IOS press, 1998. v. 46.

31 INFORMEA. *The Law and Environment Ontology (LEO) portal*. 2015. Https://www.informea.org/en/terms. Accessed 2019-04-01.

32 OPREA, M. M. Air_pollution_onto: an ontology for air pollution analysis and control. In: SPRINGER. *IFIP International Conference on Artificial Intelligence Applications and Innovations*. [S.l.], 2009. p. 135–143.

33 RASKIN, R. G.; PAN, M. J. Knowledge representation in the semantic web for earth and environmental terminology (sweet). *Computers & Geosciences*, Elsevier, v. 31, n. 9, p. 1119–1125, 2005.

34 BUTTIGIEG, P. L.; MORRISON, N.; SMITH, B.; MUNGALL, C. J.; LEWIS, S. E. The environment ontology: contextualising biological and biomedical entities. *Journal of biomedical semantics*, BioMed Central, v. 4, n. 1, p. 43, 2013.

35 FOX, M. S. The role of ontologies in publishing and analyzing city indicators. *Computers, Environment and Urban Systems*, Elsevier B.V., v. 54, p. 266–279, 2015. ISSN 01989715.

36 DAHLEH, D.; FOX, S. M. *An Environmental Ontology for Global City Indicators (ISO 37120)*. 2016. Http://eil.mie.utoronto.ca/wp-content/uploads/2015/06/GCI-Environmental-Ontology-Final-17sep2016.pdf. Accessed 05/06/2018.

37 W3C. *Semantic Sensor Network Ontology*. 2017. Https://www.w3.org/TR/vocab-ssn/#intro. Accessed 2018-06-18.

38 GANZHA, M.; PAPRZYCKI, M. Semantic interoperability in the Internet of Things : An overview from the INTER-IoT perspective. *Journal of Network and Computer Applications*, v. 81, p. 111–124, 2017.

39 W3C. *SensorOntology2009 - Semantic Sensor Network Incubator Group*. Https://www.w3.org/2005/Incubator/ssn/wiki/ SensorOntology2009. Visited 2018-10-13.

40 RUEDA, C.; BERMUDEZ, L.; FREDERICKS, J. The mmi ontology registry and repository: A portal for marine metadata interoperability. In: *OCEANS 2009*. [S.l.: s.n.], 2009. p. 1–6. ISSN 0197-7385.

41 SCHILDHAUER, M.; JONES, M.; BOWERS, S.; MADIN, J.; KRIVOV, S.; PENNINGTON, D.; VILLA, F.; LEINFELDER, B.; JONES, C.; O'BRIEN, M. *OBOE — Semantic Tools Project*. Https://github.com/NCEAS/oboe. Visited 2018-10-13.

42 MÜLLER, H.; CABRAL, L.; MORSHED, A.; SHU, Y. From restful to sparql: A case study on generating semantic sensor data. In: *SSN@ ISWC*. [S.l.: s.n.], 2013. p. 51–66.

43 RIJGERSBERG, H.; ASSEM, M. van; TOP, J. Ontology of units of measure and related concepts. *Semantic Web*, IOS Press, v. 4, n. 1, p. 3–13, 2013. ISSN 1570-0844.

44 W3C. *W3C Semantic Web Interest Group: Basic Geo (WGS84 lat/long) Vocabulary*. Https://www.w3.org/2003/01/geo. Visited 2018-06-18.

45 ATEMEZING, G.; CORCHO, O.; GARIJO, D.; MORA, J.; POVEDA-VILLALÓN, M.; ROZAS, P.; VILA-SUERO, D.; VILLAZÓN-TERRAZAS, B. Transforming meteorological data into linked data. *Semantic Web*, IOS Press, v. 4, n. 3, p. 285–290, 2013. ISSN 15700844.

46 HOBBS, J. R.; PAN, F. An ontology of time for the semantic web. *ACM Transactions on Asian Language Information Processing*, v. 3, n. 1, p. 66–85, mar 2004. ISSN 15300226.

47 WANG, C.; CHEN, N.; HU, C.; YAN, S.; WANG, W. A general sensor web resource ontology for atmospheric observation. In: IEEE. *Geoscience and Remote Sensing Symposium (IGARSS), 2011 IEEE International*. [S.l.], 2011. p. 3436–3439.

48 Maria Bermudez-Edo, Institute for Communication Systems, U. o. S.; GRANADA, U. of; Tarek Elsaleh, Institute for Communication Systems, U. o. S.; Payam Barnaghi, Institute for Communication Systems, U. o. S.; Kerry Taylor, Institute for Communication Systems, U. o. S.; UNIVERSITY, T. A. N. *IoT-Lite Ontology*. Https://www.w3.org/Submission/iot-lite. Visited 2018-10-13.

49 Working group of the SysML 1.2 Revision Task Force (RTF), W3C Semantic Sensor Network Incubator Group. *Library for Quantity Kinds and Units: schema, based on QUDV model OMG SysML(TM), Version 1.2*. Https://www.w3.org/2005/Incubator/ssn/wiki/QU_Ontology. Visited 2018-10-13.

50 DANIELE, L.; HARTOG, F. den; ROES, J. Created in close interaction with the industry: the smart appliances reference (saref) ontology. In: SPRINGER. *International Workshop Formal Ontologies Meet Industries*. [S.l.], 2015. p. 100–112.

51 GUPTA, S.; PADHY, A.; ADHIKARI, A.; DUTTA, A. A semantic web and linked data based framework for Smart City data management. *2016 13th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, p. 1–6, 2016.

52 GAUR, A.; SCOTNEY, B.; PARR, G.; MCCLEAN, S. Smart city architecture and its applications based on IoT. *Procedia Computer Science*, Elsevier Masson SAS, v. 52, n. 1, p. 1089–1094, 2015. ISSN 18770509.

53 ABID, T.; ZARZOUR, H.; LAOUAR, M. R.; KHADIR, M. T. Towards a smart city ontology. In: *2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA)*. [S.l.: s.n.], 2016. p. 1–6. ISSN 2161-5330.

54 RDF and Semantic Web developer community. *FOAF Vocabulary Specification*. Http://xmlns.com/foaf/spec. Visited 2018-10-13.

55 PETROLO, R.; LOSCRÌ, V.; MITTON, N. Towards a smart city based on cloud of things, a survey on the smart city vision and paradigms. *Transactions on Emerging Telecommunications Technologies*, v. 28, n. 1, 2017. ISSN 21613915.

56 D'AQUIN, M.; DAVIES, J.; MOTTA, E. Smart Cities' Data: Challenges and Opportunities for Semantic Technologies. *IEEE Internet Computing*, v. 19, n. 6, p. 66–70, 2015. ISSN 10897801.

57 BSI Group. *Standard - Hypercat*. Http://www.hypercat.io/standard.html. Visited 2018-04-05.

58 LEA, R.; BLACKSTOCK, M. City hub: A cloud-based IoT platform for smart cities. In: *Proceedings of the International Conference on Cloud Computing Technology and Science, CloudCom*. [S.l.: s.n.], 2015. v. 2015-Febru, n. February, p. 799–804. ISBN 978-1-4799-4093-6. ISSN 23302186.

59 BLACKSTOCK, M.; LEA, R. Iot mashups with the wotkit. In: IEEE. *Internet of Things (IOT), 2012 3rd International Conference on the*. [S.l.], 2012. p. 159–166.

60 CKAN Organization. *About – ckan*. Https://ckan.org/about. Visited 2020-08-10.

61 ABREU, D. P.; VELASQUEZ, K.; PINTO, A. M.; CURADO, M.; MONTEIRO, E. Describing the Internet of Things with an ontology: The SusCity project case study. *Proceedings of the 2017 20th Conference on Innovations in Clouds, Internet and Networks, ICIN 2017*, p. 294–299, 2017.

62 COMPTON, M.; CORSAR, D.; TAYLOR, K. Sensor data provenance: SSNO and PROV-O together at last. In: *CEUR Workshop Proceedings*. [S.l.: s.n.], 2014. v. 1401, p. 67–82. ISSN 16130073.

63 Open Geospatial Consortium. *GeoSPARQL - A Geographic Query Language for RDF Data | OGC*. 2012. Http://www.opengeospatial.org/standards/geosparql. Visited 2018-05-25.

64 LI, S.; XU, L. D.; ZHAO, S. The internet of things: a survey. *Information Systems Frontiers*, Springer, v. 17, n. 2, p. 243–259, 2015.

65 AGIWAL, M.; ROY, A.; SAXENA, N. Next generation 5g wireless networks: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, IEEE, v. 18, n. 3, p. 1617–1655, 2016.

66 MINOLI, D.; OCCHIOGROSSO, B. Practical aspects for the integration of 5g networks and iot applications in smart cities environments. *Wireless Communications and Mobile Computing*, Hindawi, v. 2019, 2019.

67 GAMAGE, H.; RAJATHEVA, N.; LATVA-AHO, M. Channel coding for enhanced mobile broadband communication in 5g systems. In: IEEE. *2017 European conference on networks and communications (EuCNC)*. [S.l.], 2017. p. 1–6.

68 SHARMA, S. K.; WANG, X. Toward massive machine type communications in ultra-dense cellular iot networks: Current issues and machine learning-assisted solutions. *IEEE Communications Surveys & Tutorials*, IEEE, v. 22, n. 1, p. 426–471, 2019.

69 LIEN, S.-Y.; HUNG, S.-C.; DENG, D.-J.; WANG, Y. J. Efficient ultra-reliable and low latency communications and massive machine-type communications in 5g new radio. In: IEEE. *GLOBECOM 2017-2017 IEEE Global Communications Conference*. [S.l.], 2017. p. 1–7.

70 RASHED, A.; IBRAHIM, A.; ADEL, A.; MOURAD, B.; HATEM, A.; MAGDY, M.; ELGAML, N.; KHATTAB, A. Integrated IoT medical platform for remote healthcare and assisted living. In: IEEE. *Electronics, Communications and Computers (JAC-ECC), 2017 Japan-Africa Conference on*. [S.l.], 2017. p. 160–163.

71 MQTT Org. *Message Queuing Telemetry Transport Protocol*. 2018. Http://mqtt.org/. Accessed 2018-10-28.

72 DATTA, S. K.; BONNET, C.; NIKAEIN, N. An iot gateway centric architecture to provide novel m2m services. In: *IEEE World Forum on Internet of Things (WF-IoT)*. [S.l.]: IEEE World Forum on Internet of Things (WF-IoT), 2014. p. 514–519.

73 PAGANELLI, F.; TURCHI, S.; GIULI, D. A web of things framework for restful applications and its experimentation in a smart city. *IEEE Systems Journal*, IEEE, v. 10, n. 4, p. 1412–1423, 2016.

74 JAN, S. R.; KHAN, F.; ULLAH, F.; AZIM, N.; TAHIR, M. Using coap protocol for resource observation in iot. *International Journal of Emerging Technology in Computer Science & Electronics, ISSN*, p. 0976–1353, 2016.

75 SHELBY, Z.; HARTKE, K.; BORMANN, C. *The constrained application protocol (CoAP)*. [S.l.], 2014. 1-122 p. Disponível em: <https://www.rfc-editor.org/rfc/rfc7252.txt>.

76 DESAI, P.; SHETH, A.; ANANTHARAM, P. Semantic gateway as a service architecture for iot interoperability. In: . [S.l.]: IEEE International Conference on Mobile Services (MS), 2015. p. 313–319.

77 XMPPORG. *eXtensible Protocol Presence Messaging*. Https://xmpp.org/about/. Accessed 2019-02-16.

78 AMQPORG. *Advanced Message Queuing Protocol*. Https://www.amqp.org/. Accessed 2019-02-18.

79 DIZDAREVIC, J.; CARPIO, F.; JUKAN, A.; MASIP-BRUIN, X. Survey of communication protocols for internet-of-things and related challenges of fog and cloud computing integration. *arXiv preprint arXiv:1804.01747*, 2018.

80 JENNINGS, C.; ARKKO, J.; SHELBY, Z. *RFC 8428 Sensor Measurement Lists (SenML)*. 2018. Https://tools.ietf.org/html/rfc8428. Accessed 2019-08-01.

81 ZHAO, L.; CHEN, Z.; YANG, Y.; WANG, Z. J.; LEUNG, V. C. Incomplete multi-view clustering via deep semantic mapping. *Neurocomputing*, Elsevier, v. 275, p. 1053–1062, 2018.

82 ZHAO, L.; CHEN, Z.; YANG, L. T.; DEEN, M. J.; WANG, Z. J. Deep semantic mapping for heterogeneous multimedia transfer learning using co-occurrence data. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, ACM, v. 15, n. 1s, p. 9, 2019.

83 LIU, J.; LI, T.; XIE, P.; DU, S.; TENG, F.; YANG, X. Urban big data fusion based on deep learning: An overview. *Information Fusion*, Elsevier, v. 53, p. 123–133, 2020.

84 IRFAN, M.; AHMAD, N. Internet of medical things: Architectural model, motivational factors and impediments. In: IEEE. *Learning and Technology Conference (L&T), 2018 15th*. [S.l.], 2018. p. 6–13.

85 DUTTA, J.; ROY, S. Iot-fog-cloud based architecture for smart city: Prototype of a smart building. In: *7th International Conference on Cloud Computing, Data Science & Engineering-Confluence*. [S.l.]: 7th International Conference on Cloud Computing, Data Science & Engineering-Confluence, 2017. p. 237–242.

86 GIANG, N. K.; LEA, R.; BLACKSTOCK, M.; LEUNG, V. On building smart city IoT applications: A coordination-based perspective. In: ACM. *Proceedings of the 2nd International Workshop on Smart*. [S.l.], 2016. p. 7.

87 LEA, R.; BLACKSTOCK, M. City hub: A cloud-based IoT platform for smart cities. In: IEEE. *2014 IEEE 6th international conference on cloud computing technology and science*. [S.l.], 2014. p. 799–804.

88 ANTONIĆ, A.; MARJANOVIĆ, M.; PRIPUŽIĆ, K.; ŽARKO, I. P. A mobile crowd sensing ecosystem enabled by cupus: Cloud-based publish/subscribe middleware for the internet of things. *Future Generation Computer Systems*, Elsevier, v. 56, p. 607–622, 2016.

89 SANTOS, P. M.; RODRIGUES, J. G.; CRUZ, S. B.; LOURENÇO, T.; D'OREY, P. M.; LUIS, Y.; ROCHA, C.; SOUSA, S.; CRISÓSTOMO, S.; QUEIRÓS, C. et al. PortoLivingLab: An IoT-based sensing platform for smart cities. *IEEE Internet of Things Journal*, IEEE, v. 5, n. 2, p. 523–532, 2018.

90 SANTOS, J.; WAUTERS, T.; VOLCKAERT, B.; TURCK, F. D. Fog computing: Enabling the management and orchestration of smart city applications in 5g networks. *Entropy*, MDPI, v. 20, n. 1, p. 4, 2017.

91 SU, X.; LIU, X.; MOTLAGH, N. H.; CAO, J.; SU, P.; PELLIKKA, P.; LIU, Y.; PETÄJÄ, T.; KULMALA, M.; HUI, P. et al. Intelligent and scalable air quality monitoring with 5g edge. *IEEE Internet Computing*, IEEE, v. 25, n. 2, p. 35–44, 2021.

92 RAHMANI, A. M.; GIA, T. N.; NEGASH, B.; ANZANPOUR, A.; AZIMI, I.; JIANG, M.; LILJEBERG, P. Exploiting smart e-health gateways at the edge of healthcare internet-of-things: A fog computing approach. *Future Generation Computer Systems*, Elsevier, v. 78, p. 641–658, 2018.

93 HOWELL, S.; REZGUI, Y.; BEACH, T. Integrating building and urban semantics to empower smart water solutions. *Automation in Construction*, Elsevier, v. 81, p. 434–448, 2017.

94 TSIROPOULOU, E. E.; PARUCHURI, S. T.; BARAS, J. S. Interest, energy and physical-aware coalition formation and resource allocation in smart iot applications. In: IEEE. *2017 51st Annual Conference on Information Sciences and Systems (CISS)*. [S.l.], 2017. p. 1–6.

95 SU, X.; ZHANG, H.; RIEKKI, J.; KERÄNEN, A.; NURMINEN, J. K.; DU, L. Connecting iot sensors to knowledge-based systems by transforming senml to rdf. *Procedia Computer Science*, Elsevier, v. 32, p. 215–222, 2014.

96 Instituto Nacional de Meteorologia (INMET). *Banco de Dados Meteorológicos do INMET*. 2020. Http://www.inmet.gov.br/portal/index.php?r=informacoes/ cartaProdutoServicoCidadaoView&id=45. Accessed 2020-01-06.

97 DIDAN, K. *MOD13Q1 MODIS/Terra Vegetation Indices 16-Day L3 Global 250m SIN Grid V006 [Data set]. NASA EOSDIS Land Processes DAAC.* 2020. Https://doi.org/10.5067/MODIS/ MOD13Q1.006, Accessed 2020-01-06.

98 CASTRO, S. de. *Weather Madrid 1997 - 2015*. Https://www.kaggle.com/juliansimon/weather_madrid_lemd _1997_2015.csv Accessed 2019-02-13.

99 DECIDESOLUCIONES. *Air Quality in Madrid (2001-2018)*. Https://www.kaggle.com/decide-soluciones/air-quality-madrid Accessed 2019-02-18.

100 RODRIGUES, M.; RIVA, J. de la. An insight into machine-learning algorithms to model human-caused wildfire occurrence. *Environmental Modelling & Software*, Elsevier, v. 57, p. 192–201, 2014.

101 TAVARES, M. d. F. D.; NAKAGOMI, B.; SOARES, V.; BOTEGA, L. C.; NERIS, V. P. de A. Paisagens protegidas e incêndios florestais em brasília: Sistema de alerta e a produção voluntária de informações geográficas. *Territorium*, n. 26 (I), p. 63–86, 2019.

102 GHORBANZADEH, O.; BLASCHKE, T.; GHOLAMNIA, K.; ARYAL, J. Forest fire susceptibility and risk mapping using social/infrastructural vulnerability and environmental variables. *Fire*, Multidisciplinary Digital Publishing Institute, v. 2, n. 3, p. 50, 2019.

103 JAAFARI, A.; TERMEH, S. V. R.; BUI, D. T. Genetic and firefly metaheuristic algorithms for an optimized neuro-fuzzy prediction modeling of wildfire probability. *Journal of environmental management*, Elsevier, v. 243, p. 358–369, 2019.

104 BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001.

105 MASSADA, A. B.; SYPHARD, A. D.; STEWART, S. I.; RADELOFF, V. C. Wildfire ignition-distribution modelling: a comparative study in the huron–manistee national forest, michigan, usa. *International journal of wildland fire*, CSIRO, v. 22, n. 2, p. 174–183, 2013.

106 GHORBANZADEH, O.; KAMRAN, K. V.; BLASCHKE, T.; ARYAL, J.; NABOUREH, A.; EINALI, J.; BIAN, J. Spatial prediction of wildfire susceptibility using field survey gps data and machine learning approaches. *Fire*, Multidisciplinary Digital Publishing Institute, v. 2, n. 3, p. 43, 2019.

107 HONG, H.; PRADHAN, B.; SAMEEN, M. I.; CHEN, W.; XU, C. Spatial prediction of rotational landslide using geographically weighted regression, logistic regression, and support vector machine models in xing guo area (china). *Geomatics, Natural Hazards and Risk*, Taylor & Francis, v. 8, n. 2, p. 1997–2022, 2017.

108 MILLER, C.; AGER, A. A. A review of recent advances in risk analysis for wildfire management. *International journal of wildland fire*, CSIRO Publishing, v. 22, n. 1, p. 1–14, 2012.

109 JAAFARI, A.; ZENNER, E. K.; PANAHI, M.; SHAHABI, H. Hybrid artificial intelligence models based on a neuro-fuzzy system and metaheuristic optimization algorithms for spatial prediction of wildfire probability. *Agricultural and forest meteorology*, Elsevier, v. 266, p. 198–207, 2019.

110 KIM, S. J.; LIM, C.-H.; KIM, G. S.; LEE, J.; GEIGER, T.; RAHMATI, O.; SON, Y.; LEE, W.-K. Multi-temporal analysis of forest fire probability using socio-economic and environmental variables. *Remote Sensing*, Multidisciplinary Digital Publishing Institute, v. 11, n. 1, p. 86, 2019.

111 NAMI, M.; JAAFARI, A.; FALLAH, M.; NABIUNI, S. Spatial prediction of wildfire probability in the hyrcanian ecoregion using evidential belief function model and gis. *International journal of environmental science and technology*, Springer, v. 15, n. 2, p. 373–384, 2018.

112 RIHAN, W.; ZHAO, J.; ZHANG, H.; GUO, X.; YING, H.; DENG, G.; LI, H. Wildfires on the mongolian plateau: Identifying drivers and spatial distributions to predict wildfire probability. *Remote Sensing*, Multidisciplinary Digital Publishing Institute, v. 11, n. 20, p. 2361, 2019.

113 SAYAD, Y. O.; MOUSANNIF, H.; MOATASSIME, H. A. Predictive modeling of wildfires: A new dataset and machine learning approach. *Fire safety journal*, Elsevier, v. 104, p. 130–146, 2019.

114 TONINI, M.; D'ANDREA, M.; BIONDI, G.; ESPOSTI, S. D.; TRUCCHIA, A.; FIORUCCI, P. A machine learning-based approach for wildfire susceptibility mapping. the case study of the liguria region in italy. *Geosciences*, Multidisciplinary Digital Publishing Institute, v. 10, n. 3, p. 105, 2020.

115 GHOLAMNIA, K.; NACHAPPA, T. G.; GHORBANZADEH, O.; BLASCHKE, T. Comparisons of diverse machine learning approaches for wildfire susceptibility mapping. *Symmetry*, Multidisciplinary Digital Publishing Institute, v. 12, n. 4, p. 604, 2020.

116 KAUR, H.; SOOD, S. K. A smart disaster management framework for wildfire detection and prediction. *The Computer Journal*, Oxford Academic, bxz091, 2020.

117 GALIZIA, L. F. d. C.; RODRIGUES, M. Modeling the influence of eucalypt plantation on wildfire occurrence in the brazilian savanna biome. *Forests*, Multidisciplinary Digital Publishing Institute, v. 10, n. 10, p. 844, 2019.

118 GOMES, L.; MIRANDA, H. S.; BUSTAMANTE, M. M. da C. How can we advance the knowledge on the behavior and effects of fire in the cerrado biome? *Forest Ecology and Management*, Elsevier, v. 417, p. 281–290, 2018.

119 JÚNIOR, A. C. P.; OLIVEIRA, S. L.; PEREIRA, J. M.; TURKMAN, M. A. A. Modelling fire frequency in a cerrado savanna protected area. *PloS one*, Public Library of Science, v. 9, n. 7, 2014.

120 SANTANA, N. C.; JÚNIOR, O. A. de C.; GOMES, R. A. T.; GUIMARÃES, R. F. Burned-area detection in amazonian environments using standardized time series per pixel in modis data. *Remote Sensing*, Multidisciplinary Digital Publishing Institute, v. 10, n. 12, p. 1904, 2018.

121 GUEDES, B. J.; MASSI, K. G.; EVERS, C.; NIELSEN-PINCUS, M. Vulnerability of small forest patches to fire in the paraiba do sul river valley, southeast brazil: Implications for restoration of the atlantic forest biome. *Forest Ecology and Management*, Elsevier, v. 465, p. 118095, 2020.

122 SOUZA, G. de. *Monitoramento sazonal e recuperação pós-fogo da vegetação do Cerrado usando dados do sensor MODIS*. 2014. Http://jbb.ibict.br//handle/1/1070. Accessed 2020-01-13.

123 SILVA, L. G. d. *Comportamento e efeito do fogo sobre os ecossistemas do bioma cerrado: modelos baseados em processos*. 2018. Https://repositorio.unb.br/handle/10482/32603. Accessed 2020-03-06.

124 JÚNIOR, C. A. dos S.; BITTENCOURT, O. O.; MORELLI, F.; SANTOS, R. *CLASSIFICAÇÃO DE ÁREAS QUEIMADAS POR MACHINE LEARNING USANDO DADOS DE SENSORIAMENTO REMOTO*. 2019. Https://proceedings.science/sbsr-2019/papers/classificacao-de-areas-queimadas-por-machine-learning-usando-dados-de-sensoriamento-remoto. Accessed 2020-02-06.

125 PEREIRA, A. A.; PEREIRA, J.; LIBONATI, R.; OOM, D.; SETZER, A. W.; MORELLI, F.; MACHADO-SILVA, F.; CARVALHO, L. M. T. D. Burned area mapping in the brazilian savanna using a one-class support vector machine trained by active fires. *Remote Sensing*, Multidisciplinary Digital Publishing Institute, v. 9, n. 11, p. 1161, 2017.

126 SUBRAMANIAN, S.; CROWLEY, M. Using spatial reinforcement learning to build forest wildfire dynamics models from satellite images. *Frontiers in ICT*, Frontiers, v. 5, p. 6, 2018.

127 RADKE, D.; HESSLER, A.; ELLSWORTH, D. FireCast: Leveraging Deep Learning to Predict Wildfire Spread. In: . [S.l.: s.n.], 2019. p. 4575–4581.

128 ZHOU, T.; DING, L.; JI, J.; YU, L.; WANG, Z. Combined estimation of fire perimeters and fuel adjustment factors in FARSITE for forecasting wildland fire propagation. *Fire Safety Journal*, v. 116, p. 103167, 2020. ISSN 0379-7112.

129 ALLAIRE, F.; MALLET, V.; FILIPPI, J.-B. Emulation of wildland fire spread simulation using deep learning. *Neural Networks*, v. 141, p. 184–198, 2021. ISSN 0893-6080.

130 JINDAL, R.; KUNWAR, A. K.; KAUR, A.; JAKHAR, B. S. Predicting the dynamics of forest fire spread from satellite imaging using deep learning. In: IEEE. *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*. [S.l.], 2020. p. 344–350.

131 PERUMAL, R.; ZYL, T. L. van. Comparison of recurrent neural network architectures for wildfire spread modelling. In: IEEE. *2020 International SAUPEC/RobMech/PRASA Conference*. [S.l.], 2020. p. 1–6.

132 KC, U.; ARYAL, J.; HILTON, J.; GARG, S. A surrogate model for rapidly assessing the size of a wildfire over time. *Fire*, Multidisciplinary Digital Publishing Institute, v. 4, n. 2, p. 20, 2021.

133 HUOT, F.; HU, R. L.; GOYAL, N.; SANKAR, T.; IHME, M.; CHEN, Y.-F. Next day wildfire spread: A machine learning data set to predict wildfire spreading from remote-sensing data. *arXiv preprint arXiv:2112.02447*, 2021.

134 JAIN, P.; COOGAN, S. C.; SUBRAMANIAN, S. G.; CROWLEY, M.; TAYLOR, S.; FLANNIGAN, M. D. A review of machine learning applications in wildfire science and management. *Environmental Reviews*, v. 28, n. 4, p. 478–505, 2020.

135 MITRAKIS, N. E.; MALLINIS, G.; KOUTSIAS, N.; THEOCHARIS, J. B. Burned area mapping in mediterranean environment using medium-resolution multi-spectral data and a neuro-fuzzy classifier. *International Journal of Image and Data Fusion*, Taylor & Francis, v. 3, n. 4, p. 299–318, 2012.

136 OTT, C. W.; ADHIKARI, B.; ALEXANDER, S. P.; HODZA, P.; XU, C.; MINCKLEY, T. A. Predicting fire propagation across heterogeneous landscapes using wyofire: A monte carlo-driven wildfire model. *Fire*, Multidisciplinary Digital Publishing Institute, v. 3, n. 4, p. 71, 2020.

137 DUFF, T. J.; CHONG, D. M.; TOLHURST, K. G. Indices for the evaluation of wildfire spread simulations using contemporaneous predictions and observations of burnt area. *Environmental Modelling & Software*, Elsevier, v. 83, p. 276–285, 2016.

138 ELIAS, F.; JUNIOR, B. H. M.; OLIVEIRA, F. J. M. de; OLIVEIRA, J. C. A. de; MARIMON, B. S. Soil and topographic variation as a key factor driving the distribution of tree flora in the amazonia/cerrado transition. *Acta Oecologica*, Elsevier, v. 100, p. 103467, 2019.

139 Instituto Nacional de Pesquisa Espacial (INPE). *Banco de dados geomorfométricos do Brasil*. 2020. Http://www.dsr.inpe.br/topodata/dados.php. Accessed 2020-01-06.

140 SIMONS, M. Interferometric synthetic aperture radar geodesy. *Treatise on Geophysics*, Elsevier B. V, v. 3, p. 391–446, 2007.

141 RISSI, M. N.; BAEZA, M. J.; GORGONE-BARBOSA, E.; ZUPO, T.; FIDELIS, A. Does season affect fire behaviour in the cerrado? *International Journal of Wildland Fire*, CSIRO, v. 26, n. 5, p. 427–433, 2017.

142 IVO, I. O.; BIUDES, M. S.; VOURLITIS, G. L.; MACHADO, N. G.; MARTIM, C. C. Effect of fires on biophysical parameters, energy balance and evapotranspiration in a protected area in the brazilian cerrado. *Remote Sensing Applications: Society and Environment*, Elsevier, v. 19, p. 100342, 2020.

143 FIDELIS, A.; ALVARADO, S. T.; BARRADAS, A. C. S.; PIVELLO, V. R. The year 2017: Megafires and management in the cerrado. *Fire*, MDPI, v. 1, n. 3, p. 49, 2018.

144 CHAI, H.; CHENG, W.; ZHOU, C.; CHEN, X.; MA, X.; ZHAO, S. et al. Analysis and comparison of spatial interpolation methods for temperature data in xinjiang uygur autonomous region, china. *Natural Science*, Scientific Research Publishing, v. 3, n. 12, p. 999, 2011.

145 CACHUCHO, R.; MEENG, M.; VESPIER, U.; NIJSSEN, S.; KNOBBE, A. Mining multivariate time series with mixed sampling rates. In: *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. [S.l.: s.n.], 2014. p. 413–423.

146 WALTON, E.; CASEY, C.; MITSCH, J.; VÁZQUEZ-DIOSDADO, J. A.; YAN, J.; DOTTORINI, T.; ELLIS, K. A.; WINTERLICH, A.; KALER, J. Evaluation of sampling frequency, window size and sensor position for classification of sheep behaviour. *Royal Society open science*, The Royal Society Publishing, v. 5, n. 2, p. 171442, 2018.

147 SHOOK, J.; GANGOPADHYAY, T.; WU, L.; GANAPATHYSUBRAMANIAN, B.; SARKAR, S.; SINGH, A. K. Crop yield prediction integrating genotype and weather variables using deep learning. *Plos one*, Public Library of Science San Francisco, CA USA, v. 16, n. 6, p. e0252402, 2021.

148 RAJBAHADUR, G. K.; WANG, S.; ANSALDI, G.; KAMEI, Y.; HASSAN, A. E. The impact of feature importance methods on the interpretation of defect classifiers. *IEEE Transactions on Software Engineering*, IEEE, 2021.

149 ALTMANN, A.; TOLOŞI, L.; SANDER, O.; LENGAUER, T. Permutation importance: a corrected feature importance measure. *Bioinformatics*, Oxford University Press, v. 26, n. 10, p. 1340–1347, 2010.

150 ANG, J. C.; MIRZAL, A.; HARON, H.; HAMED, H. N. A. Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection. *IEEE/ACM transactions on computational biology and bioinformatics*, IEEE, v. 13, n. 5, p. 971–989, 2015.

151 DASH, M.; LIU, H. Feature selection for classification. *Intelligent data analysis*, Elsevier, v. 1, n. 1-4, p. 131–156, 1997.

152 ZHAI, Y.; SONG, W.; LIU, X.; LIU, L.; ZHAO, X. A chi-square statistics based feature selection method in text classification. In: IEEE. *2018 IEEE 9th International conference on software engineering and service science (ICSESS)*. [S.l.], 2018. p. 160–163.

153 URBANOWICZ, R. J.; MEEKER, M.; CAVA, W. L.; OLSON, R. S.; MOORE, J. H. Relief-based feature selection: Introduction and review. *Journal of biomedical informatics*, Elsevier, v. 85, p. 189–203, 2018.

154 KUMARI, B.; SWARNKAR, T. Filter versus wrapper feature subset selection in large dimensionality micro array: A review. Citeseer, 2011.

155 BERGSTRA, J.; BENGIO, Y. Random search for hyper-parameter optimization. *Journal of machine learning research*, v. 13, n. 2, 2012.

156 HOSSIN, M.; SULAIMAN, M. N. A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process*, Academy & Industry Research Collaboration Center (AIRCC), v. 5, n. 2, p. 1, 2015.

157 KEANE, R. E.; KARAU, E. Evaluating the ecological benefits of wildfire by integrating fire and ecosystem simulation models. *Ecological Modelling*, Elsevier, v. 221, n. 8, p. 1162–1172, 2010.

158 VALERO, M. M.; RIOS, O.; MATA, C.; PASTOR, E.; PLANAS, E. An integrated approach for tactical monitoring and data-driven spread forecasting of wildfires. *Fire safety journal*, Elsevier, v. 91, p. 835–844, 2017.

159 ARCA, B.; GHISU, T.; CASULA, M.; SALIS, M.; DUCE, P. A web-based wildfire simulator for operational applications. *International journal of wildland fire*, CSIRO, v. 28, n. 2, p. 99–112, 2019.

160 NEMBRINI, S.; KÖNIG, I. R.; WRIGHT, M. N. The revival of the gini importance? *Bioinformatics*, Oxford University Press, v. 34, n. 21, p. 3711–3718, 2018.

161 JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. *An introduction to statistical learning*. [S.l.]: Springer, 2013. v. 112.

162 BENESTY, J.; CHEN, J.; HUANG, Y.; COHEN, I. Pearson correlation coefficient. In: *Noise reduction in speech processing*. [S.l.]: Springer, 2009. p. 1–4.

163 MILES, J. Tolerance and variance inflation factor. *Wiley statsref: statistics reference online*, Wiley Online Library, 2014.

164 GIANNAROS, T. M.; LAGOUVARDOS, K.; KOTRONI, V. Performance evaluation of an operational rapid response fire spread forecasting system in the southeast mediterranean (greece). *Atmosphere*, Multidisciplinary Digital Publishing Institute, v. 11, n. 11, p. 1264, 2020.

APPENDICES

A - ARTICLE PUBLISHED IN INTERNATIONAL JOURNAL OF COMMUNICATION SYSTEMS.

IoT-based Platform for Environment Data Sharing in Smart Cities

Jesús Noel Suárez Rubí[a], Paulo Roberto de Lira Gondim[a,*]

[a] *Faculdade de Tecnologia, Campus Universitário Darcy Ribeiro, Asa Norte, Brasilia, DF, Brazil*

Abstract

The technological development and dissemination of IoT equipment have led to large volumes of environmental data which, in some cases, are incomplete, follow different formats of representation and even have different semantic approaches. All such aspects and the heterogeneity of different IoT components (e.g. network interfaces, communication protocols, data structure and data semantics) have caused interoperability issues which might hamper the effectiveness of support decision systems for smart cities, where the use of big data and machine learning techniques has been considered, in addition to the exploration of smart city data. This article proposes an environment IoT-based platform for smart cities that grants interoperability from data capture to knowledge extraction and visualization through the use of Semantic Web Technologies, and the definition of an ontology for environment indicators. The components of the platform include IoT devices, gateways, cloud and fog computing, which are used for a better application of big data techniques. A real environment quality monitoring use case was considered for the validation of the platform. Metrics, such as latency and resources consumption, were analyzed for three communication protocols, namely MQTT, CoAP and REST. CoAP adapter provided the best results regarding latency, RAM and CPU consumption.

Keywords: Smart City, Linked Data, Platform, Environment, Ontologies

1. Introduction

The rapid growth of urban populations has demanded studies that identify, prevent and act in situations of threatened quality of life (QoL) [1]. In Smart Cities (SC) [2], QoL is commonly dealt with by indicators that measure the effectiveness of services and sustainability of a city in domains/verticals, such as Environment, Healthcare, Security, Transport, Economy, Education and Government. Particularly, the Environment vertical has drawn special attention in recent years. Indicators of environmental pollutants (e.g., atmospheric greenhouse gases, fine particular matter PM, noise, solid waste, among others) and water (acidity and mercury) must be monitored for the detection of adverse situations associated with overpopulated regions. In this sense, SCs [2] must provide interoperable tools that collect, store and disseminate indicator-related data, and several sensors, frameworks and SC platforms have emerged for such purposes. However, some challenges still hamper better management and analyses of data of those indicators.

[*] Corresponding author
Email address: pgondim@unb.br (Paulo Roberto de Lira Gondim)

B - ARTICLE PUBLISHED IN COMPUTERS AND ELECTRICAL ENGINEERING JOURNAL

Forestry 4.0 and Industry 4.0: Use Case on Wildfire Behaviour Predictions

Jesús N. S. Rubí[a], Paulo H. P. de Carvalho[a], Paulo R. L. Gondim[a]

[a] *Department of Electrical Engineering, Faculty of Technology, University of Brasilia, Asa Norte, Brasilia, 70910-900, DF, Brazil*

Abstract

Forest industries deserve special attention due to relations between environmental impact and social and economic development. The increase of forest fires caused by the untenable exploitation has motivated the application of concepts such as Industry/Forestry 4.0 and Internet of Forest Things (IoFT) towards improving the performance of current supply chains and assuming an environmental responsibility. This research focuses on the application of IoFT for the prediction of wildfires behavior and proposes a semantic platform for heterogeneous IoFT data aggregation that grants interoperability through semantic technologies. The dataset considered climatic- and vegetation-related data gathered by Brazilian government sensors and satellite information on fires, and Machine Learning predicted the areas affected after a fire event. Both platform and predictions were validated and Random Forest predicted the area with 89% accuracy, showing better performance than Deep Neural Network, with 79%.

Keywords: , Industry 4.0, Forestry 4.0, Semantic, Platform, IoFT, Ontology, Wildfires, Machine Learning.

1. Introduction

The social-economic development of countries with significant forest resources depends on an adequate exploration that preserves the environment and the life in its several expressions. Such an exploration/preservation balance has been threatened by several issues related to population growth and the global competition in world markets [1]. In this sense, the assurance of environmental sustainability of production processes and preservation of ecosystems and natural resources will be challenging in the coming years.

Forest fires are one of such cases in which the unsustainable exploitation of resources causes the degradation of biodiversity and environment. As an example, the wildfires growth in Brazil has been correlated to forest industry and production chains such as firewood, coal, solid wood, paper and cellulose; therefore, tools that minimize the environmental downside effects exerted by the aggressive nature of such industry should be studied.

The increasing number of fire spots in the Brazilian Federal District (FD), located inside the Cerrado biome, highlights the importance of fire-related studies on the region. The Cerrado, with 60.5% of natural vegetation cover, is one of the most important worldwide biomes, given its rich biodiversity. Its long dry periods and vegetation type and the extensive exploration for the production of charcoal, firewood, and paper have led to conditions that stimulate the spread of fires.

Preprint submitted to Computers & Electrical Engineering *June 28, 2022*

C - ARTICLE PUBLISHED IN ENGINEERING APPLICATIONS OF ARTIFICIAL INTELLIGENCE JOURNAL

Application of machine learning models in the behavioral study of forest fires in the Brazilian Federal District region

Jesús N. S. Rubí[a,*], Paulo H. P. de Carvalho[a], Paulo R. L. Gondim[a]

[a]*Department of Electrical Engineering, Faculty of Technology, University of Brasilia, Zip code 70910-900, Brasilia DF, Brazil*

Abstract

Ecosystems, settlements, and human lives are put at risk by forest fires every year. Several models proposed for the prediction of their occurrence and behavior have aimed at identifying their conditioning factors, risks, and post-effects. However, the application of such models to other regions is impracticable or very difficult, due to the distinct geographic characteristics of the areas and the unavailability of data. This research is devoted to the prediction of both spread and behavior of wildfires at a specific time and/or in specific regions for helping fire management agencies minimize the damages caused. The Brazilian Federal District, inserted in the Cerrado biome, is the focus of the analyses, due to its large number of fire occurrences and reduced quantity of studies conducted on the region. A dataset was compiled from Brazilian governmental open data for the prediction of the wildfire behavior and used for the training of several Machine Learning models that consider the fire point of ignition to predict the areas that will be impacted. It includes observations on climate features from 5 monitoring stations and satellite data on fires that occurred over the past two decades and was enriched with other topographic, hydrographic, and anthropogenic features, such as urbanization index, distance to rivers/roads, and Normalized Difference Vegetation Index (NDVI). According to the results, the AdaBoost model predicted the area affected by the wildfire with 91% accuracy, showing better performance than Random Forest (RF) 88%, Artificial Neural Network (ANN) 86%, and Support Vector Machine (SVM) 81%.

Keywords: Wildfires Fires, Behaviour, Spread, Machine Learning, Classification, Prediction, Performance

1. Introduction

Forest fires have increased worldwide, and improvements in their modeling are a key aspect for a better understanding of procedures for their prevention and combat (Monedero et al. (2019)). The life quality of the inhabitants of the Brazilian Federal District (FD) region has been negatively affected by the fire activity, which has led the government to spend resources on their fighting. The past decades have witnessed an increment in the number of fire spots, according to satellite data, thus highlighting the importance of fire-related studies on the region (see Figure 1).

The FD region is inserted in the Brazilian savanna (the Cerrado biome), comprised of 11,627 species of plants and which has been affected by a large number of fires. The dry climate together with the savanna vegetation create a favorable scenario for fire dissemination. Therefore, research into forest fires in the DF will both leverage the local firefighting decision making and policies and probably decrease the number of fires in the Cerrado region.

Few investigations aimed at understanding forest fires in the DF region have been conducted (de Bem et al. (2019)). Gomes et al. (2018) identified three major challenges, expressed on three different scales, namely predictive (multiple drivers must be considered for studies of fire), 2) spatial (changes that occur from site (local) to biome level), and 3) temporal (changes from short to long-term dynamics). The authors also reported a lack of proposals on the three scales that consider the joint simulations of fire risk, behaviour, and impacts.

This manuscript focuses on the predictions of the fire-line behaviour given a point of ignition and climate, topographic, hydrographic, and anthropogenic data. The proposal determines the direction and extension of the fire-line and evaluates the predicted fire scars.

No consensus on modelling methodologies of forest fire behaviour has been achieved. Although approaches involving complex mathematical models have been published (Jiang et al. (2021), Rossa and Fernandes (2018) and Chen et al. (2018)), their static characteristics hamper the representation of highly dynamic processes such as fire-line progress. Most of those empirical and semi-empirical models have been applied in laboratories and controlled field-scale experiments, which commonly consider two types of numerical approaches. The first is based on the complex modelling of physical and chemical processes (Chen et al. (2018), Mueller et al. (2014)), whereas the other involves the rate of spread-correlating features, such as slope, wind, and vegetation type (Zhai et al. (2020)). However, both have showed poor accuracy in real fire events and required high computational costs and simulation times, which are impractical for real-time decision support.

The reproduction of prediction results is usually difficult

[*]Corresponding author
Email address: nsuarezrubi@gmail.com (Jesús N. S. Rubí)

D - ARTICLE SUBMITTED TO ENVIRONMENT SYSTEMS AND DECISIONS JOURNAL

A Performance Comparison of Machine Learning Models for Wildfire Risk Prediction in the Brazilian Federal District Region

J. N. S. Rubí[a], Paulo R. L. Gondim[a]

[a]*Department of Electrical Engineering, Faculty of Technology, University of Brasilia, Asa Norte, Brasilia, 70.910-900, Federal District, Brazil*

Abstract

Despite a steady increase in the frequency of wildfires and the total area burned in Brazil, such hazards have risen in the Federal District, since wildfires have reached several of its protected areas. This study compares 8 machine learning models that predict wildfire risk worldwide so that they can be adopted in the aforementioned region, considering correlations among climate conditions, spatial location, topographic features, anthropogenic characteristics, and fire occurrence. A dataset enriched with Brazilian governmental open data was comprised of observationson 16 climate features of 5 monitoring stations and satellite data on fires occurred over the past two decades and topographic, hydrographic and anthropogenic features, such as Normalized Difference Vegetation Index (NDVI), urbanization index, and distance to rivers/roads. According to the results, fire risk can be predicted with 99% accuracy and the models showed more sensitive to NDVI, atmospheric pressure, and relative humidity, as demonstrated by a study on the impactof features.

Keywords: Wildfires, Fires, Risk, Machine Learning, Classification, Prediction, Performance, SVM, Neural Networks, Random Forest, Naive Bayes, KNN, Linear Regression, Logistic Regression, AdaBoost, NDVI, Climatic

1. Introduction

Global climate changes and the risk of forest ecosystems have been the focus of recent discussions. In recent years, countries such as Australia, South Africa, and Brazil have drawn worldwide attention, since they have caused losses in forested areas and human and economic damages and devastated biodiversity, thus leading to the development of more studies on wildfire [1].

This research was motivated by the increased fire activity in the Brazilian Federal District (FD) region observed since the year 2000 [2], which has affected its native species, even in protected zones [3], led to a poorer quality of life of its inhabitants, and caused local governments to spend resources on fighting fires.

Few studies on the prediction of wildfires in that area, inserted in the Brazilian savanna (the Cerrado biome), have been developed [4], and a model for the prediction of fire occurrence would help institutions to reduce the costs of fire combat and its consequences.

Machine Learning (ML) techniques that process large volumes of fire-related data can be designed from the current computational resources. The several studies that have proposed models for wildfire risk prediction can be supported by free remotely sensed data for precisely locating

December 3, 2022