Avaluació de models basats en intel·ligència artificial per a la predicció espacial del risc d'incendi forestal

Fidel Bonet Vilela

Màster Universitari en Enginyeria Informàtica
Àrea d'Intel·ligència Artificial
Professor col·laborador: Dr. Samir Kanaan Izquierdo
Professor: Dr. Carles Ventura Royo

31 de desembre de 2019

FITXA DEL TREBALL FINAL

Títol del treball	*Avaluació de models basats en intel·ligència artificial per a la predicció espacial del risc d'incendi forestal.*
Nom de l'autor	*Fidel Bonet Vilela*
Nom del consultor/a	*Dr. Samir Kanaan Izquierdo*
Nom del PRA	*Dr. Carles Ventura Royo*
Data de lliurament (mm/aaaa)	*12/2019*
Titulació o programa	*Màster Universitari en Enginyeria Informàtica*
Àrea del Treball Final	*Intel·ligència artificial*
Idioma del treball	*Català*
Paraules clau	*Aprenentatge automàtic, sistemes d'informació geogràfica, big data*

Resum del Treball

L'objectiu d'aquest treball final de màster és l'obtenció de models de predicció espacial del risc d'incendi forestal mitjançant l'ús combinat d'intel·ligència artificial, sistemes d'informació geogràfica i big data.

S'han utilitzat diversos conjunts de dades d'incendis, meteorològiques, orogràfiques i de vegetació per tal de predir les zones de risc d'incendi i estimar-ne la probabilitat mitjançant tècniques d'aprenentatge automàtic: classificació per a estimar el risc d'incendi forestal, regressió per a predir la mida dels incendis en el moment d'ignició i agrupament per a obtenir zones de risc d'incendi segons les condicions meteorològiques.

Per altra banda, el projecte ha permès identificar els atributs amb més pes alhora de predir el risc d'incendi i obtenir models robustos amb precisions de fins al 90% en la predicció del risc i del 99% en l'agrupament de nous exemples en categories de risc en funció de les condicions meteorològiques.

Així mateix, els millors resultats s'han obtinguts amb l'ús d'aprenentatge profund. Concretament, s'han utilitzat algorismes genètics per tal d'optimitzar l'arquitectura d'un perceptró multicapa en l'estimació del risc.

En darrer lloc, els resultats del projecte permeten obtenir mapes de risc amb prou detall per a diversos àmbits (comarques, municipis, espais naturals, etc.) i vàlids per àrees concretes com ara un parc natural on els resultats assolits han permès estimar el risc d'incendi per a les diverses zones del parc i, fins i tot, en determinats indrets sensibles com, per exemple, els principals senders i les zones d'estacionament de vehicles.

Abstract

The goal of this final master's project is to obtain spatial prediction models of forest fire risk through the combined use of artificial intelligence, geographic information systems and big data.

Several sets of fire, meteorological, orographic and vegetation data have been used to predict fire risk areas and to estimate the probability by means of several machine learning techniques: classification to predict the risk of forest fire, regression to predict the size of the fires at the moment of ignition and clustering to obtain fire risk areas according to the meteorological conditions.

On the other hand, the project has made it possible to identify the attributes with the greatest weight when predicting fire risk and to obtain robust models with 90% accuracy in predicting risk and 99% in grouping new examples into risk categories according to weather conditions.

Furthermore, the best results have been obtained with the use of deep learning. Specifically, genetic algorithms have been used to optimize the architecture of a multilayer perceptron in risk estimation.

Finally, the results of the project allow us to obtain risk maps with sufficient detail for various areas (counties, municipalities, natural spaces, etc.) and valid for specific areas such as a natural park where the results achieved have allowed us to estimate the wildfire risk for the various areas of the park and even in certain sensitive places such as the main paths and vehicle parking areas.

ÍNDEX

LLISTAT DE TAULES

LLISTAT D'IMATGES

1 Introducció

Tot seguit s'analitza tant el context i la justificació d'aquest treball, com els seus principals objectius, el mètode seguit i la planificació del treball. Per acabar, s'enumeren breument els productes obtinguts i la descripció dels següents capítols d'aquesta memòria.

1.1 Context i justificació del treball

En aquest treball final de màster es pretén analitzar l'ús de diversos conjunts de dades meteorològiques, orogràfiques, de vegetació i d'incendis per a l'obtenció de models de predicció espacial del risc d'incendi forestal en el territori de Catalunya mitjançant l'ús de tècniques d'aprenentatge automàtic, anàlisi geoespacial i *big data*. Els models obtinguts i els mapes que se'n derivin haurien de poder ser utilitzats per a l'alerta primerenca, la planificació dels recursos i l'assignació de tasques en l'extinció d'incendis.

Quant a la rellevància per a la societat del camp d'estudi escollit, aquesta rau en el gran impacte dels incendis forestals. Pourghasemi (2014) i Coffield (2019) en destaquen diversos àmbits: en primer lloc, els grans danys al medi ambient, com ara la contribució al cicle global del carboni, el qual es relaciona amb el canvi climàtic i amb la fosa de neu i gel a altes latituds; en segon lloc, els danys a la salut humana, en particular l'augment de les morts prematures i de les hospitalitzacions en pacients amb malalties respiratòries i cardiovasculars; en tercer lloc, els danys a la propietat, sovint amb importants perjudicis econòmics, i, finalment, el perill per a la vida de les persones.

Per altra banda, l'impacte global dels incendis forestals ha crescut darrerament (Jaafari, 2019) i es preveu que, a causa del canvi climàtic, continuï incrementant-se en els propers anys. Com assenyala Radke (2019), s'espera que n'augmenti: la durada, la freqüència i la severitat. Concretament, l'ONU apunta en el seu cinquè informe d'avaluació del canvi climàtic un augment significatiu de l'extensió i la freqüència dels incendis al sud d'Europa, sobretot a la conca mediterrània, a partir de la dècada del 1970 en relació a les dècades anteriors i n'assenyala els següent motius: l'acumulació de combustible, el canvi climàtic i els esdeveniments climàtics extrems (IPCC, 2014).

En particular, pel que fa al territori català, segons apunten investigadors del Centre de Recerca Ecològica i Aplicacions Forestals (Ramon, 2019) s'espera l'augment dels grans incendis forestals degut sobretot als següents factors: l'increment d'un 10% de la superfície forestal en els seixanta darrers anys, el canvi d'ús de la terra, l'escassa gestió forestal i el canvi climàtic. En particular, el canvi climàtic es relaciona amb: l'augment de vents càlids, la redistribució de les precipitacions, l'increment de les temperatures i l'assecament de la vegetació.

Per aquest fet, disposar de models que permetin estimar la probabilitat dels incendis així com la seva evolució i dimensions, facilitaria tant la gestió del territori com dels propis incendis per tal d'evitar o disminuir els danys en el territori, les zones residencials i la vida dels equips d'extinció i també reduir els costos de l'extinció (Sayad, 2019). Per altra banda, disposar d'eines per a la predicció de les zones amb més risc permetria als bombers la realització de cremes controlades i evitar-ne la progressió tallant l'avenç en determinades àrees (Rolnick, 2019).

Amb aquest objectiu, en aquest projecte es proposa l'ús de la intel·ligència artificial i, més concretament, de l'aprenentatge automàtic per a l'estimació del risc d'incendi ja que aquesta tècnica permet obtenir models útils amb una dificultat i cost computacional inferior als

necessaris en altres tècniques com son els models matemàtics i les simulacions.

D'altra banda, els sistemes d'informació geogràfica (SIG) permeten la preparació dels atributs espacials necessaris per a l'obtenció dels models d'aprenentatge així com la millora de la comprensió dels patrons espacials dels processos que controlen els incendis (Jaafari, 2019). És per aquest motiu que en el TFM es proposa utilitzar de forma conjunta ambdues disciplines, intel·ligència artificial i tècniques geoespacials, per a l'estimació del risc d'incendi forestal.

A més, malgrat que les condicions meteorològiques tenen un impacte molt gran en el risc d'incendi, sovint la informació que se'n recull per a l'entrenament dels models d'aprenentatge automàtic és limitada. És per això que es proposa utilitzar no només dades meteorològiques del dia d'inici de l'incendi sinó també dels dies anterior i posterior així com el conjunt de dies amb condicions meteorològiques similars, per, d'aquesta forma, aportar més informació als procés d'aprenentatge.

Per altra banda, un segon conjunt d'atributs decisius per a estimar el risc d'incendi son els topogràfics, sobretot: altitud, orientació i pendent. Avui dia, per a la prevenció d'incendis forestals a Catalunya no s'utilitza un únic model que inclogui també aquests tipus d'atributs. Per tant, es proposa desenvolupar un model d'aprenentatge automàtic que tingui en compte paràmetres orogràfics amb prou detall.

En últim terme, també es planteja l'ús de tècniques *big data* pel gran volum de dades que intervenen en el la fase d'aprenentatge dels models predictius.

Convé destacar que, per bé que s'utilitzaran dades del territori català per a l'entrenament dels models de predicció, es tindrà en compte en tot moment que les solucions proposades puguin ser aplicades a altres territoris.

1.2 Objectius del treball

A continuació s'exposen els objectius generals i específics del projecte per a, tot seguit, analitzar-ne la priorització.

D'una banda, s'han establert dos objectius generals: la utilització d'algorismes d'aprenentatge automàtic per a la predicció d'incendis forestals i l'obtenció de models específics per al territori català. El primer s'ha considerat més prioritari que el segon.

Objectiu general	Descripció	Prioritat
G1	Utilització d'algorismes d'aprenentatge automàtic per a l'obtenció de models predictius de risc d'incendi forestal.	Alta
G2	Obtenció d'un model de predicció de risc d'incendi forestal per al territori català.	Mitjana

Taula 1. Prioritització dels objectius generals.

D'altra banda, s'han fixat nou objectius específics. Per una banda, els principals son el tractament mitjançant sistemes d'informació geogràfica de sèries espai-temporals i, específicament, de dades topogràfiques per a l'entrenament de models d'aprenentatge

automàtic, la comparació de diverses tècniques d'intel·ligència artificial per a la predicció del risc d'incendi i, finalment, l'obtenció de models predicatius. D'altra banda, en un segon nivell de prioritat s'hi troben la utilització de tecnologia *big data* per al tractament de dades espai-temporals i, concretament en el marc de treball Apache Spark, la valoració de l'ús de dades meteorològiques del dia previ i posterior als incendis i l'ús d'eines geoespacials per al tractament i la interpretació de les prediccions obtingudes. En darrer lloc, s'ha considera un objectiu de prioritat baixa l'obtenció de models de regressió per a l'obtenció de la mida final dels incendis en el moment de la ignició, ja que es prioritza l'obtenció de models d'estimació de la probabilitat d'aquests. La següent taula mostra aquesta priorització dels objectius específics.

Objectiu específic	Descripció	Prioritat
E1	Tractament mitjançant sistemes d'informació geogràfica de sèries espai-temporals per a l'entrenament d'algorismes d'aprenentatge automàtic.	Alta
E2	Utilització de dades orogràfiques per a l'obtenció de models de predicció del risc d'incendi forestal.	Alta
E3	Utilització de tecnologia *big data* per al tractament de grans volums de dades espai-temporals.	Mitjana
E4	Utilització d'algorismes d'aprenentatge automàtic en el marc de treball Apache Spark.	Mitjana
E5	Obtenció de models predictius de risc d'incendis forestals.	Alta
E6	Obtenció de models de regressió per a l'estimació de la mida final dels incendis en el moment de la ignició.	Baixa
E7	Comparació de tècniques d'aprenentatge automàtic per a la predicció del risc d'incendi forestal.	Alta
E8	Valoració de la utilitat de l'ús de dades meteorològiques del dia previ i posterior als incendis en els models d'aprenentatge.	Mitjana
E9	Utilització d'eines geoespacials per al tractament i interpretació dels resultats obtinguts.	Mitjana

Taula 2. Priorització dels objectius específics.

1.3 Enfocament i mètode seguit

Aquest treball s'ha desenvolupat principalment utilitzant llibreries d'aprenentatge automàtic en el llenguatge Python tant en un entorn local com en l'entorn de treball per a computació distribuïda Apache Spark.

En primer lloc, s'han analitzat mètodes de categorització per a l'agrupament de dies segons les seves característiques meteorològiques i la seva relació amb els incendis forestals. Tot seguit, s'han analitzat mètodes de regressió per a l'estimació de la mida final dels incendis en el moment de la ignició. Per acabar, s'han valorat mètodes de classificació per a l'estimació del risc d'incendi. Així mateix, en tots tres casos se n'ha avaluat els resultats obtinguts i la seva adequació al problema plantejat. En darrer lloc, s'han implementat els resultats obtinguts per a generar mapes de risc d'incendi forestal.

1.4 Planificació del Treball

En la planificació del treball s'ha tingut en compte els recursos de maquinari, programari i dades necessaris, les tasques indispensables per assolir els objectius plantejats així com la planificació temporal d'aquestes.

1.4.1 Recursos

Els recursos necessaris per a la realització del projecte han estat, sobretot: dades, maquinari i programari.

Dades

S'ha utilitzat dades d'incendis forestals, orogràfiques, meteorològiques i de vegetació, usos i cobertes del sòl tant generades en el propi projecte com obtingudes de diverses fonts: el Servei de prevenció d'Incendis Forestals i el Departament de Territori i Sostenibilitat de la Generalitat de Catalunya, el Servei Meteorològic de Catalunya, l'Institut Cartogràfic i Geològic de Catalunya, l'ECMWF, el Climate Data Store i la NASA. En l'apartat 6 es descriuen en detall les dades utilitzades.

Cap de les anteriors dades està protegida per les lleis de protecció de dades de caràcter personal. Tanmateix, les dades d'incendis forestals tenen la consideració de dades confidencials.

Maquinari

En la següent taula es detallen els recursos de maquinari emprats en el desenvolupament del projecte.

Característica	Descripció
Processador	3,06 GHZ Intel Core 2 Duo
Memòria RAM	4 GiB
Sistema operatiu	macOS High Sierra, versió 10.13.6

Taula 3. Característiques del maquinari utilitzat.

Programari

En darrer lloc, la següent taula mostra un resum del principal programari utilitzat per a la realització del projecte on es descriu la funció a la qual s'ha destinat.

Programari	Principals biblioteques	Funció	Versió
Conda		Gestió d'entorns de codi obert	4.6.14
Spyder		Entorn integrat de desenvolupament per al llenguatge de programació Python	3.3.1
Python		Implementació d'algorismes d'aprenentatge automàtic, tractament dels conjunts de dades, tractament de dades geoespacials i generació de gràfics	3.5
	NumPy	Computació científica	1.11.0

	Pandas	Manipulació i anàlisi de dades	1.12.0
	Scipy	Computació científica	0.17.0
	Scikit-learn	Aprenentatge automàtic	0.21.3
	Matplotlib	Visualització de dades	1.12.0
	Seaborn	Visualització de dades	0.8.1
	Keras	Implementació de xarxes neuronals	1.0.2
	Tensorflow	Implementació d'algorismes d'aprenentatge profund	1.10.0
	DEAP	Entorn de computació evolutiu. Utilitzat per a la implementació d'algorismes genètics	1.3.0
Apache Spark		Implementació d'algorismes d'aprenentatge en un entorn distribuït	0.2.1
	Pyspark	API Python per a Apache Spark	2.4.4
	Elephas	Extensió de Keras que permet l'execució de models d'aprenentatge profund en Spark	0.4.2
	MLLib	Aprenentatge automàtic	1.0.0
QGIS		Tractament de dades espacial. Generació de mapes.	3.4
	PyQGIS	API Python de QGIS	3.0

Taula 4. Programari utilitzat.

1.4.2 Descripció de les tasques.

A partir dels objectius específics exposats anteriorment s'han obtingut les tasques principals del treball. La següent taula en mostra les seves descripcions així com la relació de cadascuna amb l'objectiu corresponent. Així mateix, per a ordenar-les s'ha tingut en compte la priorització detallada en l'apartat 1.2 per tal de dur a terme abans les prioritàries sempre que no hi hagués una dependència entre tasques que ho impedís.

Objectiu específic		Tasca	Descripció
E1. Tractament mitjançant sistemes d'informació geogràfica de sèries espai-temporals per a l'entrenament d'algorismes d'aprenentatge automàtic.	T1.1	Recol·lecció de dades	*Recol·lecció de dades topogràfiques, meteorològiques, de vegetació i incendis.*
	T1.2	Processament de dades	*Processament geoespacial de les dades per a l'obtenció dels conjunts de dades.*
	T1.3	Anàlisi estadístic previ de les dades disponibles	*Anàlisi estadístic dels diversos conjunts de dades previ a l'entrenament.*
	T1.4	Anàlisi de correlació de les dades	*Anàlisi de la correlació de les dades i reducció de la dimensionalitat.*
	T1.5	Obtenció dels conjunts de dades	*Obtenció de conjunts de dades adaptats a les diverses tècniques d'aprenentatge. Obtenció de conjunts reduïts de prova.*
E2. Utilització de dades orogràfiques per a l'obtenció de models de predicció del risc d'incendi forestal.	T2.1	Recol·lecció de dades topogràfiques	*Recol·lecció de dades topogràfiques.*
	T2.2	Processament de dades	*Processament geoespacial de les dades.*
	T2.3	Obtenció de dades derivades	*Obtenció d'orientació i pendent del terreny*
E3. Utilització de tecnologia *big data* per al tractament de grans volums de dades espai-temporals.	T3.1	Anàlisi d'alternatives d'entorns *big data*	*Anàlisi de les diverses alternatives big data disponibles per a l'aplicació de tècniques d'aprenentatge automàtic a grans conjunts de dades*
	T3.2	Preparació de l'entorn *big data*	*Preparació de l'entorn big data escollit per a l'obtenció dels models d'aprenentatge.*

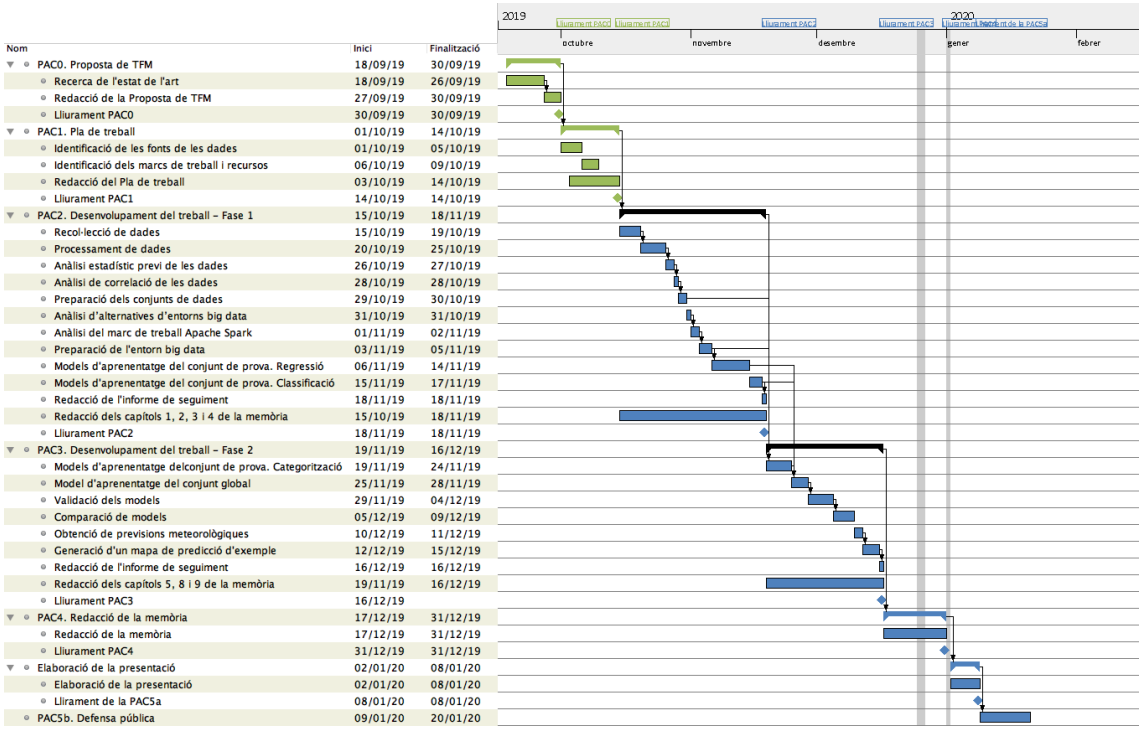
E4. Utilització d'algorismes d'aprenentatge automàtic en el marc de treball Apache Spark.	T4.1	Anàlisi del marc de treball Apache Spark	*Estudi del marc de treball Apache Spark i la seva adequació al projecte.*
	T4.2	Implementació dels algorismes en l'entorn Apache Spark	*Implementació dels diversos algorismes d'aprenentatge automàtic en l'entorn Apache Spark*
E5. Obtenció de models predictius de risc d'incendis forestals.	T5.1	Obtenció dels models d'aprenentatge. Classificació.	*Obtenció de models de predicció de la probabilitat d'incendi forestal.*
	T5.2	Obtenció dels models d'aprenentatge. Categorització	*Obtenció de zones amb condicions meteorològiques anàlogues per a la generació d'indicadors de risc d'incendi segons l'històric d'incendis en situacions similars.*
E6. Obtenció de models de regressió per a l'estimació de la mida.	T6.1	Obtenció dels models d'aprenentatge. Regressió.	*Obtenció de models de regressió per a l'estimació de la mida final dels incendis en el moment de la ignició.*
E7. Comparació de tècniques d'aprenentatge automàtic per a la predicció del risc d'incendi forestal.	T7.1	Obtenció del model d'aprenentatge per al conjunt final	*Obtenció dels models d'aprenentatge sobre el conjunt de dades final de tot el territori català.*
	T7.2	Validació dels models	*Utilització de diverses tècniques de validació de models d'aprenentatge.*
	T7.3	Comparació de models	*Comparació dels diversos models d'aprenentatge automàtic obtinguts.*
E8. Valoració de la utilitat de l'ús de dades meteorològiques del dia previ i posterior als incendis en els models d'aprenentatge.	T8.1	Comparació de models en funció del paràmetres meteorològiques utilitzats	*Comparació dels diversos models d'aprenentatge automàtic obtinguts amb la finalitat d'avaluar la idoneïtat de tenir en compte les condicions tant del moment de l'inici de l'incendi com les condicions prèvies i posteriors.*
E9. Utilització d'eines geoespacials per al tractament i interpretació dels resultats obtinguts.	T9.1	Obtenció de previsions meteorològiques	*Obtenció de dades de previsions meteorològiques per a generar nous casos que permetin l'ús dels models obtinguts.*
	T9.2	Generació d'un mapa de predicció d'exemple	*Generació d'un mapa de predicció d'exemple a partir dels models de predicció obtinguts i de dades de predicció meteorològica.*
Preparació del lliurament final i defensa	T10.1	Redacció de la memòria	*Redacció del contingut de la memòria.*
	T10.2	Preparació de la presentació	*Preparació de la presentació.*
	T10.3	Preparació del lliurament	*Preparació del lliurament: memòria, codi i presentació.*
	T10.4	Defensa pública	*Defensa del treball final de màster.*

Taula 5. Tasques.

1.4.3 Planificació temporal.

El següent diagrama de Gantt mostra la planificació temporal de les tasques, les seves dependències i les fites corresponents a lliuraments. Com es pot observar, s'ha programat la realització d'algunes de les tasques en paral·lel: la redacció del Pla de treball al mateix temps que s'identifiquen les fonts de les dades així com els marcs de treball i els recursos disponibles; la redacció dels capítols 1, 2, 3 i 4 de la memòria mentre es desenvolupa la resta d'activitats de la Fase 1, i la redacció dels capítols 5, 8 i 9 mentre es desenvolupen les activitats de la Fase 2.

Per altra banda, s'han considerat tres dies no hàbils per a la realització del treball: 25 i 26 de desembre i 1 de gener.



Nom	Inici	Finalització
PAC0. Proposta de TFM	18/09/19	30/09/19
Recerca de l'estat de l'art	18/09/19	26/09/19
Redacció de la Proposta de TFM	27/09/19	30/09/19
Lliurament PAC0	30/09/19	30/09/19
PAC1. Pla de treball	01/10/19	14/10/19
Identificació de les fonts de les dades	01/10/19	05/10/19
Identificació dels marcs de treball i recursos	06/10/19	09/10/19
Redacció del Pla de treball	03/10/19	14/10/19
Lliurament PAC1	14/10/19	14/10/19
PAC2. Desenvolupament del treball – Fase 1	15/10/19	18/11/19
Recol·lecció de dades	15/10/19	19/10/19
Processament de dades	20/10/19	25/10/19
Anàlisi estadístic previ de les dades	26/10/19	27/10/19
Anàlisi de correlació de les dades	28/10/19	28/10/19
Preparació dels conjunts de dades	29/10/19	30/10/19
Anàlisi d'alternatives d'entorns big data	31/10/19	31/10/19
Anàlisi del marc de treball Apache Spark	01/11/19	02/11/19
Preparació de l'entorn big data	03/11/19	05/11/19
Models d'aprenentatge del conjunt de prova. Regressió	06/11/19	14/11/19
Models d'aprenentatge del conjunt de prova. Classificació	15/11/19	17/11/19
Redacció de l'informe de seguiment	18/11/19	18/11/19
Redacció dels capítols 1, 2, 3 i 4 de la memòria	15/10/19	18/11/19
Lliurament PAC2	18/11/19	18/11/19
PAC3. Desenvolupament del treball – Fase 2	19/11/19	16/12/19
Models d'aprenentatge delconjunt de prova. Categorització	19/11/19	24/11/19
Model d'aprenentatge del conjunt global	25/11/19	28/11/19
Validació dels models	29/11/19	04/12/19
Comparació de models	05/12/19	09/12/19
Obtenció de previsions meteorològiques	10/12/19	11/12/19
Generació d'un mapa de predicció d'exemple	12/12/19	15/12/19
Redacció de l'informe de seguiment	16/12/19	16/12/19
Redacció dels capítols 5, 8 i 9 de la memòria	19/11/19	16/12/19
Lliurament PAC3	16/12/19	
PAC4. Redacció de la memòria	17/12/19	31/12/19
Redacció de la memòria	17/12/19	31/12/19
Lliurament PAC4	31/12/19	31/12/19
Elaboració de la presentació	02/01/20	08/01/20
Elaboració de la presentació	02/01/20	08/01/20
Llirament de la PAC5a	08/01/20	08/01/20
PAC5b. Defensa pública	09/01/20	20/01/20

Imatge 1. Planificació inicial.

1.4.4 Avaluació de riscos.

S'han identificat un total de cinc riscos principals: manca de les dades necessàries per a l'entrenament dels models d'aprenentatge, dificultats en el desplegament en l'entorn *big data*, dificultats amb el marc Apache Spark, el gran nombre de mètodes d'aprenentatge automàtic utilitzats i les situacions laborals i personals que poden afectar a l'esforç dedicat al projecte. La següent taula detalla aquests riscos i n'estableix el nivell estimat.

Codi	Risc	Causa	Conseqüència	Probab.	Impac.	Nivell
R01	No disponibilitat de dades necessàries	Dades necessàries no obertes o no disponibles a temps	No poder utilitzar algunes de les tècniques d'aprenentatge automàtic previstes. Menor qualitat dels resultats obtinguts.	Baixa	Mitjà	Mitjà
R02	Dificultats en el desplegament en l'entorn *big data*	Manca de familiaritat amb l'entorn escollit	Allargament de l'etapa de desplegament en l'entorn *big data*	Mitjana	Mitjà	Mitjà
R03	Dificultats amb el marc Apache Spark	Desconeixement del marc Apache Spark	Endarreriment en l'inici del desplegament en l'entorn *big data*	Baixa	Mitjà	Mitjà

R04	Gran nombre de mètodes d'aprenentatge utilitzats	Projecte no acotat a un sol algorisme	Endarreriment en la fase d'implementació dels algorismes i en la validació de resultats	Mitjana	Alt	Alt
R05	Situacions laborals i familiar	Canvis en l'àmbit laboral o familiar	Reducció de l'esforç dedicat al projecte	Mitjana	Mitjà	Mitjà

Taula 6. Avaluació de riscos.

Un cop identificats els riscos del projecte, s'ha desenvolupat el pla de contingència per mitigar-los o corregir-los i establir les accions que caldrà dur a terme si es materialitzen.

Codi	Risc associat	Acció	Tipus	Data límit	Risc residual
A1R01	R01	Avaluar les dades disponibles en una etapa inicial del projecte	Mitigador	20-10-2019	Baix
A2R01		Cerca de diverses fonts alternatives per a un determinat tipus de dades	Corrector	20-10-2019	Molt baix
A1R02	R02	Implementació dels algorismes en un entorn local	Corrector	06-11-2019	Molt baix
A1R03	R03	Utilització d'un entorn de desenvolupament amb un temps d'aprenentatge baix	Mitigador	03-11-2019	Mitjà
A1R04	R04	Reduir el nombre de tècniques utilitzades	Mitigador	15-11-2019	Baix
A1R05	R05	Reducció de l'abast del projecte eliminat els objectius E3, E4 (desplegament en entorn *big data*) i E6 (obtenció de models de regressió de la mida dels incendis)	Mitigador	03-11-2019	Baix

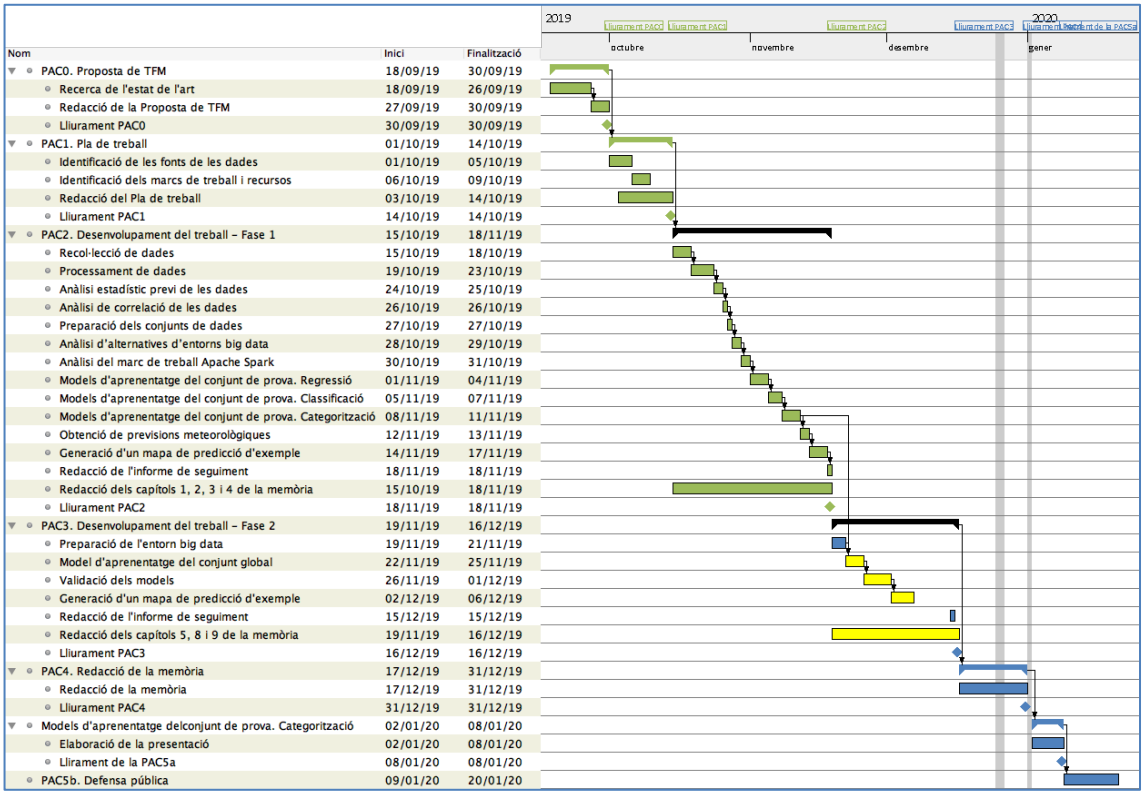
Taula 7. Pla de contingència.

1.4.5 Seguiment de la planificació.

Durant la realització de la fase 1 de desenvolupament del treball, s'han produït canvis a la planificació inicial ja que s'han avançat nombroses tasques. El motiu d'aquest avanç és que s'ha dedicat més temps del previst a la fase 1 amb la finalitat de mitigar el riscs R05 identificat en el pla de projecte i evitar així haver d'aplicar l'acció de contingència A1R05 consistent en la reducció de l'abast del projecte eliminant diversos objectius.

Per altra banda, s'ha materialitzat el risc R01 ja que no s'ha disposat de les dades meteorològiques previstes a l'inici del projecte i s'han utilitzat les accions de contingència A1R01 i A2R01 per a corregir-lo.

El següent diagrama de Gantt mostra la planificació un cop realitzats els canvis anteriors així com l'estat de la consecució dels treballs en la data de finalització de les fase 1. D'altra banda, en la fase 2 s'ha complert la replanificació sense cap desviament.



Nom	Inici	Finalització
PAC0. Proposta de TFM	18/09/19	30/09/19
Recerca de l'estat de l'art	18/09/19	26/09/19
Redacció de la Proposta de TFM	27/09/19	30/09/19
Lliurament PAC0	30/09/19	30/09/19
PAC1. Pla de treball	01/10/19	14/10/19
Identificació de les fonts de les dades	01/10/19	05/10/19
Identificació dels marcs de treball i recursos	06/10/19	09/10/19
Redacció del Pla de treball	03/10/19	14/10/19
Lliurament PAC1	14/10/19	14/10/19
PAC2. Desenvolupament del treball – Fase 1	15/10/19	18/11/19
Recol·lecció de dades	15/10/19	18/10/19
Processament de dades	19/10/19	23/10/19
Anàlisi estadístic previ de les dades	24/10/19	25/10/19
Anàlisi de correlació de les dades	26/10/19	26/10/19
Preparació dels conjunts de dades	27/10/19	27/10/19
Anàlisi d'alternatives d'entorns big data	28/10/19	29/10/19
Anàlisi del marc de treball Apache Spark	30/10/19	31/10/19
Models d'aprenentatge del conjunt de prova. Regressió	01/11/19	04/11/19
Models d'aprenentatge del conjunt de prova. Classificació	05/11/19	07/11/19
Models d'aprenentatge del conjunt de prova. Categorització	08/11/19	11/11/19
Obtenció de previsions meteorològiques	12/11/19	13/11/19
Generació d'un mapa de predicció d'exemple	14/11/19	17/11/19
Redacció de l'informe de seguiment	18/11/19	18/11/19
Redacció dels capítols 1, 2, 3 i 4 de la memòria	15/10/19	18/11/19
Lliurament PAC2	18/11/19	18/11/19
PAC3. Desenvolupament del treball – Fase 2	19/11/19	16/12/19
Preparació de l'entorn big data	19/11/19	21/11/19
Model d'aprenentatge del conjunt global	22/11/19	25/11/19
Validació dels models	26/11/19	01/12/19
Generació d'un mapa de predicció d'exemple	02/12/19	06/12/19
Redacció de l'informe de seguiment	15/12/19	15/12/19
Redacció dels capítols 5, 8 i 9 de la memòria	19/11/19	16/12/19
Lliurament PAC3	16/12/19	16/12/19
PAC4. Redacció de la memòria	17/12/19	31/12/19
Redacció de la memòria	17/12/19	31/12/19
Lliurament PAC4	31/12/19	31/12/19
Models d'aprenentatge delconjunt de prova. Categorització	02/01/20	08/01/20
Elaboració de la presentació	02/01/20	08/01/20
Llirament de la PAC5a	08/01/20	08/01/20
PAC5b. Defensa pública	09/01/20	20/01/20

Imatge 2. Replanificació i situació de l'avanç del projecte en el lliurament de la fase 1.

1.5 Breu sumari de productes obtinguts

Un cop finalitzat el projecte s'han obtingut diversos mapes a partir dels models d'aprenentatge implementats:

- Similitud de les condicions meteorològiques del dia 29-10-2019 i els sis dies posteriors.
- Similitud de les condicions meteorològiques del dia 27-07-2017 i els sis dies posteriors.
- Similitud de les condicions meteorològiques del dia 01-07-1994 i els sis dies posteriors.
- Estimació del risc d'incendi forestal del dia 28-07-2017 obtinguts amb els models: kNN, SVM, arbres de decisió, *Random forest*, AdaBoost, bayesià ingenu gaussià, perceptró multicapa i combinació de mètodes.
- Estimació del risc d'incendi forestal d'àmbit municipal, comarcal i dels principals espais naturals del dia 28-07-2017 obtingut amb el model perceptró multicapa.
- Estimació del risc d'incendi forestal del dia 30-10-2019 obtingut amb el model perceptró multicapa.
- Estimació del risc d'incendi forestal del dia 02-07-1994 obtinguts amb els models: perceptró multicapa i SVM.
- Estimació del risc d'incendi forestal del Parc Natural de Sant Llorenç del Munt i l'Obac del dia 28-07-2017 obtingut amb el model perceptró multicapa.

- Estimació del risc d'incendi forestal en els principals senders i zones d'estacionament de vehicles del Parc Natural de Sant Llorenç del Munt i l'Obac del dia 28-07-2017 obtingut amb el model perceptró multicapa.

1.6 Breu descripció dels altres capítols de la memòria

Els capítols 2, 3 i 4 analitzen diversos aspectes d'interès per al projecte des d'una vessant teòrica. En primer lloc, en el capítol 2 s'introdueix el concepte d'aprenentatge automàtic, el camp de la intel·ligència artificial utilitzat en el projecte per a, tot seguit, analitzar diversos algorismes d'agrupament, classificació i optimització útils per als objectius plantejats. En segon lloc, en el capítol 3 s'analitza el paradigma *big data* definint-lo, avaluant tant els seus principals entorns de treball com algunes de les principals eines d'aprenentatge automàtic disponibles i comparant entorns de treball i biblioteques per a escollir-ne els més adients per a projecte. En darrer lloc, el capítol 4 analitza l'estat de l'art en l'ús d'aprenentatge automàtic per a la predicció d'incendis forestals.

Un cop analitzades les bases teòriques del treball els següents capítols descriuen el procés dut a terme en el projecte. D'entrada, en el capítol 5 es defineix el problema i la zona d'estudi. Tot seguit, en el capítol 6 es detallen les fonts de dades, l'anàlisi previ dut a terme i la seva preparació per als algorismes d'aprenentatge. A continuació, en el capítol 7 es descriuen els diferents models d'aprenentatge i en el capítol 8 se n'avalua el rendiment. En últim terme, en el capítol 9 es descriu la implementació dels models obtinguts, és a dir, la generació dels diversos mapes de risc d'incendi forestal.

Per altra banda, en el capítol 10 es detallen les conclusions obtingudes en el projecte, en el capítol 11 s'enumeren els termes i acrònims més important utilitzats en aquesta memòria i, finalment, en el capítol 12 es mostra la bibliografia emprada durant la realització del treball.

2 Aprenentatge automàtic

L'aprenentatge automàtic (en anglès, *machine learning*) és un camp de la intel·ligència artificial que estudia tècniques per a proveir a les màquines de la capacitat d'aprendre i millorar a partir de l'experiència.

Una possible classificació dels mètodes d'aprenentatge automàtic es basa en el tipus d'informació disponible del resultat que ha d'aportar el sistema. Si es coneix la resposta del sistema es parla d'aprenentatge supervisat. En aquest cas es pot disposar d'un atribut solució de tipus numèric que caldrà reproduir en els problemes de regressió, un atribut binari o categòric en els problemes de classificació o bé, si s'empra algorismes de cerca, és possible resoldre problemes. Per contra, si no es coneix la sortida del sistema sinó només una penalització o gratificació segons la resposta d'aquests es parla d'aprenentatge per reforç. Finalment, en l'aprenentatge no supervisat no es disposa de cap mena d'informació sobre les sortides; en aquest cas, els sistemes extreuen el coneixement de la informació d'entrada disponible.

En aquest projecte es pretén emprar algorismes d'aprenentatge automàtic per a tres objectius principals: agrupar els dies segons les seves característiques meteorològiques i relacionar aquests grups amb els incendis forestals, estimar la mida dels incendis en el moment de la ignició i, per acabar, realitzar una predicció espai-temporal del risc d'incendi forestal. El primer d'aquests objectius es pot tractar mitjançant l'ús de l'aprenentatge no supervisat, concretament de categorització o agrupament. En canvi, el segon i tercer objectius son problemes d'aprenentatge supervisat ja que es disposa d'informació dels incendis i, per tant, es tractarà com un problema de regressió en el cas d'estimar la mida dels incendis i de classificació en la predicació del risc.

Tot seguit es defineixen breument els principals algorismes tant supervisats com no supervisats i s'analitza l'adequació d'aquests al problema platejat. En darrer lloc, també es descriu un dels principals algorismes d'optimització: els algorismes genètic.

2.1 Algorismes no supervisats d'agrupament

Els algorismes d'agrupament o categorització (en anglès, *clustering*) son mètodes d'aprenentatge no supervisat ja que no es disposa prèviament d'informació de les classes o grups sinó que el propi mètode és qui ha de descobrir aquests grups a partir de les dades disponibles dividint el conjunt inicial en subconjunts d'objectes amb característiques semblants entre si i diferents de la resta de categories. Un cop obtinguts els grups amb característiques similar és possible utilitzar-los per a classificar noves dades en algun d'ells mitjançant mètodes supervisats de classificació.

Tot seguit es descriuen breument els algorismes d'agrupament utilitzat en aquest projecte: k-mitjanes, agrupament jeràrquic, propagació de l'afinitat, DBSCAN i agrupament espectral.

2.1.1 K-mitjanes

L'algorisme k-mitjanes (en anglès, *k-means*) obté per a cadascuna de les categories un centroide que la descriu i, tot seguit, assigna cada exemple al grup amb centre més proper. Alguns dels principals avantatges d'aquest mètode son la seva senzillesa i rapidesa d'execució. Per contra,

pot fallar en alguns casos, el resultat pot variar en cada execució ja que depèn dels centroides inicials generats de forma aleatòria i no és capaç de determinar el nombre de grups o clústers i, per tant, caldrà conèixer-lo prèviament.

Existeixen diversos mètodes per a escollir els centroides inicials que, combinats amb k-mitjanes, permeten obtenir millors resultats ja que minimitzen la possibilitat que el mètode convergeixi vers un mínim local com, per exemple, k-mitjanes++.

Finalment, cal tenir en compte que existeixen diverses versions de l'algorisme que cerquen reduir el cost computacional: l'algorisme Elkan, que permet reduir-lo mitjançant l'ús de la inequalitat triangular, i el mètode k-mitjanes mini lots (en anglès, *mini-batch*) que utilitza tan sols un subconjunt d'exemples en cada iteració.

2.1.2 Agrupament jeràrquic

Els mètodes d'agrupament jeràrquic poden ser aglomeratius quan parteixen d'una fragmentació completa de les dades i fusionen grups progressivament o divisius en cas contrari.

Per a decidir el grups en què cal dividir o unir es té en compte la cohesió d'aquests. En el cas dels algorismes aglomeratius, que seran els utilitzats en aquest projecte, es poden emprar els següent criteris d'enllaç de grups: complet quan s'utilitza la distancia màxim entre els elements del grup, simple quan s'utilitza la distància mínima, mitjà quan s'utilitza la mitjana entre elements i de guarda (en anglès, *ward*) quan es minimitza la variància dins dels grups.

Aquest mètode acostuma a donar pitjors resultats que altres algorismes d'agrupament ja que és un algorisme voraç i, per tant, no té en compte la conveniència futura de cada decisió.

Un mètode jeràrquic utilitzat en aquest projecte tant per a l'agrupament com per a l'obtenció del nombre de grups és el *Balanced iterative reducing and clustering using herarchies* (BIRCH). Aquest algorisme, especialment indicat per a conjunts de dades grans, permet l'agrupament incremental i dinàmic.

2.1.3 Propagació de l'afinitat

Aquest mètode es basa en l'enviament de missatges entre parelles d'exemples fins aconseguir la convergència cercant els exemples representatius de cadascun dels grups. Degut al fet que no necessita conèixer a priori el nombre de grups s'ha utilitzat per estimar el nombre idoni de clústers en el conjunt de dades.

2.1.4 Density-based spatial clustering of applications with noise (DBSCAN)

Aquest mètode cerca clústers representats per àrees d'alta densitat d'exemples separats per àrees de baixa densitat, deixant sense classificar aquells exemples en àrees poc denses als quals considera dades atípiques (en anglès, *outliers*). En conseqüència, aquest enfocament permet obtenir clústers de qualsevol forma en contraposició a altres mètodes més rígids. N'existeix una generalització per a múltiples rangs anomenada OPTICS.

Com en el cas de l'algorisme de propagació de l'afinitat no és necessari conèixer a l'avança el nombre de grups i, per tant, pot ser utilitzar per a estimar-ne la quantitat.

2.1.5 Agrupament espectral

Aquesta darrera tècnica d'agrupament cerca els valors propis més petits ja que indiquen on les dades presenten escassa variabilitat i, per tant, l'existència d'un grup. Es tracta d'un mètode molt potent ja que és capaç d'agrupar tenint en compte les diferències de densitat en les dades.

2.2 Algorismes supervisats de classificació i regressió

Els algorismes de classificació i regressió son mètodes d'aprenentatge supervisat on es disposa d'un conjunt d'objectes dels quals es coneix el valor de sortida. En el cas de la classificació, es coneix la classe a la que pertany cada objecte.

D'aquesta forma, i centrant-nos en els algorismes de classificació, a partir d'un conjunt d'entrenament d'on es coneix la classe a la que pertany cada objecte, l'algorisme permet classificar nous objectes dels que no se'n coneix la classe.

Existeixen quatre tipus d'algorismes de classificació segon el principi d'inducció utilitzat per a generar els models, basats en: probabilitats, distàncies, regles o *kernels*. En aquest treball s'han utilitzat algorismes de les quatre tipologies per avaluar-ne la seva idoneïtat per a l'estimació de la probabilitat de ris d'incendis forestals i la predicció de la mida final dels incendis.

2.2.1 Naïve Bayes

Naïve Bayes és un dels principals mètodes de classificació fonamentat en models probabilístics. Basat en el teorema de Bayes, assumeix que la presència o absència d'una determinada característica no està relacionada amb la presència o absència d'una altra característica.

Aquest mètode s'utilitza sobretot per a la categorització de textos. Per tant, per a poder-lo utilitzar en aquest projecte on els atributs son tots numèrics s'ha emprat la versió gaussiana de l'algorisme. Per altra banda, és important utilitzar conjunts d'entrenament balancejats, ja que en cas contrari pot obtenir resultats pobres atès que tendirà a classificar en la classe amb més exemples. Per contra, presenta alguns avantatges interessants en el cas d'utilitzar conjunts de dades grans com l'eficiència computacional.

2.2.2 K veïns més propers

El mètode k veïns més propers (en anglès, *k-nearest neighbour*, *kNN*), és un mètode basat en els algorismes de categorització que classifica un objecte a partir de l'objecte (o objectes) més proper. Es tracta doncs, d'un mètode basat en distàncies on cal definir la semblança entre els objectes del domini. Considerant que els atributs de què es disposa son tots ells numèrics s'utilitzarà la distància l'euclidiana.

2.2.3 Arbres de decisió

Els arbres de decisió és un dels principals mètodes basats en regles. Amb aquest algorisme es construeix un arbre on en cada node s'avalua un concepte i en cada fulla s'hi associa una classe. D'aquesta forma, cada nou objecte s'avaluarà seguint els diversos nodes fins a una fulla que el classificarà.

Un dels principals avantatges d'aquest mètode és la facilitat d'interpretació del model obtingut, permetent analitzar la importància de cada atribut en aquest. Per contra, no és apte per conjunts de dades amb molts atribut ja n'augmenta el cost computacional.

2.2.4 *Random forest*

Es tracta d'una variant dels arbres de decisió que utilitza l'agregació de bootstrap[1] per a millorar l'estabilitat i la precisió de la classificació així com reduir la variància i evitar el sobreentrenament. Per altra banda, s'empra la selecció aleatòria d'atributs per a obtenir un conjunt d'arbres de decisió.

Tenint en compte que els arbres de decisió acostumen a estar influïts per soroll, el que pretén el mètode *Random forest* és obtenir la mitjana d'aquests. Per aconseguir-ho, s'utilitza un subconjunt dels casos com a conjunt d'entrenament i la resta de casos com a conjunt de prova per a estimar l'error de l'arbre. Per a cada node, s'empra un subconjunt de les variables d'entrada escollides aleatòriament per determinar la decisió d'aquest. Un cop obtinguts els diversos arbres utilitzant aquesta metodologia es classifica cada cas nou en tots els arbres i s'assigna com a predicció la classe que té una quantitat major d'incidències.

2.2.5 AdaBoost

Mètode basat en un conjunt elevat de regles senzilles per crear classificadors robustos. Els seus principals avantatges son que treballa eficientment en espais grans d'atributs i que permet interpretar el coneixement generat per les regles.

Com a principal inconvenient es pot destacar el seu cost computacional. Per altra banda, la seva versió bàsica només permet classificar entre dues classes. És per aquest motiu que no s'utilitzarà per a classificar els dies en els clústers obtinguts mitjançant agrupament. Per contra, sí que es podrà utilitzar per classificar entre les classes incendi i no incendi.

2.2.6 Màquines de vectors de suport

Les màquines de vectors de suport (en anglès, *suport vector machines*, *SVM*) és un mètode basat en *kernels* on se cerca un hiperplà millor que el dels classificadors lineals. És possible la classificació de múltiples classes mitjançant la tècnica *un contra un*.

Aquest mètode permet obtenir bons resultats en molts tipus de problemes diferents, com el reconeixement de patrons i de sèries temporals, entre d'altres.

[1] Tècnica emprada per a determinar la variabilitat i el biaix d'un estimador mitjançant l'extracció repetida de mostres aleatòries amb reemplaçament a partir de les observacions de què es disposa (TERMCAT).

2.2.7 Xarxes neuronals i aprenentatge profund

Les xarxes neuronals s'inspiren en les neurones dels éssers vius per a simular les propietats dels sistemes neuronals mitjançant models matemàtics. Aquests sistemes tenen la capacitat d'aprendre i son flexibles i tolerants a les fallades.

Les xarxes neuronals senzilles estan formades per tres capes de neurones: una capa d'unitats d'entrada, una d'unitats ocultes i una d'unitats de sortida. Un problema que planteja aquesta estructura és que per a poder aprendre de qualsevol conjunt de dades caldrà tenir prou unitats a la capa oculta i això pot comportar un cost computacional elevat i una baixa capacitat de generalització. A fi de solucionar-ho, en l'aprenentatge profund s'utilitzen diverses capes ocultes.

Els principals tipus de xarxes neuronals profundes son: el perceptró multicapa (en anglès, *multilayer perceptron*, MLP), les xarxes neuronals convolucionals (CNN), les xarxes recurrents, els autocodificadors i les xarxes generatives antagòniques. S'utilitzarà el perceptró multicapa atès que ha estat extensament emprat tant per a tasques de regressió com de classificació. En aquest tipus de xarxes neuronals cada capa aprèn característiques més complexes que parteixen de l'aprenentatge de les capes anteriors.

2.3 Optimització

L'optimització permet cercar els valors que satisfan unes determinades expectatives com ara la minimització d'una funcions com els cost o el temps d'execució o bé la maximització d'altres funcions com l'eficiència, la productivitat, etc. En aquest projecte ens centrarem en l'ús de l'optimització per a minimitzar l'error d'un altre tipus d'algorismes d'aprenentatge automàtic: el perceptró multicapa.

Alguns dels principals algorismes d'optimització son: el descens de gradients, el salt de valls, els algorismes genètics, les colònies de formigues, l'optimització amb eixams de partícules i la cerca tabú. S'ha optat per l'ús d'algorismes genètics ja que aquests algorismes d'intel·ligència evolutiva permeten resoldre un ampli ventall de problemes d'optimització, tant clàssics com combinatoris. Per altra banda, acostumen a oferir solucions bones en un temps acceptable.

2.3.1 Algorismes genètics

Aquests algorismes cerquen la millor solució a un problema simulant la selecció natural sobre un conjunt d'individus. Cada possible solució ve representada pels gens d'un individu que es corresponen a les variables del problema. Tot seguit, aquest conjunt d'individus es fa evolucionar de forma que les millors solucions al problema es reprodueixin amb més probabilitat.

Tot i els avantatges que s'han exposat més a munt per escollir aquest mètode d'optimització cal destacar-ne alguns inconvenients. Per una banda, requereixen una gran quantitat de paràmetres (individus, estratègies de creuament, mutació i selecció, etc.) i operacions (creuament, mutació, selecció, elitisme, funció objectiu) sense unes regles clares per a establir-los. Per altra banda, a causa de la gran quantitat de solucions al problema que proposen els algorismes genètics és necessari calcular la funció d'idoneïtat molt sovint i això, en molts casos, pot significar un alt cost computacional.

3 Big data

En aquest capítol s'analitza el paradigma de dades massives (en anglès, *big data*), es descriuen els principals entorns de treball existents i biblioteques d'aprenentatge automàtic per a dades massives i s'analitza quins d'aquests s'adapten millor a les necessitats del projecte.

3.1 Definició

L'actual generació de grans volums de dades de fonts i formats heterogenis (estructurats, semiestructurats i no estructurats) i a una gran velocitat provocada per l'aparició de noves tendències tecnològiques com la informàtica en núvol (en anglès, *cloud computing*), l'extensió de tot tipus de dispositius intel·ligents i la Internet de les coses (en anglès, *internet of things*) evidencien les limitacions de les tecnologies tradicionals tant pel que fa a la capacitat d'emmagatzematge com a la gestió d'aquestes dades (Oussous, 2017). Per aquest fet, esdevé necessari l'ús de sistemes d'emmagatzematge distribuït així com la computació paral·lela per al seu tractament.

Una de les definicions més esteses del paradigma de dades massives és la proposada per Laney (2001), que el defineix com a "el conjunt de tècniques i tecnologies per al tractament de dades en entorns de gran volum, varietat d'orígens i en els que la velocitat de resposta és crítica". Per tant, es presenten tres dimensions, sovint anomenades les "3 *v*" del *big data*:

- *El* ***volum***. Les dades digitals que es generen creixen de forma exponencial, passant de la generació de gigabytes d'informació diària fa uns anys a terabytes actualment. Això és degut al gran volum de dades que es generen provinents de fonts tan diverses com: Internet, els dispositius mòbils, les imatges satèl·lit, la Internet de les coses, les xarxes socials, els sensors sense fils, les càmeres, la comunicació màquina a màquina (en l'anglès, *machine-to-machine*, M2M), etc.

- *La* ***velocitat***. Les dades son generades a una gran velocitat, sovint en temps real, i han de ser processades ràpidament per a obtenir-ne informació útil.

- *La* ***varietat*** . Les dades provenen de diferents fonts sovint incompatibles: estructurades com bases de dades o fulls de càlcul, semiestructurades com pàgines web o documents XML o bé no estructurades com documents de text, imatges, àudio o vídeo.

Un dels principals reptes d'aquest paradigma és l'anàlisi de dades massives, possible avui en dia gràcies a tècniques com: la mineria de dades, la visualització, l'anàlisi estadístic i l'aprenentatge automàtic (Oussous, 2017).

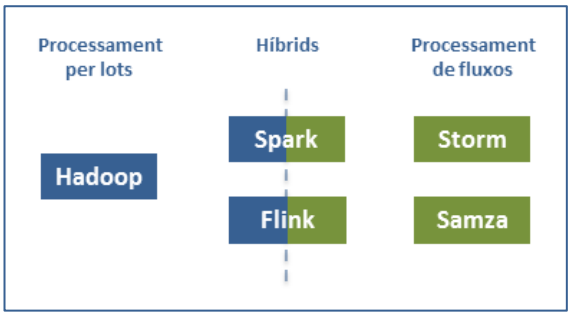
3.2 Principals entorns de treball

Ellingwood (2016) classifica els entorns de treball *big data* en tres grups en funció del mode de processament:

- ***Processament per lots*** (en anglès, *batch processing*). El processament es duu a terme sobre un gran volum de dades que no canvia, sovint dades històriques. Per tant, no és indicat quan el temps és un factor important. Entre aquest tipus destaca Apache Hadoop.

- ***Processament de fluxos de dades*** (en anglès, *stream processing*). Les dades son processades individualment a mesura que entren al sistema. Per tant, és l'indicat quan es requereix processament proper al temps real. S'hi troben els projectes Apache Storm i Apache Samza.

- ***Entorns de treball híbrids***. Suporten els dos tipus de processaments anteriors. Hi destaquen Apache Spark i Apache Flink.



Imatge 3. Classificació dels principals entorns de treball big data.

Tot seguit es descriuen breument els principals entorns de treball *big data* de processament per lots, de fluxos de dades i híbrids.

3.2.1 Apache Hadoop

Apache Hadoop és un entorn de treball (en anglès, *framework*) de codi obert que permet el processament distribuït de grans volums de dades amb un clúster d'ordinadors possibilitant la resolució de problemes *big data* sense la necessitat d'utilitzar maquinari d'altes prestacions.

Les principals característiques d'Apache Hadoop son la seva escalabilitat, l'alta disponibilitat gràcies a la capacitat de detectar i gestiona fallades en el clúster, l'ús del model de programació MapReduce per al processament distribuït de les dades i el sistema d'arxius propi Hadoop Distributed Filesystem (HDFS). Hadoop llegeix les dades del seu propi sistema de fitxers, les divideix en fragments que tot seguit distribueix pels diversos nodes del clúster, processa cada subconjunt de dades en el respectiu node, combina els resultats de cada node i emmagatzema el resultat altre cop al sistema de fitxers.

Així, els tres principals components de Hadoop son:

- ***Hadoop Distributed Filesystem*** (HDFS) és el sistema de fitxers distribuït d'Apache Hadoop i s'integra amb altres sistemes d'emmagatzematge com el sistema de fitxers local o Amazon S3.

- ***YARN,*** sistema de gestió dels recursos del clúster i monitoratge i planificació de tasques. Permet l'execució de tot tipus de programes distribuïts més enllà de MapReduce.

- ***MapReduce*** és un model de programació per al processament de dades que utilitza computació paral·lela. Hadoop distribueix el processament dels programes MapReduce als

diversos nodes d'un clúster que emmagatzemen les dades que han de ser processades amb el sistema de gestió de recursos YARN. D'aquesta forma és possible realitzar processament per lots sobre grans volums de dades obtenint resultats en un temps raonable.

Més enllà de MapReduce, YARN i HDFS, existeix un ampli ecosistema de projectes al voltant de Hadoop, entre els que podem destacar: la base de dades distribuïda orientada a columnes del tipus clau-valor HBase, els sistemes de consultes distribuïdes SQL Impala i Hive i la plataforma d'indexació i cerca dels documents emmagatzemats en HDFS Solr.

Per acabar, el seus principals inconvenients son la seva lentitud en comparació amb altres entorns i la corba d'aprenentatge del model de programació MapReduce.

3.2.2 Apache Storm

Apache Storm és un entorn de treball de codi obert de processament de fluxos de grans volums de dades amb una latència molt baixa i, per tant, indicat per al processament proper al temps real. D'altra banda, el processament de les dades es pot realitzar en qualsevol llenguatge i utilitza Zookeeper per a la coordinació entre els diversos nodes.

Els seus principals avantatges son la baixa latència en el processament de dades i la possibilitat d'integrar-lo amb Hadoop. Per contra, no garanteix que un missatge no pugui ser duplicat.

3.2.3 Apache Samza

Apache Samza és un entorn de treball de processament de fluxos que utilitza el sistema de missatges Apache Kafka per a garantir la tolerància a les fallades, memòria intermèdia i emmagatzematge d'estat. Per altra banda, funciona sobre clústers Hadoop i utilitza el coordinador de clústers YARN per a gestionar-ne els recursos.

Els principals avantatges de Samza son la baixa latència, la possibilitat que diversos subscriptors accedeixin als missatges resultants de cadascun del passos que es duen a terme en el processament de les dades d'entrada, la no pèrdua de dades en el cas que aquestes arribin al sistema a una velocitat més alta de la que son processades, l'emmagatzematge de l'estat i un nivell d'abstracció superior al d'altres entorns de treball de processament de fluxos com Storm. Per contra, com a principal desavantatge cal destacar que no suporta tants llenguatges de programació com Storm.

3.2.4 Apache Spark

Apache Spark[2] és un entorn de treball de codi obert amb capacitats de processament tant per lots com de fluxos de dades. Malgrat que es basa en el funcionament de MapReduce, Spark utilitza computació totalment en memòria i optimitzacions de processament per tal de fer més ràpids els processos per lots. D'aquesta forma, Spark només interacciona amb el sistema de fitxers per a obtenir les dades inicials i emmagatzemar els resultats finals; la resta del procés es realitza en la memòria. Pel que fa al model de processament de fluxos, aquest l'ofereix Spark

[2] http://spark.apache.org/

Streaming gestionant els fluxos de dades com a petits processos per lots que son processats utilitzant el model per lots.

Aquest entorn pot ser utilitzat independentment o conjuntament amb Hadoop per a substituir el motor MapReduce, pot utilitzar diversos magatzems de dades com Cassandra, Hbase, HDFS, Amazon S3 i Apache Hive, consta de biblioteques per al processament de dades com SQL, Spark Streaming, Spark R, MLlib per a l'aprenentatge automàtic i GraphX per al processament gràfic. Finalment, disposa de diverses interfícies de programació d'aplicacions: Java, R, Python i Scala.

En definitiva, el principal avantatge d'Spark sobre MapReduce és la major velocitat. Per altra banda, és molt versàtil ja que pot ser desplegat com un clúster independent o integrat dins d'un clúster Hadoop i també ser emprat en mode per lots o bé de fluxos de dades. Finalment, disposa de biblioteques d'aprenentatge automàtic i consultes interactives i és més fàcil d'implementar que MapReduce.

Per contra, el seu principal desavantatge és l'augment de la latència en sistemes de processament de fluxos quan aquests son molt grans. En conseqüència, és inadequat quan es requereix una latència baixa. Per altra banda, Spark utilitza més recursos que altres entorns i, per tant, no és apropiat quan aquests han de ser compartits amb altres sistemes. En darrer lloc, es requereix més RAM que altres sistemes, per bé que dur a terme els processos en un temps més curt compensa aquest major cost ja que pot disminuir el cost total en hores d'ús del clúster.

3.2.5 Apache Flink

Apache Flink és un entorn de treball de codi obert de processament de fluxos de dades amb la possibilitat de treballar en mode per lots com a un cas especial de processament de fluxos. Per altra banda, té la capacitat de guardar estats per a poder-se recuperar en cas de fallades i garanteix l'ordre dels esdeveniments que entren al sistema.

Els principals avantatges d'aquest entorn son la baixa latència, l'alt rendiment, el fet de no necessitar determinades optimitzacions manuals que sí que requereix Spark, la disponibilitat de diversos mecanismes d'optimització de tasques, un visor web per gestionar-les, consultes similars a SQL, biblioteques d'aprenentatge automàtic, computació en memòria i la possibilitat d'executar processos desenvolupats per a altres sistemes com Hadoop o Storm. Tanmateix, el seu principal desavantatge és que és menys consolidat que altres sistemes.

3.3 Principals eines d'aprenentatge automàtic en l'ecosistema Hadoop

Tot seguit, s'analitzen les principals característiques de les eines d'aprenentatge automàtic més esteses en l'ecosistema Apache Hadoop: Mahout, MLlib, H_2O i SAMOA amb la finalitat d'avaluar quina és la més adient per a la implementació dels models d'aprenentatge previstos.

3.3.1 Apache Mahout

Apache Mahout [3] és una biblioteca de codi obert d'algorismes d'aprenentatge automàtic escalables escrita en Java que pot ser utilitzada quan el volum de dades que cal processar és

[3] http://mahout.apache.org/

molt gran. És una de les biblioteques més completes i conegudes d'aprenentatge. D'altra banda, permet als usuaris desenvolupar els seus propis algorismes distribuïts.

Tot i que originalment les seves principals implementacions utilitzaven Apache Hadoop, actualment està centrat en Spark. Els principals algorismes resolen problemes d'àlgebra lineal distribuïda, preprocessament, regressió, recomanació i categorització.

3.3.2 MLlib

Apache Spark MLlib[4] és una biblioteca per a l'aprenentatge automàtic escalable sobre l'entorn Apache Spark que suporta diversos llenguatges de programació com: Java, Scala, Python i R.

Un dels seus principals avantatges és el rendiment, molt superior al que es pot assolir amb MapReduce. Per altra banda, pot funcionar sobre diversos tipus de clústers com: EC2, Mesos i Hadoop i pot accedir a centenars d'orígens de dades com: HDFS, Cassandra, Hive i HBase. Finalment, disposa d'algorismes de classificació, regressió, categorització, sistemes recomanadors i preprocessament de dades.

3.3.3 H_2O

Tot i que H_2O[5] disposa d'una edició empresarial, també és una eina de codi obert amb la característica de disposar d'una interfície gràfica d'usuari. Malgrat que existeixen altres eines d'aprenentatge amb IGU com Weka, RapidMiner i KNIME, aquestes no estan pensades per entorns *big data*. Per altra banda, H_2O implementa diverses eines per a xarxes neuronals profundes, pot ser programat en Java, Python, R i Scala i disposa del seu propi motor de processament tot i que també es pot integrar amb altres motors com Spark i Storm.

3.3.4 SAMOA

Apache SAMOA[6] és una plataforma dissenyada per a l'aprenentatge sobre dades distribuïdes en temps real. Disposa d'algorismes de classificació, categorització, regressió, mineria de patrons i *boosting*[7], entre d'altres. En darrer lloc, és un entorn pensat per a grans volums de dades que s'actualitzen constantment i on cal obtenir els resultats de l'anàlisi a temps real.

3.4 Comparació d'entorns de treball i biblioteques d'aprenentatge

En la següent aula es resumeixen les principals característiques dels entorns de treball *big data* analitzats en l'apartat anterior. Podem concloure que l'entorn que millor s'adapta a les necessitats del projecte és Spark per la seva versatilitat en poder treballar en mode de processament de lots i de fluxos, la seva rapidesa de processament tant de conjunts de dades grans com petits, la disponibilitat de nombroses biblioteques d'aprenentatge automàtic i la possibilitat d'utilitzar diversos llenguatges.

[4] http://spark.apache.org/mllib/
[5] https://www.h2o.ai/
[6] https://samoa.incubator.apache.org/
[7] https://en.wikipedia.org/wiki/Boosting_(machine_learning)

Característica		**Entorn**				
		Hadoop	*Storm*	*Samza*	*Spark*	*Flink*
Mode de processament		Lots	Fluxos	Fluxos	Lots i fluxos	Lots i fluxos
Escalabilitat		Horitzontal	Horitzontal	Horitzontal	Horitzontal	Horitzontal
Mode de computació		Basat en accessos a disc	En memòria	Basat en accessos a disc i en memòria[8]	En memòria	En memòria
Autoescalat		Sí	No	Sí	Sí	No
Temps de processament	*Conjunts de dades grans*	Menys ràpid	-	-	El més ràpid	Més lent
	Conjunts de dades petits	Més lent	-	-	El més ràpid	Menys ràpid
Biblioteques d'aprenentatge automàtic		Mahout	SAMOA, H_2O	SAMOA	Mahout, MLlib, H_2O	Flink-ML, SAMOA
Llenguatges suportats		Java	Qualsevol	Java	Java, R, Python, Scala	Java, Scala

Taula 8. Comparació dels principals entorns de treball big data (adaptat, en part, d'Alkatheri, 2019).

Per altra banda, totes del biblioteques analitzades implementen tots els algorismes d'aprenentatge automàtic previstos excepte Apache Mahout que no disposa d'algorismes supervisats de classificació[9]. Així mateix, de les diverses biblioteques d'aprenentatge automàtic disponibles en l'entorn Spark, MLlib és una de les més extenses i robustes. En conseqüència és la més indicada per a la implantació dels algorismes previstos.

[8] Accés en memòria per als elements més freqüents.
[9] Els algorismes de classificació implementats en MapReduce ja no son mantinguts.

4 Estat de l'art

En un recent article (Rolnick, 2019) publicat per un grup de prestigiosos experts en aprenentatge automàtic s'analitza com afrontar el canvi climàtic amb l'ús de tècniques d'intel·ligència artificial tant pel que fa a la seva mitigació com a l'adaptació a les seves conseqüències. En particular, un dels principals camps apuntats per l'estudi és la predicció del risc. En el cas del risc d'incendis, l'aprenentatge automàtic avançat pot ajudar a la seva predicció amb eines de monitoratge precises i de baix cost. És per aquest motiu que darrerament s'han realitzar diversos esforços per aprofundir en l'ús de l'aprenentatge automàtic per a la predicció dels incendis forestals, àmbit en què es vol incidir en aquest projecte.

En les darreres dècades s'han utilitzat diversos mètodes per a l'avaluació del risc d'incendi. La literatura destaca tres grans àmbits: en primer lloc els sistemes d'informació geogràfica, en segon lloc la simulació i els models probabilístics i, en tercer lloc, més recentment s'han utilitzat diverses tècniques d'intel·ligència artificial com, per exemple: màquines de vectors de suport, *Random Forest*, xarxes neuronals i agents[10], entre d'altres. En molts casos, les tècniques d'intel·ligència artificial han superar a les anteriors (Jaafari, 2019).

Un dels inconvenients dels models matemàtics i les simulacions (Radke, 2019) son la seva dificultat i cost computacional elevat. Per altra bada, moltes de les tècniques utilitzades no empren dades històriques per ajudar en la predicció. És per aquest motiu que es proposa l'ús d'intel·ligència artificial per a obtenir els models de predicció d'incendis.

La següent taula mostra un recull de projectes en què s'han utilitzat tècniques d'aprenentatge automàtic per al modelatge de diversos aspectes dels incendis forestals: la predicció de la superfície cremada, la predicció de la seva mida final, l'estimació de la susceptibilitat d'incendi, la predicció espacial, la predicció de la seva propagació i l'avaluació dels motius dels incendis, entre d'altres.

Pel que fa als mètodes d'aprenentatge automàtic utilitzats es pot comprovar com en els darrers anys hi ha hagut diversos projectes de recerca encaminats a l'avaluació de diferents mètodes per establir la seva idoneïtat per al modelatge dels factors relacionats amb els incendis forestals. Algunes de les tècniques que més apareixen en la literatura analitzada son les xarxes neuronals profundes (tant les convolucionals com el perceptró multicapa) i el mètode *Random forest*.

[10] Sistemes computacionals autònoms, capaços de raonar i aprendre i també de comunicar-se amb altres agents. Es tracta d'un dels principals camps de la intel·ligència artificial.

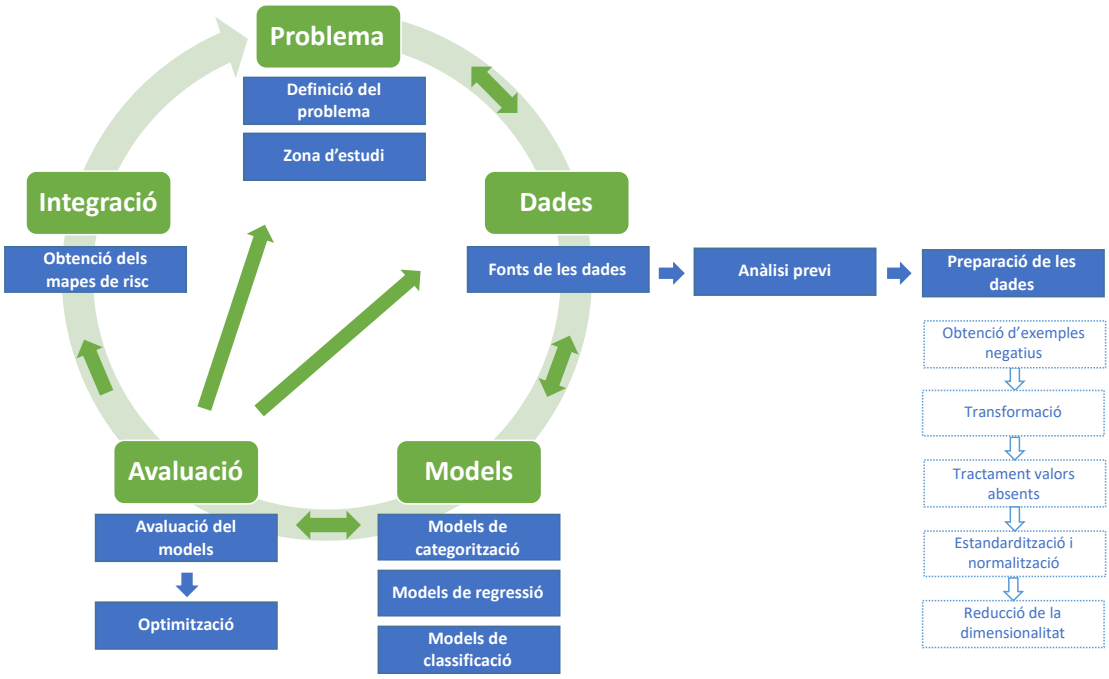
Referència	Finalitat	Mètodes utilitzats
Castelli, 2019	Predicció de la superfície cremada	Programació genètica
Coffield, 2019	Predicció de la mida final de l'incendi en el moment de la ignició	Arbres de decisió, *Random forst*, kNN, *boosting* i perceptró multicapa
Ghorbanzedeh, 2019	Predicció espacial de la susceptibilitat d'incendi forestal	Xarxes neuronals, màquines de vectors de suport, i *Random forest*
Jaafari, 2019	Predicció espacial de la probabilitat d'incendi forestal	ANFIS[11] amb tres tipus d'optimització: algorismes genètiques, eixam de partícules i l'algorismes *Shuffled frog leaping*
Kin, 2019	Probabilitat d'incendi forestal	Màxima entropia i *Random forest*
Pourghasemi, 2014	Obtenció de mapes de susceptibilitat d'incendi forestal	Procés de jerarquia analítica modificat[12] i Lògica difusa de Mamdani[13]
Radke, 2019	Predicció de la propagació dels incendis forestals	Xarxes neuronals convolucionals amb optimització RMSProp i algorisme de gradient estocàstic
Safi, 2013	Predicció d'incendis forestals	Xarxes neuronals
Sayad, 2019	Monitoratge, predicció i prevenció d'incendis forestals	Xarxes neuronals, màquines de vectors de suport, teledetecció
Su, 2018	Identificació dels motius d'incendis forestals	*Random forest*
Subramanian, 2018	Modelatge de les dinàmiques dels incendis forestals	Aprenentatge per reforç
Zhang, 2019	Modelatge de la susceptibilitat d'incendi forestal	Xarxes neuronals convolucionals

Taula 9. Resum de projectes on s'ha utilitzat aprenentatge automàtic per al modelatge d'incendis forestals.

[11] Adaptive neuro-fuzy inference system.
[12] En anglès, *modified analytical herarchy process*.
[13] En anglès, *Mamdani fuzzy logic.*

5 Presentació del problema

El següent diagrama mostra el procés seguit en aquest projecte. D'entrada, s'ha definit el problema i la zona d'estudi. Tot seguit, s'han obtingut i preparat les dades, identificant-ne les fonts, analitzant-les i preparant-les per al seu ús en els tres tipus de models d'aprenentatge previstos. Un cop obtinguts els models, aquests han estat avaluats i optimitzats. Per acabar, s'han integrat els resultat obtinguts generant mapes de risc d'incendi forestal.



Imatge 4. Fases del procés.

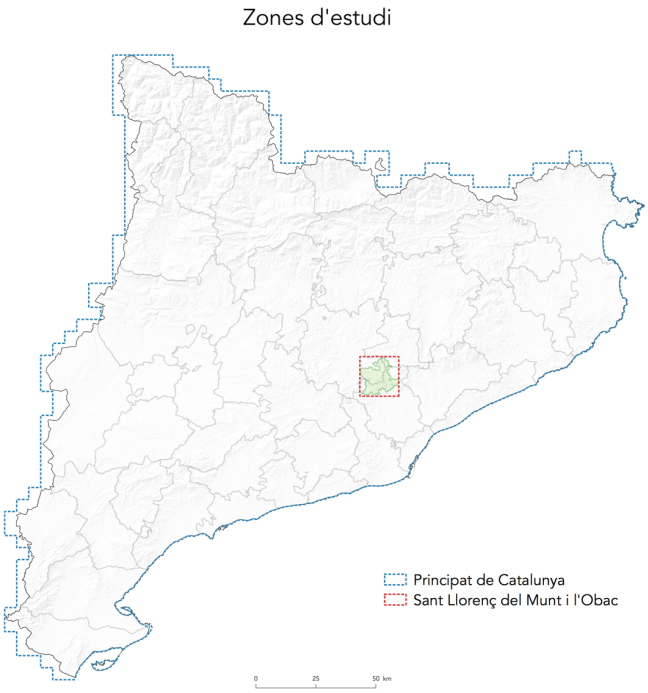
5.1 Definició del problema

Un cop analitzats els principals algorismes d'aprenentatge automàtic útils per al propòsit del projecte, el paradigma del *big data* i l'estat de l'art s'ha iniciat el procés d'obtenció dels diversos models d'aprenentatge automàtic.

La finalitat d'aquest projecte és l'ús d'algorismes d'aprenentatge automàtic per a donar resposta a tres problemàtiques relacionades amb les incendis forestals: la predicció del risc d'incendi forestal amb algorismes supervisats de classificació, l'estimació de la mida de l'incendi en el moment d'ignició d'aquest amb algorismes de regressió i la predicció de les zones de risc d'incendi segons les seves característiques meteorològiques amb l'ús d'algorismes no supervisats de categorització.

5.2 Zona d'estudi

S'han establert dues zones d'estudi de mides diferent: el conjunt del Principat de Catalunya i la zona del Parc Natural de Sant Llorenç del Munt i l'Obac. La primera zona[14] té una superfície de 34.104,8 km^2 i la segona de 267,5 km^2.



Imatge 5. Zones d'estudi.

La finalitat de treballar amb dues àrees d'estudi és, per una bada, poder entrenar els algorismes amb conjunts de dades de diverses mides per a la realització de les proves i, per l'altra, avaluar la capacitat dels algorismes de realitzar prediccions amb diverses resolucions espacial.

[14] La zona d'estudi del Principat de Catalunya compren aquest territori i zones adjacents com es mostra en el mapa.

6 Dades

Un cop definit els problema que es vol resoldre així com la zona d'estudi, s'ha iniciat el procés d'identificació de les fonts de dades, l'obtenció d'aquestes, l'anàlisi previ i la seva preparació per a l'entrenament dels models d'aprenentatge.

6.1 Fonts de les dades

La següent taula mostra un resum de les dades utilitzades en el projecte i les fonts d'aquestes. En el cas de les dades provinents del Climate Data Store, l'ECMWF i l'EarthData l'accés s'ha realitzat amb les diverses API (interfícies de programació d'aplicacions) que ofereixen aquest serveis.

Dades		Fonts
Incendis forestals	Coordenades	Servei de Prevenció d'Incendis Forestals del la Generalitat de Catalunya[15]
	Data inici	
	Hora inici	
	Superfície cremada	
Orografia	Altitud	ICGC[16] i generació pròpia
	Orientació	Generació pròpia
	Pendent	Generació pròpia
Meteorologia	Humitat relativa	Climate Data Store[17] i ECMWF[18]
	Humitat específica	
	Temperatura	
	Component est *u* del vent	
	Component nord *v* del vent	
	Anàlegs	Servei Meteorològic de Catalunya[19]
Vegetació i usos i cobertes del sòl	NDVI	EarthData[20] (NASA)
	Usos i cobertes del sòl	Departament de Territori i Sostenibilitat de la Generalitat de Catalunya[21]

Taula 10. Fonts de les dades utilitzades.

Per a l'elecció de l'anterior conjunt de dades i de les seves característiques s'han analitzat els projectes de la taula 9 tot avaluant les dades utilitzades en aquests, la resolució temporal i espacial així com la qualitat dels seus resultats. S'ha optat per a utilitzat un conjunt extens de dades de diversos àmbits i amb la millor resolució espacial i temporal possible a fi d'obtenir models predictius robustos.

Per altra banda, com es detallarà en l'apartat 7.2, s'han obtingut noves dades orogràfiques derivades amb la finalitat de millorar els resultats en l'estimació de la mida dels incendis forestals en el moment de la ignició.

[15] http://agricultura.gencat.cat/
[16] Institut Cartogràfic i Geològic de Catalunya. http://www.icgc.cat
[17] https://cds.climate.copernicus.eu
[18] European Centre for Medium-Range Weather Forecasts. https://www.ecmwf.int/
[19] https://www.meteo.cat/
[20] https://earthdata.nasa.gov/
[21] http://territori.gencat.cat

6.2 Anàlisi previ de les dades

Per tal de comprendre millor les característiques de les dades abans de la seva utilització en els algorismes d'aprenentatge, s'han analitzat els diversos conjunts disponibles: dades d'incendis forestals, orogràfiques, meteorològiques, vegetació i cobertes i usos del sòl.

La següent taula en mostra un resum on s'indica els atributs recollits, el sensor en el cas de les dades provinents de satèl·lits, la seva resolució espacial i temporal, el format original, el domini i les unitats. Com es pot observar, s'han utilitzat dades per a l'índex NDVI provinents de diverses fonts: del sensor MODIS[22] (*Moderate-Resolution Imaging Spectroradiometer*) dels satèl·lit Terra per al període comprès entre els anys 2000 i 2018 i del sensor AVHRR[23] (*Advanced Very High Resolution Radiometer*), predecessor de MODIS, per al període comprès entre els anys 1987 i 1999.

Tipus	Atribut	Sensor	Resolució espacial	Resolució temporal	Format	Domini	Unitat
Incendis forestals	Coordenada X	-	1:5.000	-	Taula	265905 a 523806	m
	Coordenada Y	-	1:5.000	-	Taula	4491796 a 4747696	m
	Data inici	-	-	-	Taula	1987/1/2 a 2018/10/1	Dia
	Hora inici	-	-	-	Taula	00:00 a 23:59	Hora
	Superfície cremada	-	0,01 Ha	-	Taula	1 a 25776	Ha
Orografia	Altitud	-	15x15m	-	Ràster	0 a 3119	m
	Orientació	-	15x15m	-	Ràster	0 a 360	Graus
	Pendent	-	15x15m	-	Ràster	0 a 89,83	Graus
Meteorologia	Humitat relativa	MODIS	0,25x0,25°	1 hora	Ràster	0 a 100	%
	Humitat específica	MODIS	0,25x0,25°	1 hora	Ràster	0 a 0,0175	Kg
	Temperatura	MODIS	0,25x0,25°	1 hora	Ràster	271 a 314	Kelvin
	Component est (U) del vent	MODIS	0,25x0,25°	1 hora	Ràster	-10,28 a 11,58	$m\ s^{-1}$
	Component nord (V) del vent	MODIS	0,25x0,25°	1 hora	Ràster	-19,57 a 14,55	$m\ s^{-1}$
	Anàlegs	-	Tot Catalunya	1 dia	Taula	1948/5/2 a 2009/10/26	Dia
Vegetació i usos i cobertes del sòl	NDVI	MODIS	250x250m	16 dies	Ràster	-1 a 1	NDVI
		AVHRR	0,05x0,05°	15 dies	Ràster	-1 a 1	NDVI
	Usos i cobertes del sòl	-	10x10m	-	Ràster	0 a 40	Classe

Taula 11. Dades utilitzades en el projecte.

6.2.1 Incendis

Les dades d'incendis utilitzades s'han centrat en els de tipus forestal ja que la finalitat és estimar el risc d'aquesta classe d'incendis. El conjunt està formada per un total de 4.249 casos

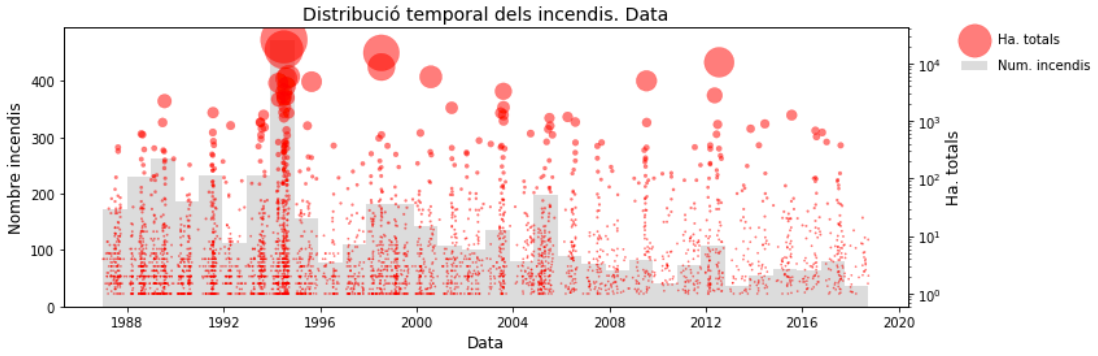
[22] https://modis.gsfc.nasa.gov/
[23] https://earth.esa.int/web/guest/missions/3rd-party-missions/current-missions/noaa-avhrr

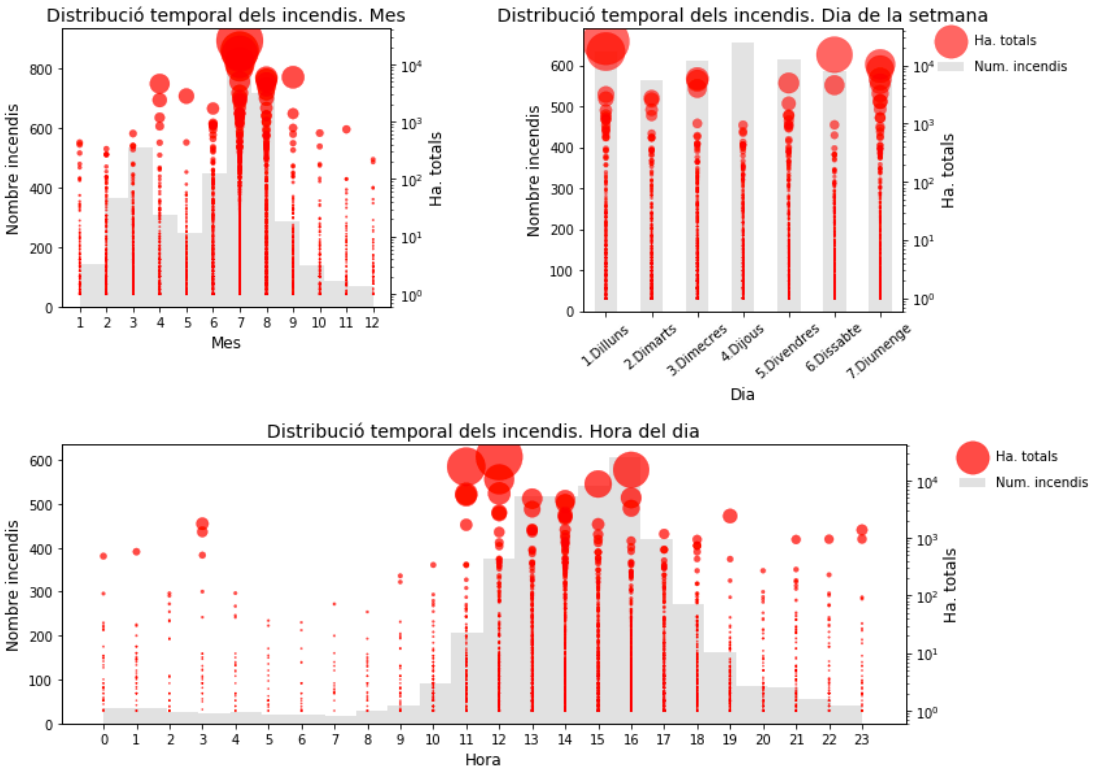
corresponents a la totalitat dels incendis forestals d'una o més hectàrees cremades registrats en un període de 32 anys entre el 02-01-1987 i el 01-10-2018. En la següent taula es resumeixen els principals valors estadístics per a cadascun dels atributs del conjunt d'incendis.

Estadístic	**Data d'inici**	**Hora d'inici**	**X** *(m)*	**Y** *(m)*	**Superfície cremada** *(ha.)*		
					Forestals	*No forestals*	*Totals*
Mínim	02-01-1987	00:00	265905	4491796	1	0	1
1r quantil					1,5	0	1,5
Mediana					2,5	0	2,9
Mitjana			379994	4614808	44,87	11,09	55,96
3r quantil					6,0	0,1	7,5
Màxim	01-10-2018	23:58	523806	4747696	16.832,8	8.943,2	25.776
Desviació estand.			65333	5266	458,69	191,14	625,89
Valors únics	2567	622					
Valor més freqüent	04-07-1994	16:00					
Freq. valor més freqüent	20	235					
Valors absents	0	1	0	0	0	0	0

Taula 12. Principals valors estadístics dels atributs del conjunt d'incendis.

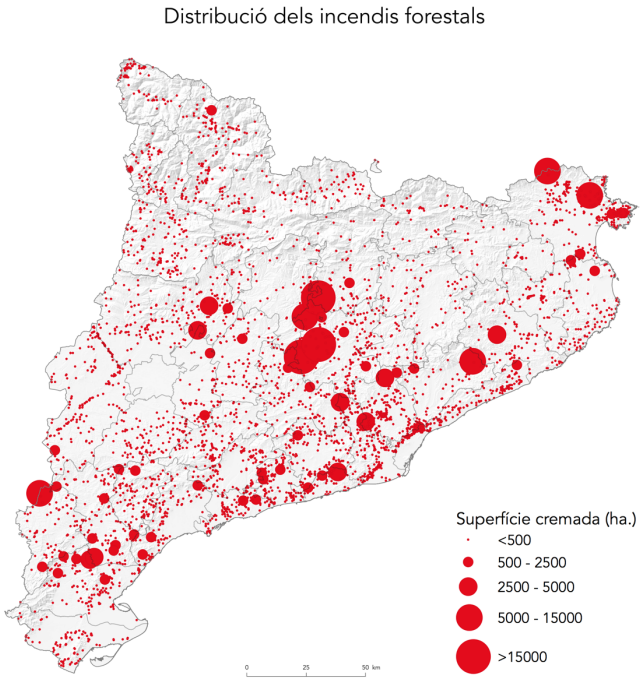
En les següents gràfiques es mostra la distribució temporal dels incendis per data, mes, dia de la setmana i hora. Convé destacar que la majoria d'incendis son de petites dimensions, amb una mediana tan sols de 2,9 i una mitjana de 55,96 hectàrees cremades. Pel que fa a la seva distribució, a la dècada dels 90 del segle passat és quan es varen concentrar la majoria d'incendis grans. Per altra banda, els mesos amb més concentració d'incendis son el juliol seguit de l'agost. En relació als dies de la setmana, tot i que s'observa una distribució força uniforme, els dilluns son els que concentren els incendis més grans. Finalment, la majoria d'incendis es concentren en les hores centrals, aproximadament entre les 11:00 i les 17:00 mentre que els de major magnitud s'han iniciat entre les 11:00 i les 16:00.





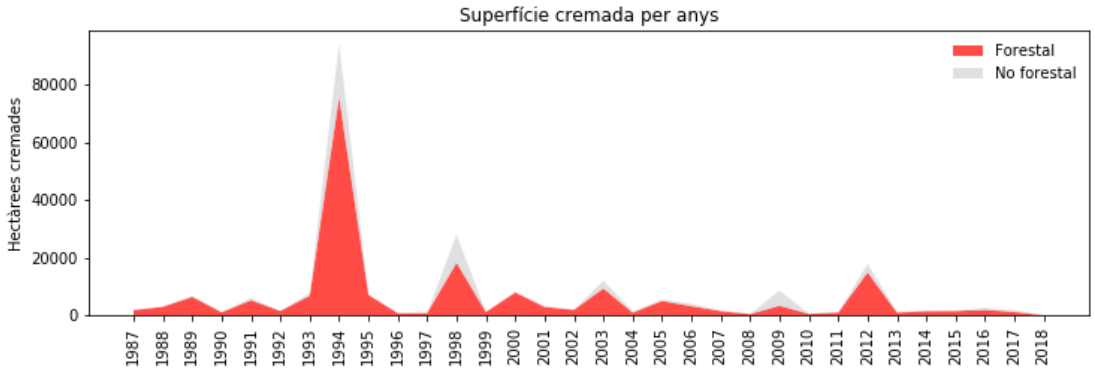
Imatge 6. Histogrames de distribució dels incendis per data, mes, dia de la setmana i hora.

El següent mapa mostra la distribució espacial dels incendis així com la superfície cremada, on s'observa que, per bé que es distribueixen pràcticament per tot el territori, els gran incendis es concentren en determinades zones, com és ara la Catalunya central i els extrems nord-est i sud.



Imatge 7. Distribució espacial dels incendis forestals.

Finalment, s'ha analitzat l'evolució de la superfície cremada, tant forestal com no forestal, al llarg dels anys on s'observa que els anys amb més superfície cremada en el període d'estudi[24] han estat: 1994, 1998 i 2012.



Imatge 8. Superfície cremada per anys.

Sovint, es classifiquen els incendis per mida considerant aquells inferiors a les 500 ha. com a normals i els superiors a aquesta superfície com a grans incendis. La següent taula mostra la distribució dels incendis registrats segons aquesta classificació on es pot comprovar que pràcticament la meitat dels incendis registrats son inferiors a les 3 hectàrees.

Tipus	**Mida** (ha.)	**Núm. d'incendis**
Incendi normal	entre 1 i 3	2137
	entre 3 i 500	2047
Gran incendi	> 500	63

Taula 13. Distribució d'incendis per mida.

6.2.2 Orografia

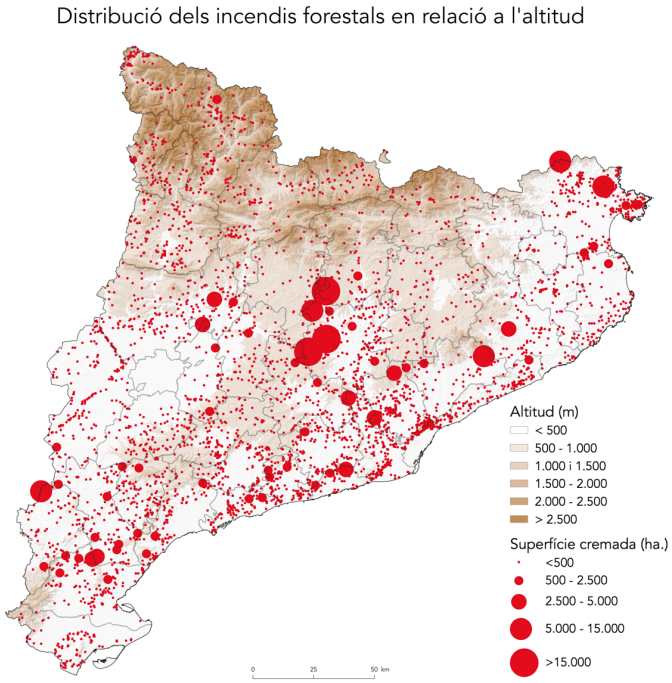
Les dades orogràfiques s'han obtingut a partir del darrer model digital d'elevacions de l'Institut Cartogràfic i Geològic de Catalunya de 15x15m de resolució. A partir d'aquest s'han generat mapes d'orientació i de pendent i se n'ha extret la informació corresponent a cadascun dels punts d'ignició. La següent taula en mostra les principals característiques.

Estadístic	**Altitud** *(m)*	**Orientació** *(graus)*	**Pendent** *(%)*
Mínim	0	0,5	0
1r quantil	146,2	103,7	5,3
Mediana	303,2	173,5	12,4
Mitjana	429,9	176,0	13,9
3r quantil	573,0	247, 6	20,7
Màxim	2540,3	360,0	89,8
Desviació estàndard	400,03	93,3	10,2
Valors absents	1	11	0

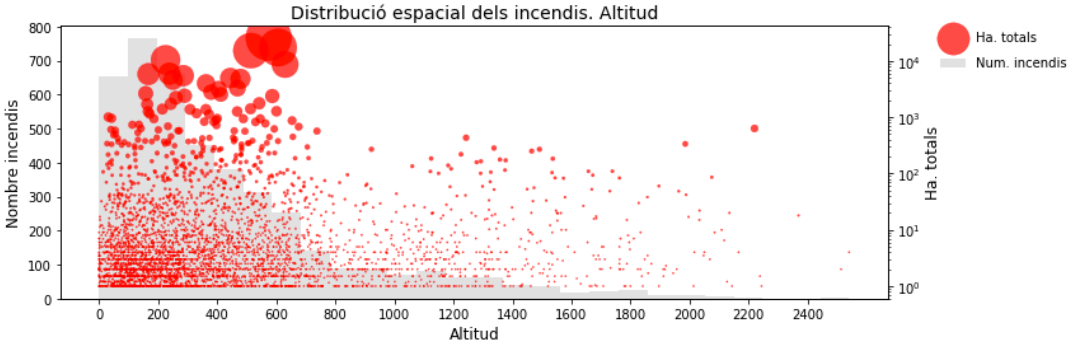
Taula 14. Principals valors estadístics dels atributs de les dades orogràfiques.

[24] Cal tenir en compte que de l'any 2018, darrer any del període d'estudi, nomes es disposa de dades dels nou primers mesos.

El següent mapa mostra la relació espacial entre els incendis i l'altitud on s'observa que la majoria d'incendis, i especialment els grans incendis, es concentres en cotes baixes amb una altitud mitjana de 303m. Aquest fenomen també es pot apreciar en l'histograma que l'acompanya, on els incendis amb més freqüència es donen en les cotes més baixes i tot seguit disminueix la freqüència a mesura que augmenta l'altitud. Per altra banda, la mida dels incendis també mostra relació amb la cota ja que en les altituds més baixes aquestes son majors, concentrant-se els més grans aproximadament entre les cotes 150m i 800m.

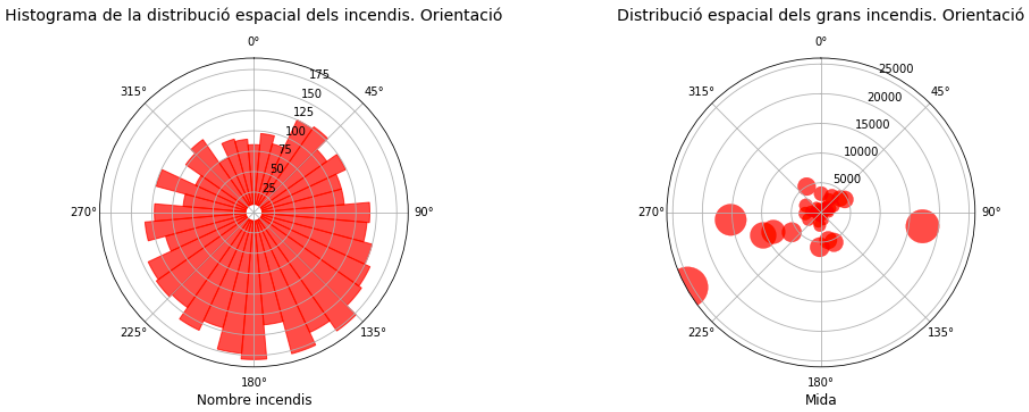


Imatge 9. Distribució dels incendis en relació a l'altitud.



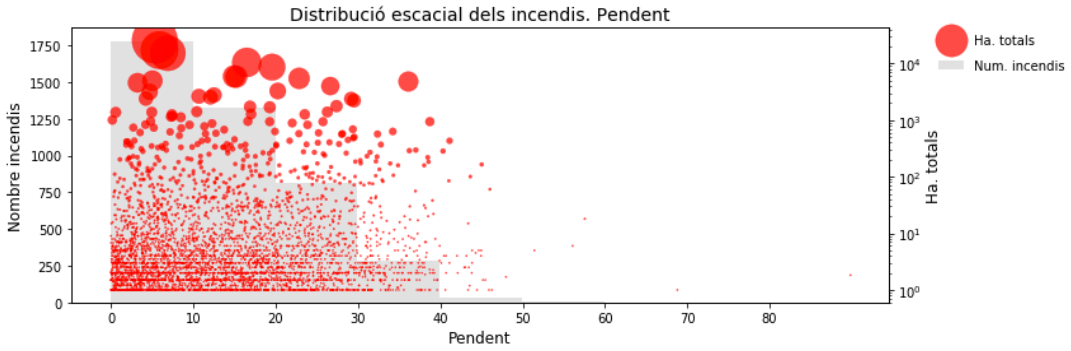
Imatge 10. Histograma de la distribució espacial dels incendis per altitud.

Quant a l'orientació, la majoria d'incendis s'inicien en orientacions sud, amb predomini de l'orientació sud-est. Pel que fa als incendis més grans, aquests s'inicien principalment en orientacions oest i oest-sud-oest.



Imatge 11. Histogrames de la distribució espacial dels incendis per orientació.

Per acabar, les dades de pendent mostren com els incendis es concentren sobretot en zones de pendents petites i mitjanes i son molt escadussers en zones d'altes pendents. També s'observa com els incendis més grans tendeixen a concentrar-se en les zones de pendents més baixes i en disminueix la mida en augmentar la pendent. El següent gràfic il·lustra aquest comportament.



Imatge 12. Histogrames de la distribució espacial dels incendis per pendent.

6.2.3 Meteorologia

Per a l'entrenament dels diversos algorismes d'aprenentatge s'han utilitzat dos conjunts de dades meteorològiques: un primer conjunt provinent del Servei Meteorològic de Catalunya i un segon de l'ECMWF. Malgrat que inicialment s'havia previst utilitzar les dades del Servei Meteorològic de Catalunya no ha estat possible disposar-ne fins a la recta final del projecte. Així és que s'ha optat per l'obtenció de dades meteorològiques alternatives per al conjunt dels experiments i s'ha cenyit l'ús de dades del Servei Meteorològic als algorismes de categorització.

Així, el primer conjunt es troba en forma d'anàlegs: grup de dies anteriors on els estats meteorològics son més similars als del dia actual o futur. Aquest mètode permet la predicció tant de la temperatura com de la precipitació. S'han utilitzat tant per a la temperatura com per a la precipitació el membre de control dels anàlegs ja que es correspon a la predicció més fiable.

Es disposa d'un total de 153 anàlegs corresponents al mesos: juny, juliol, agost i setembre de l'any 2018 i juliol de 2017 format per un conjunt de 11.339 registres. Aquestes dades, doncs,

cobreixen parcialment el període d'estudi i tenen una resolució espacial baixa ja que fan referència globalment al conjunt del territori d'estudi. En conseqüència, s'han utilitzat en la categorització de dies segons les condicions meteorològiques però no en a la resta de models.

Per altra banda, el segon conjunt de dades meteorològiques utilitzat s'ha obtingut dels reanàlisis ERA5 de l'ECMWF. Aquests reanàlisis permeten generar dades climàtiques a partir tant d'observacions com de models. ERA5 disposa d'estimacions horàries des de l'any 1979 fins a l'actualitat de diverses variables meteorològiques a diferents nivells de pressió per a qualsevol punt del planeta.

Per a cobrir les necessitats del projecte s'han obtingut les dades horàries entre els anys 1987 i 2019 a un nivell de pressió atmosfèrica de 1000 hPa, és a dir, en superfície, per al territori de Catalunya. Per a cada incendi s'han obtingut les dades referents al lloc i al moment d'inici d'aquest.

Els paràmetres utilitzats han estat: humitat relativa, humitat específica, temperatura, component *u* (est) del vent i component *v* (nord) del vent. La següent taula resumeix els seus principals estadístics per als diversos punts d'ignició.

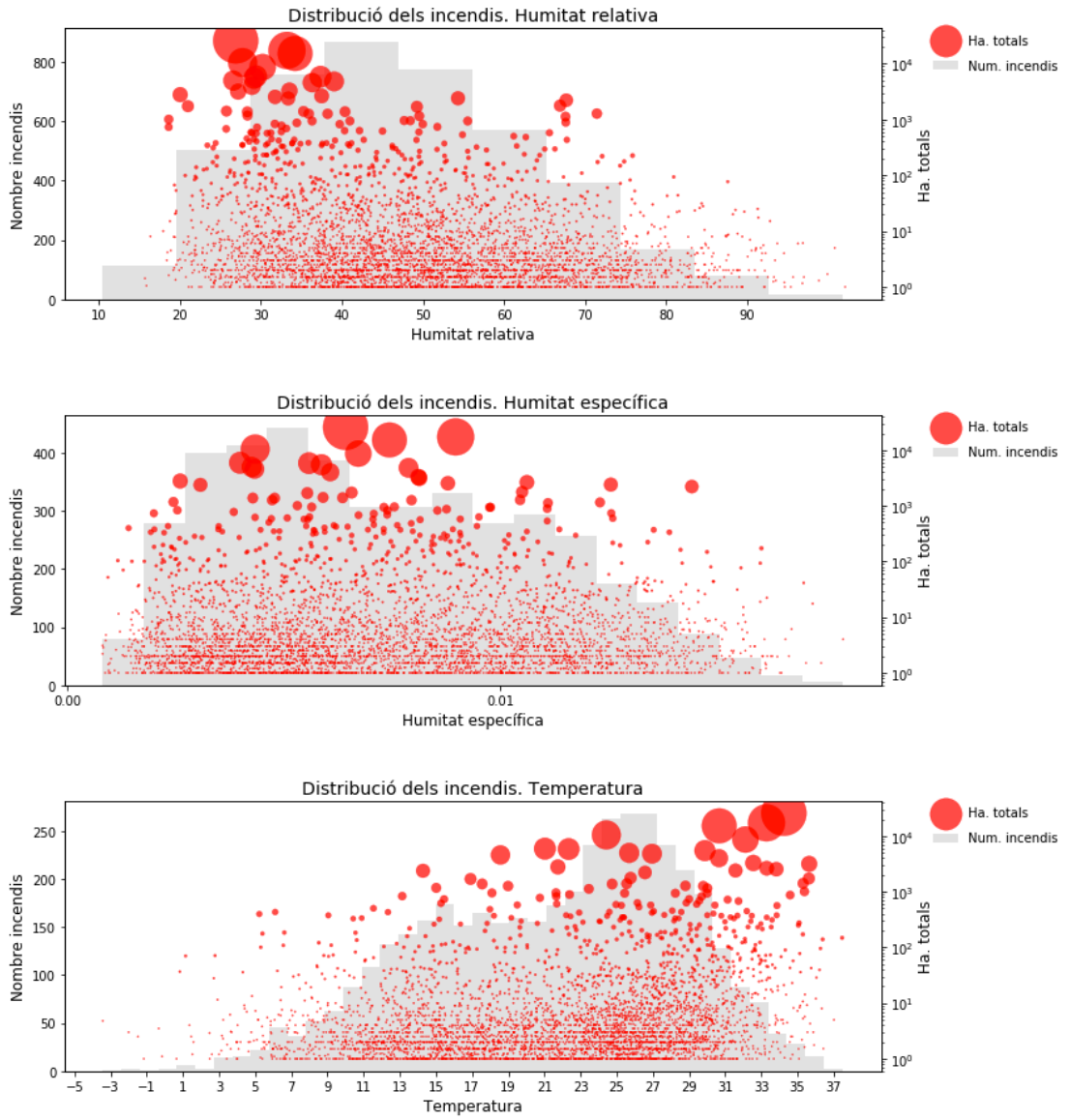
Estadístic	Humitat relativa *(%)*	Humitat específica *(kg)*	Temperatura *(^0C)*[25]	Component ***u*** del vent *(m s^{-1})*	Component ***v*** del vent *(m s^{-1})*
Mínim	10,47	0,00081	-3,43	-9,90	-19,57
1r quantil	34,46	0,00437	15,96	-0,79	-1,27
Mediana	45,73	0,00690	22,73	0,49	1,02
Mitjana	47,13	0,00737	21,49	0,97	0,40
3r quantil	58,67	0,01021	27,02	2,38	2,77
Màxim	101,85	0,01793	37,47	14,73	13,67
Desviació estàndard	16,74	0,00361	7,14	2,89	3,57
Valor absent[26]	1	1	1	1	1

Taula 15. Principals valors estadístics dels atributs de les dades meteorològiques.

En els següents gràfics es mostra la distribució dels incendis en relació a la humitat relativa, la humitat específica i la temperatura en el lloc i moment de l'inici de l'incendi. Com es pot observar, la majoria dels grans incendis s'han produït amb humitats relatives baixes i humitats específiques per sota de la mitjana. Per altra banda, els incendis es concentren en les temperatures més altes, especialment els grans incendis.

[25] Tot i que la unitat de temperatura de les dades originals és Kelvin i que s'ha mantingut en els models, s'han convertit a graus Celsius per tal de mostrar-ne els principals estadístics.

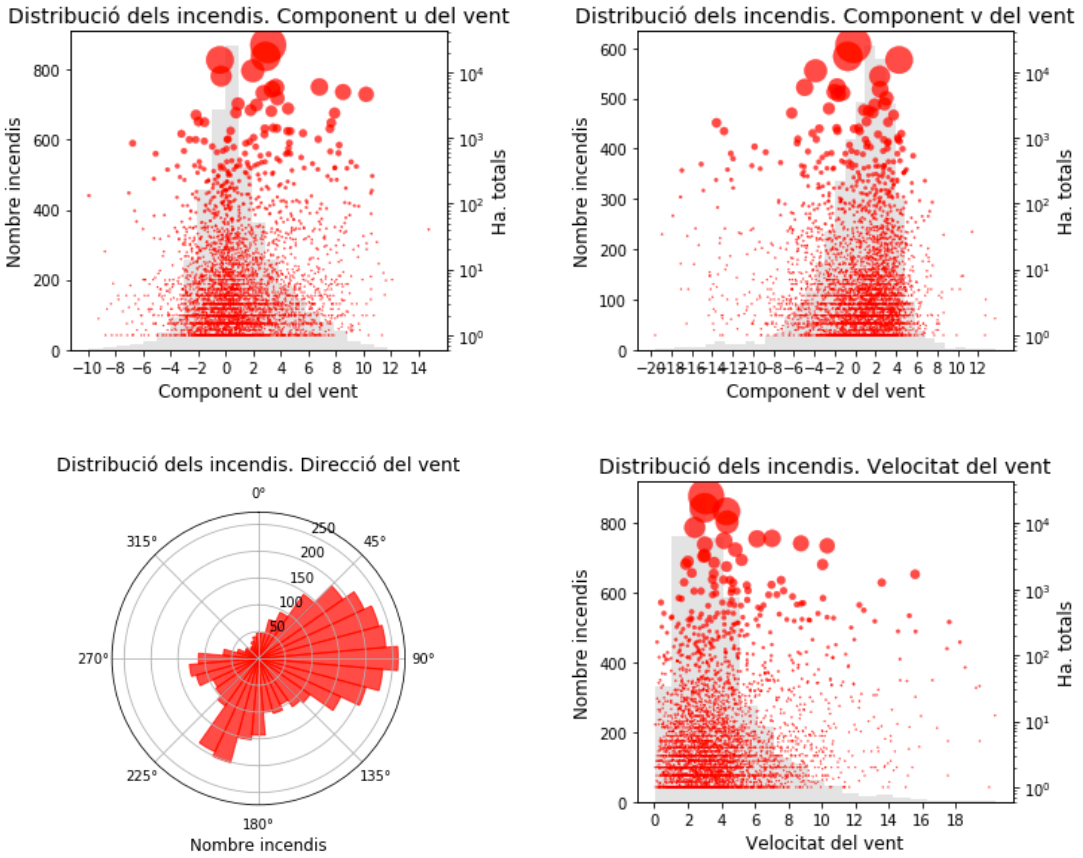
[26] L'únic valor absent en els atributs meteorològics és degut a un exemple que no disposa d'hora d'inici de l'incendi. En eliminar aquest exemple no vàlid, la resta d'exemples no tindran cap valor absent per aquests atributs.



Imatge 13. Histogrames de la distribució dels incendis segons la humitat relativa i específica i la temperatura.

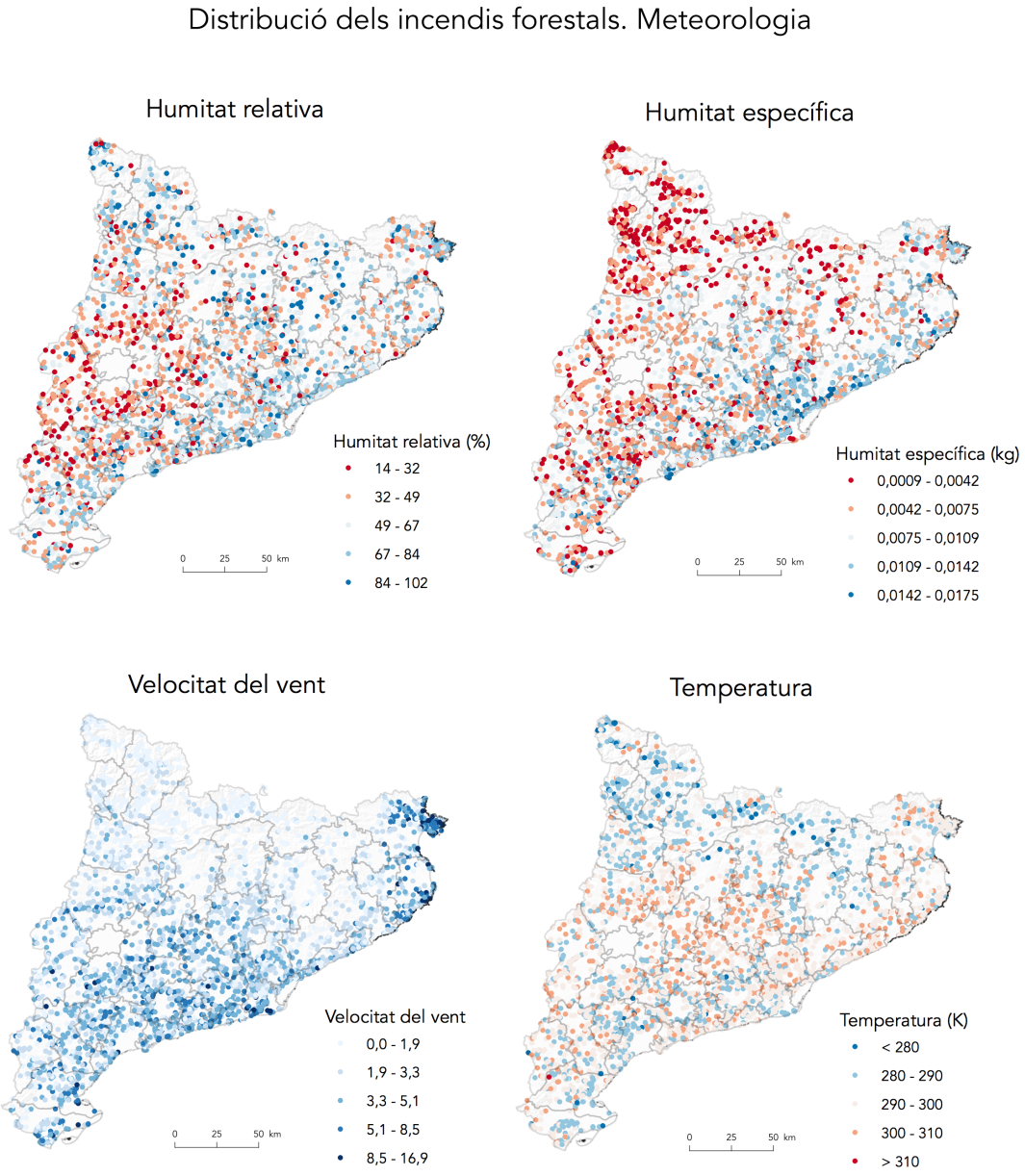
Per altra banda, la distribució dels incendis segons les components *u* i *v* del vent indica un predomini dels valors positius per a la component *u*, és a dir, la direcció est i dels valors negatius per a la component *v*, és a dir, la direcció sud. Per tant, predominen les components est i sud del vent en el moment i lloc d'inici dels incendis. Per altra banda, s'inicien amb velocitats del vent baixes[27].

[27] Les dades referents a la velocitat i a la direcció del vent s'han calculat a partir de les components *u* i *v* del vent per a poder realitzar aquet anàlisi previ. Tot i així, per a l'entrenament dels models no s'han utilitzat aquestes dades derivades sinó les components *u* i *v* originals.



Imatge 14. Histogrames dels incendis segons les components, la direcció i la velocitat del vent.

Finalment, es mostren diversos mapes que ajuden a interpretar la distribució espacial dels incendis en funció de: la humitat relativa, la humitat específica, la temperatura i la velocitat del vent. Podem observar com els incendis en la depressió central es produeixen en situacions d'humitat relativa més baixa. Per altra banda, els incendis produïts a la costa registren una humitat específica més elevada que els que s'inicien a l'interior. Pel que fa a la velocitat del vent, en la costa nord i el sud del país es registren velocitats més altes que en la zona pirenaica. En darrer lloc, els incendis iniciats amb temperatures més baixes es produeixen al nord i al sud i amb temperatures més altes aquests es concentren al centre.



Imatge 15. Relació entre la distribució espacial dels incendis i la humitat, la velocitat del vent i la temperatura.

6.2.4 Vegetació i usos i cobertes del sòl

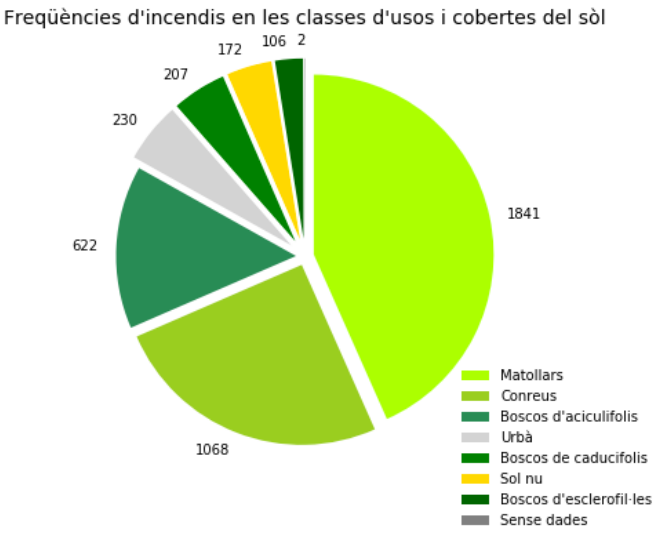
Per a disposar d'informació del combustible vegetal en els models d'aprenentatge s'han utitizat dos conjunts de dades: dades sobre l'estat de la vegetació i els usos i cobertes del sòl.

Les dades de vegetació s'han obtingut a partir de l'índex NDVI (de l'anglès *normalized difference vegetation index*, índex de vegetació de diferència normalitzada) atès que permet estimar la qualitat, quantitat i el desenvolupament de la vegetació a partir de dades de la intensitat de radiació de diverses bandes de l'espectre electromagnètic. Finalment, s'ha assignat a cada incendi l'índex corresponent al dia disponible immediatament anterior al de la ignició ja que l'índex del mateix dia o dies posteriors pot estar afectar pel propi incendi.

Quant als usos i cobertes del sòl, s'han utilitzat els mapes disponibles corresponents als anys: 1987, 1992, 1997, 2002, 2007, 2012 i 2017, els quals contenen informació sobre els diversos tipus de coberta obtinguda per classificació automàtica a partir d'imatges Landsat[28].

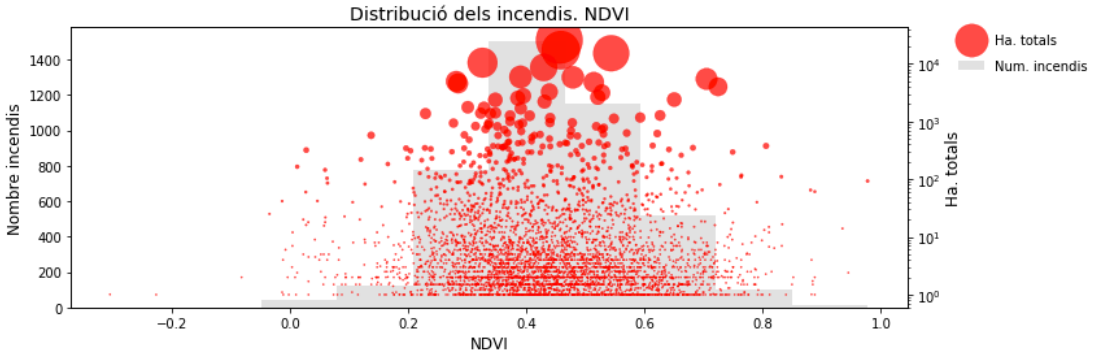
La següent taula mostra un resum dels principals estadístic de les dades de vegetació i dels usos i cobertes del sòl dels punts d'ignició i el gràfic que l'acompanya la freqüència dels incendis iniciats en cadascuna de les vuit classes d'usos i cobertes del sòl. En conjunt, la majoris d'incendia s'inicien en zones de matollar o conreus, seguit de les zones de bosc.

Estadístic	NDVI	Usos i cobertes
Mínim	-0,305	
1r quantil	0,347	
Mediana	0,436	
Mitjana	0,442	
3r quantil	0,537	
Màxim	0,978	
Desv. Est.	0,144	
Valors únics		8 classes
Valor més freqüènt		Matollars
Freq. valor més freqüent		1841
Registres no vàlids	13	0



Taula 16. Estadístics dels atributs vegetació, usos i cobertes del sòl. Freqüència d'incendis en les classes d'usos i cobertes.

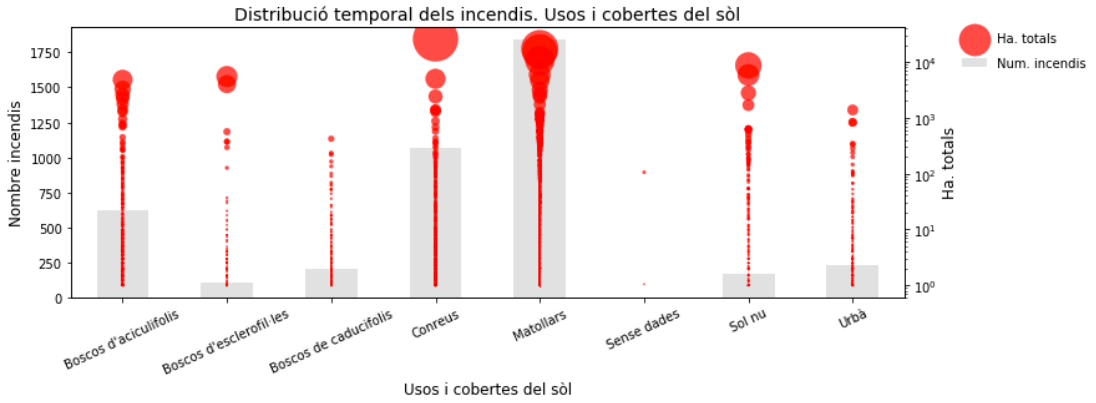
Per altra banda, la majoria dels incendis s'inicien en indrets amb un índex NDVI entre 0,2 i 0,7 amb una mitjana de 0,44 corresponents a llocs amb vegetació, contretament matollar per als valors més baixos i vegetació alta per als valors més elevats de l'índex.



Imatge 16. Histograma de la distribució dels incendis en relació a l'índex NDVI.

[28] https://en.wikipedia.org/wiki/Landsat_program

Finalment, s'ha observat que els incendis més gran s'han iniciat en matollars i conreus. Per altra banda, els incendis iniciats tant en zones urbanes com en boscos caducifolis no acostumen a ser grans incendis.



Imatge 17. Histograma de la distribució dels incendis en relació als usos i cobertes del sòl.

6.3 Preparació de les dades

El tractament previ de les dades ha consistit en: l'obtenció d'exemples negatius, la transformació de dades, el tractament dels valors absents, l'estandardització i normalització de dades, la reducció de la dimensionalitat i la generació dels conjunts d'exemples.

6.3.1 Obtenció d'exemples negatius

El conjunt d'exemples d'incendis s'ha ampliat amb una quantitat equivalent de nous casos aleatoris d'exemples de no incendi per tal d'obtenir un conjunt equilibrat on no predomini una de les dues classes. Després d'haver generat un total de 5.309 exemples aleatoris s'ha obtingut la correlació de Pearson en relació als exemples positius amb la finalitat d'estimar si existien exemples excessivament similars als incendis. En conseqüència, s'ha eliminat el 20% dels nous casos negatius amb major similitud als incendis obtenint un conjunt final de 4.250 no incendis.

Per avaluar la idoneïtat d'aquesta metodologia on s'ha tingut en compte la similitud de Pearson alhora d'acceptar nous exemples negatius, s'han comparat els resultats obtinguts en diversos classificadors utilitzant, per un costat, el conjunt inicial i, per l'altre, el conjunt on s'ha tingut en compte la similitud de Pearson. La següent taula mostra els resultats obtinguts per la mètrica *F1 score* on s'aprecia una millora un cop eliminat el 20% dels nous exemples amb major similitud.

Mètode	*Configuració*	Mesura F1[29]	
		Conjunt original	Conjunt Pearson
KNN	Veïns=8	0,82	0,84
SVM	*Kernel* gaussià[30]	0,82	0,83
Arbres de decisió	Profunditat màxima=5	0,77	0,80
Random forest	Profunditat màxima=5	0,81	0,84
AdaBoost		0,83	0,85
Xarxa neuronal	Capa oculta: 10 unitats	0,84	0,85
Mitjana		**0,815**	**0,835**

Taula 17. Comparació del comportament dels conjunts original i un cop aplicada la similitud de Pearson.

6.3.2 Transformació

Les principals transformacions de dades dutes a terme han estat el remostreig (en anglès, *resampling*) i la reclassificació de capes ràster així com l'obtenció de valors orogràfics a diversos radis dels punts d'inici dels incendis.

A causa de la diferència de resolució espacial entre les diverses dades disponible s'ha optat per aplicar un remostreig a aquelles amb menys resolució. Concretament, les dades meteorològiques procedents del sensor MODIS amb una resolució inicial de 0,25° x 0,25° han estat remostrejades a una resolució de 0,01° x 0,01° mitjançant el mètode B-Spline[31].

En canvi, les dades d'usos i cobertes del sòl disponibles contenien un excés de resolució temàtica per a les finalitats del projecte i, per tant, s'han reclassificat per a obtenir un nombre de classes més reduït. La següent taula mostra les classes originals així com la seva reclassificació.

Classe	2017	2012-2007	2002-1997 1992-1998	Reclassificació
NODATA	✓	✓	✓	NODATA
Aigües continentals	✓	✓	✓	Sòl nu
Aigües marines	✓	✓	✓	
Congestes	✓	✓	✓	
Infraestructures viàries	✓	✓	✓	Urbà
Urbanitzacions	✓	✓	✓	
Zones urbanes	✓	✓	✓	
Zones industrials i comercials	✓	✓	✓	
Conreus herbacis de secà	✓	✓	✓	Conreus
Conreus herbacis de regadiu	✓	✓	✓	
Fruiters de secà	✓	✓	✓	
Fruiters de regadiu	✓	✓	✓	
Vinyes	✓	✓	✓	
Prats supraforestals	✓	✓	✓	Matollar
Matollars	✓	✓		
Bosquines i prats			✓	
Prats de terra mitjana	✓	✓		
Prats de terra baixa	✓	✓		
Boscos d'esclerofil·les	✓	✓	✓	Boscos d'esclerofil·les
Boscos de caducifolis	✓	✓	✓	Boscos de caducifolis
Boscos d'aciculifolis	✓	✓	✓	Boscos d'aciculifolis

[29] En anglès, *F1 score*.
[30] També anomenat *RBF kernel*.
[31] https://ca.wikipedia.org/wiki/B-spline

Vegetació de zones humides	✓	✓	✓	Matollar
Zones amb vegetació escassa o nul·la	✓	✓	✓	Sòl nu
Zones cremades	✓	✓	✓	
Sorrals i platges	✓	✓	✓	
Arrossars	✓			Conreus
Cítrics	✓			

Taula 18. Classes originals dels mapes d'usos i cobertes del sòl i la seva reclassificació.

Per acabar, s'han obtingut valors d'altitud, orientació, pendent i l'índex de rugositat de la superfície per a diversos radis al voltant dels punts d'inici dels incendis.

6.3.3 Tractament dels valors absents

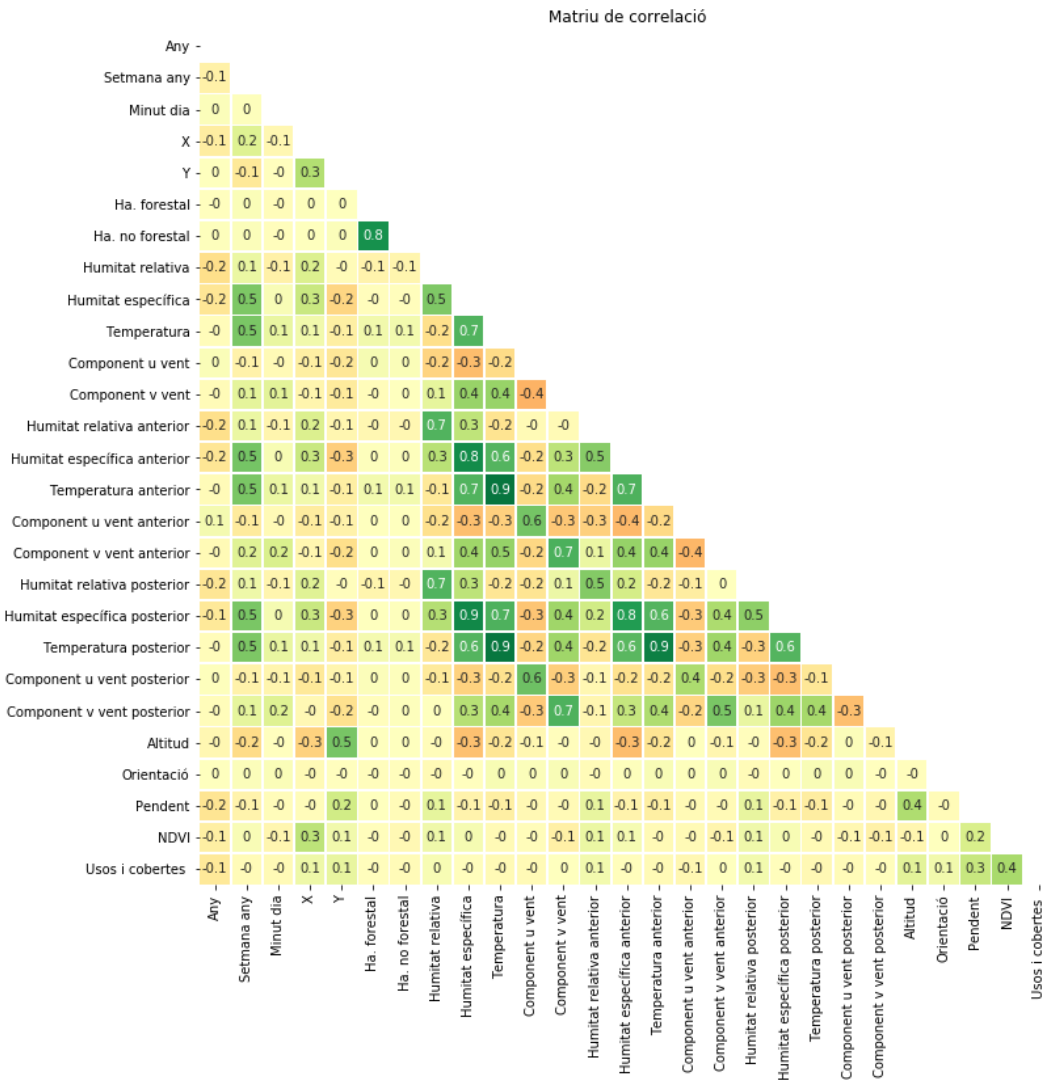
El tractament de valors absents ha estat necessari en diversos atributs amb valors desconeguts. En primer lloc, en el cas dels exemples sense hora d'inici de l'incendi, aquests han estat eliminats del conjunt d'exemples ja que en no disposar d'aquest atribut no era possible obtenir-ne els atributs meteorològics amb prou precisió. En segon lloc, els casos sense dades d'orientació en trobar-se en una zona sense pendent s'han substituït per la mitjana dels valors disponibles. En tercer lloc, als exemples sense índex NDVI se'ls ha assignat el valor de la cel·la ràster més propera. En últim terme, als casos amb altituds errònies en incendis produïts a la costa se'ls ha assignat la cota zero.

6.3.4 Estandardització i normalització de dades

Totes les dades han estat estandarditzades abans de ser utilitzades en els diversos algorismes d'aprenentatge per evitar els efectes del desplaçament i escales en els valors dels atributs. Per altra banda, les dades han estat normalitzades abans d'entrenar les xarxes neuronals per afavorir el temps de convergència de l'optimització d'aquestes.

6.3.5 Reducció de la dimensionalitat

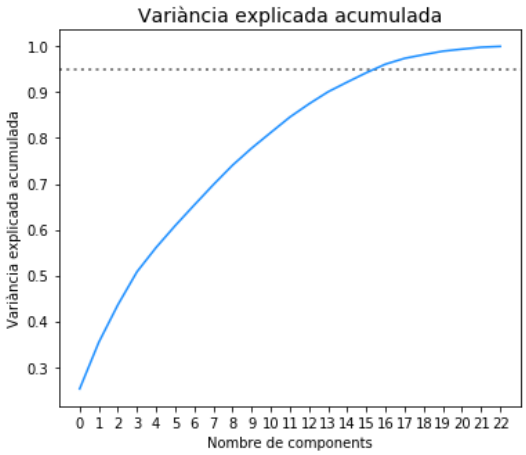
Abans d'iniciar l'extracció de característiques per a la reducció de la dimensionalitat, s'ha analitzat la relació lineal entre els diversos atributs amb els coeficients de Pearson on s'han identificat nombroses correlacions. De fet, moltes d'aquestes eren previsibles, com en el cas de les variables meteorològiques dels diversos moments tinguts en compte en el conjunt d'exemples: l'inici de l'incendi, 24 hores abans i 24 hores després. També s'aprecien correlacions previsibles entre l'índex NDVI i els usos i les cobertes del sòl, l'altitud i la pendent així com entre la superfície forestal i no forestal cremades. Per altra banda, s'identifiquen correlacions entre la humitat específica i la temperatura així com entre les components del vent. En aquest darrer cas, s'ha detectat una correlació negativa, tot i que poc significativa, entre les dues components del vent.



Imatge 18. Matriu de correlació de Pearson dels atributs.

La correlació detectada planteja la utilització d'una anàlisi de components principals (en anglès, *principal component analysis*, PCA) per a minimitzar la correlació estadística entre els atributs per tal de representar les dades en un espai de dimensionalitat inferior utilitzant les components principals que expliquin un cert percentatge de la seva variabilitat. Concretament, se cercaran els components principals que expliquin, com a mínim, el 95% de la variabilitat.

El resultat de l'anàlisi indica que amb els disset primers components principals és possible explicar el 96,13% de la variabilitat del conjunt de dades. El següent gràfic mostra la variància explicada acumulada per a diferents quantitats de components principals.



Taula 19. Variància explicada acumulada.

La projecció del conjunt original a un espai PCA de dimensionalitat reduïda presenta importants avantatges com l'augment del rendiment dels algorismes en reduir-se el nombre d'atributs i la reducció del sobreentrenament (en anglès, *overfitting*) ja que aquest es pot produir quan hi ha un nombre elevat d'atributs. Tot i així, en l'apartat 7.3 s'ha analitzat el comportament de cadascun dels algorismes en aplicar PCA ja que no sempre aquest millora els resultats obtinguts.

6.3.6 Generació dels conjunts d'exemples

S'han generat tres conjunts diferents per a resoldre els tres problemes plantejats: la predicció del risc d'incendi forestal, la predicció de la mida en el moment de la ignició i l'agrupament de les zones de risc segons les condicions meteorològiques. Concretament, per a resoldre el tercer problema s'han emprat dos conjunt: un primer amb dades meteorològiques provinents d'anàlegs i un segon amb dades provinents del reanàlisi.

Pel que fa al conjunt d'exemples positius, gràcies als diversos mètodes de tractament de valors absents, només ha estat necessari eliminar un exemple. Pel que fa als casos de no incendi, com s'ha comentat anteriorment, s'han eliminat el 20% dels exemples originals.

Problema	Conjunt	Mida
Predicció del risc d'incendi forestal	Incendis + no incendis	8.498
Predicció de la mida en el moment de la ignició	Incendis	4.247
Agrupament de les zones de risc segons les condicions meteorològiques	Incendis + no incendis	8.498
	Anàlegs	153

Taula 20. Mida dels conjunts d'exemples.

D'altra banda, els conjunts s'han dividit en un conjunt d'entrenament format pel 70% dels casos i un conjunt de test amb el 30% d'aquests. S'ha assegurat que la quantitat d'exemples pertanyents a les diverses classes en cadascun dels conjunts fos proporcional. Finalment, s'ha emprat validació creuada en comptes d'un conjunt de validació ja que la quantitat de dades no és molt alta.

7 Models d'aprenentatge

Com s'ha indicat anteriorment, s'han plantejat tres possibles models d'aprenentatge: models de categorització per a l'agrupament de zones segons les condicions meteorològiques, models de regressió per a estimar la mida dels incendis en el moment d'ignició i models de classificació que permetin estimar el risc d'incendi forestal.

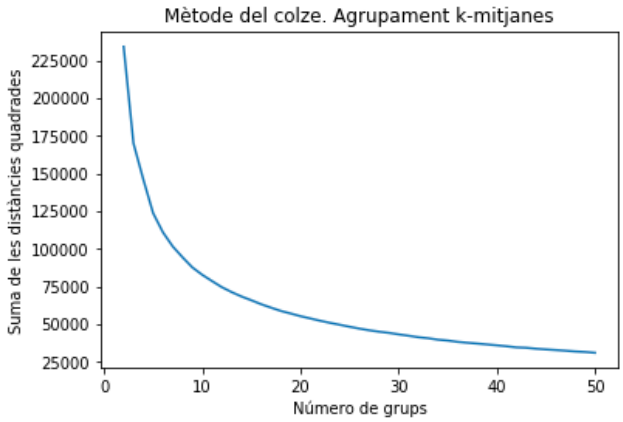
7.1 Algorismes d'agrupament de zones de risc d'incendi forestal segons les condicions meteorològiques

Com s'ha detallat en l'apartat 6.2.3, per realitzar els experiments s'han utilitzat dos conjunts de dades meteorològiques: reanàlisis i anàlegs. Tot seguit es detallen els experiments duts a terme amb cadascun dels dos tipus de dades.

7.1.1 Basats en reanàlisis ERA5

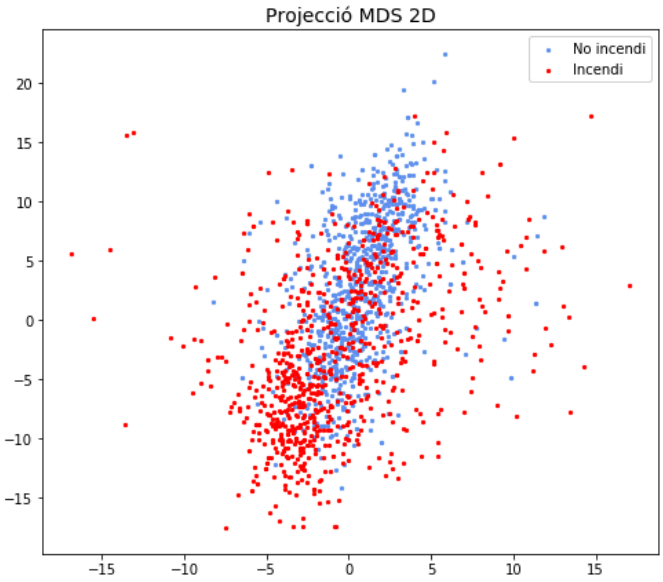
Per a poder aplicar els algorismes d'agrupament, en primer lloc ha estat necessari analitzat el nombre de grups existents en el conjunt de dades meteorològiques atès que aquest no es coneixia.

El primer mètode utilitzat ha estat el del colze ja que sovint permet identificar el nombre de clústers de les dades. El següent gràfic mostra el resultat obtingut amb el mètode d'agrupament k-mitjanes i un nombre creixent de grups. Tot i que s'aprecia una inflexió entre els 5 i 10 grups, no s'identifica amb claredat un colze i, per tant, no permet estimar el nombre de grups de les dades de forma precisa.



Imatge 19. Mètode del colze aplicat als reanàlisis.

Tot seguit, s'ha projectat un subconjunt del 20% de les dades originals en un espai 2D de dimensionalitat reduïda amb escalament multidimensional per, així, poder-les visualitzar. El següent gràfic en mostra la projecció utilitzant una mètrica euclidiana on es confirma el que s'havia obtingut amb el mètode del colze, és a dir, que els grups no son clarament diferenciables.



Imatge 20. Reprojecció del conjunt de dades en un espai 2D.

Per aquest fet, s'ha optat per utilitzar mètodes d'agrupament que permetin estimar directament el nombre idoni de clústers: DBSCAN, OPTICS, propagació de l'afinitat (en anglès, *affinity propagation*) i BIRCH.

DBSCAN, el primer d'aquests mètodes, ha obtingut resultats no vàlids ja que el mètode tendeix a agrupar tots els casos en un sol clúster si la distància establerta per a les dades atípiques és alta o bé deixa molts exemples sense classificar si aquesta distància és petita[32]. La utilització del mètode OPTICS, generalització del mètode DBSCAN, no n'ha millorat els resultats.

Distància	Núm. clústers	*Outliers*
0.1	0	8494
0.2	11	8070
0.3	15	4044
0.4	5	1796
0.5	1	859
1	1	76
2	1	0

Taula 21. Nombre de clústers i outliers obtinguts amb el mètode DBSCAN.

Per aplicar el segon mètode, propagació de l'afinitat, s'ha utilitzat un subconjunt del 50% dels exemples disponibles mantenint la mateixa proporció d'exemples positius i negatius per a l'obtenció dels grups atès que la complexitat computacional de l'algorisme és d'ordre $O(N^2)$. Per avaluar-ne els resultats, s'ha utilitzat: el coeficient de silueta i l'índex Davies Bouldin. El primer té en compte la mitjana de la similitud dels casos en un clúster així com la distància d'aquests al clúster més proper per estimar la bondat de l'agrupament. Els seus valors varien entre 1 i -1, essent els més alts els que indiquen una bona configuració dels clústers. El segon

[32] Cal tenir en compte que el conjunt està format per 8.494 grups i en algunes configuracions s'arriba a considerar tots aquests dades atípiques.

correspon al rati entre les distàncies dins dels clústers i entre clústers. En aquest cas els valors més baixos indiquen un millor agrupament.

En els diversos experiments duts a terme, els que han obtingut els millors valors per a les dues mètriques indiquen un total de 10 clúster. Altres configuracions amb bons resultats han generat 9, 11 i 12 clúster. La següent taula mostra els millors resultats dels experiments duts a terme on s'assenyala: els principals paràmetres de l'algorisme, el nombre de clústers obtinguts i els dos coeficients emprats.

Preference	*Damping*	Núm. clústers	Coeficient de silueta	Índex Davies Bouldin
-60	0,9	12	0,400	1,141
-65	0,95	11	0,404	1,146
-70	0,95	10	0,413	1,129
-85	0,95	9	0,411	1,174

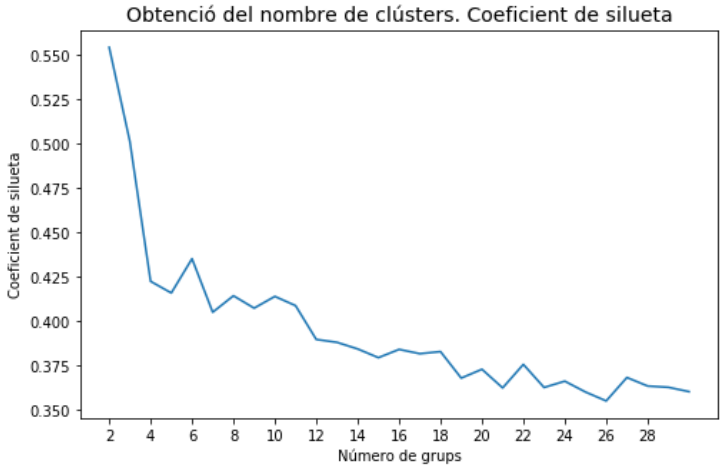
Taula 22. Millors resultats de la bateria de proves de l'algorisme de propagació de l'afinitat.

Tot seguit, s'ha utilitzat el mètode BIRCH per a comparar el resultat amb l'aconseguit amb l'anterior mètode on s'ha obtingut el mateix nombre de clústers. La següent taula en mostra els millors resultats.

Factor de fuga	Llindar	Núm. clústers	Coeficient de silueta	Índex Davies Bouldin
25	0,5	28	0,346	1,179
	0,6	12	0,370	1,195
	0,62	11	0,392	1,197
	0,64	10	0,406	1,155
	0,7	6	0,395	1,229
	0,5	26	0,347	1,172

Taula 23. Resultat de la bateria de proves de l'algorisme Birch.

Finalment, s'ha analitzat el coeficient de silueta amb un nombre creixent de clústers per al mètode d'agrupament k-mitjanes amb l'objectiu avaluar-ne el comportament. Com s'aprecia en els següent gràfic, els millors resultats s'assoleixen per a configuracions de 2 i 3 grups per bé que cal descartar-los ja que és un nombre poc útil per a la finalitat que es persegueix. També cal tenir en compte que es tracta d'un resultat previsible ja que, com s'ha pogut observar en la reprojecció del conjunt d'exemples en un espai 2D, els diversos casos son molt similars entre si i, per tant, un nombre molt baix de grups comporta que les distàncies entre els seus membre sigui petita. És per aquest motiu que s'obtenen valors per al coeficient de silueta alts. Per altra banda, s'observa que entre els 4 i 11 grups s'obtenen resultats similars per a, tot seguit, davallar a partir dels 12 grups. Així, es confirmar que un total de 10 grups ofereix una bona configuració.

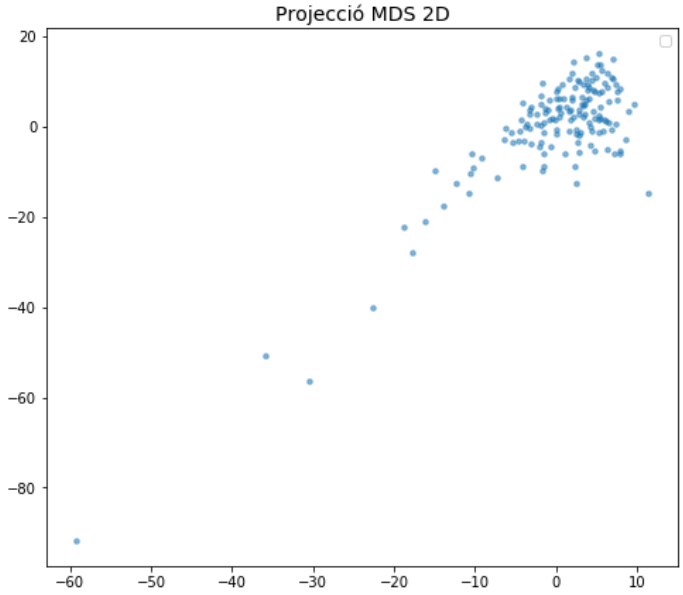


Imatge 21. Coeficient de silueta.

Un cop obtingut el nombre de grups s'han utilitzat les dades meteorològiques del conjunt d'exemples per a l'entrenament de diversos models de categorització: k-mitjanes, agrupament jeràrquic algomeratiu, BIRCH, propagació de l'afinitat i agrupament espectral. En l'apartat 8.1 s'avaluen els resultats obtinguts.

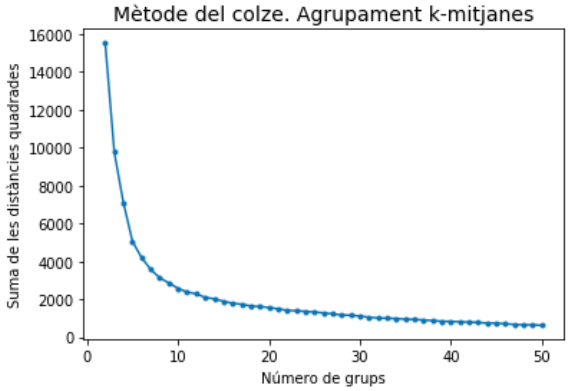
7.1.2 Basats en anàlegs

Tot i que ja existeix un agrupament inicial en els propis anàlegs on cadascun d'aquests està format per un grup d'una mitjana de 74 dies similars, s'ha realitzat un segon agrupament d'aquests vist que molts dels anàlegs es corresponen a condicions meteorològiques força semblants. El següent gràfic mostra la projecció 2D de dismensionalitat reduïda dels centroides de cadascun dels anàlegs, on s'aprecia la gran semblança de la majoria dels anàlegs.



Imatge 22. Projecció en un espai 2D dels centroides dels anàlegs.

Com en el cas dels reanàlisis, per establir el nombre de clústers s'ha iniciat l'anàlisi amb el mètode del colze. Malauradament, el resultat obtingut no indica de forma clara el nombre de grups.



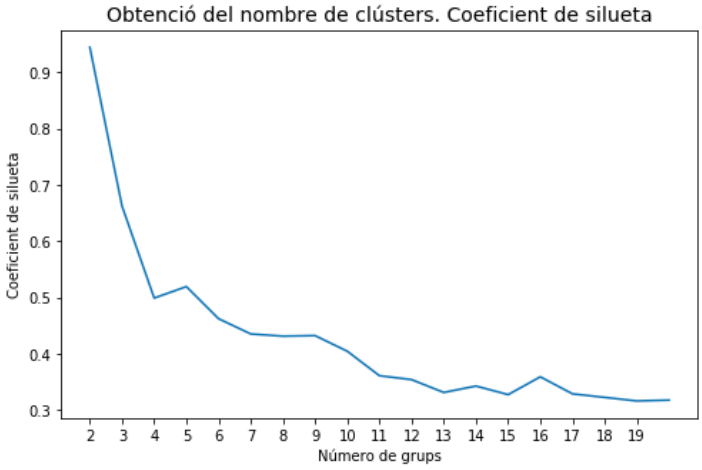
Imatge 23. Mètode del colze aplicat als anàlegs.

És per aquest motiu que, com ja s'havia realitzat amb les dades meteorològiques provinents del reanàlisi del *Climate Data Store*, s'han utilitzat els mètodes d'agrupament per avaluar el nombre idoni de clústers: DBSCAN, OPTICS, propagació de l'afinitat, *Mean shift* i BIRCH. Els mètodes DBSCAN i OPTICS no han obtingut resultats vàlids ja que, o bé han agrupat tots els exemples en un sol clúster o bé han obtingut tants grups com exemples. La resta de mètodes han obtingut 5, 8 i 4 grups respectivament. Degut a què les millors mètriques s'han assolit amb els mètodes propagació de l'afinitat i BIRCH s'ha escollit el resultat d'un d'aquests, concretament, el mètode de propagació de l'afinitat i, per tant, un nombre final de 5 clústers.

Mètode	Coeficient de silueta	Índex Davies Bouldin	Núm. clústers
DBSCAN	-	-	-
OPTICS	-	-	-
Propagació de l'afinitat	0,814	0,321	5
Mean shift	0,544	0,447	8
BIRCH	0,842	0,347	4

Imatge 24. Estimació del nombre de clústers dels anàlegs.

L'anàlisi del coeficient de silueta amb a un nombre creixent de clústers per al mètode de categorització k-mitjanes mostra la bondat dels resultats obtinguts. Deixant de banda un nombre excessivament petit com 2 o 3 clústers, la mètrica obté valors alts per a 4 o 5 clústers disminuint tot seguit.



Imatge 25. Coeficient de silueta.

7.2 Algorismes de regressió per a la predicció de la mida dels incendis forestals en el moment de la ignició

El segon tipus d'algorismes d'aprenentatge automàtic que es proposa en aquest projecte té com a objectiu avaluar la possibilitat de predir la mida dels incendis en el moment d'ignició.

Amb aquesta finalitat, s'han implementat diversos algorismes de regressió: kNN, màquines de vectors de suport, AdaBoost, regressió logística, regressió *Kernel ridge* i xarxes neuronals. Per altra banda, també s'han combinat arbres de decisió amb AdaBoost amb la finalitat de millorar els resultats obtinguts. Aquets s'avaluen en l'apartat 8.2.

7.3 Algorismes de classificació per a la predicció del risc d'incendi forestal

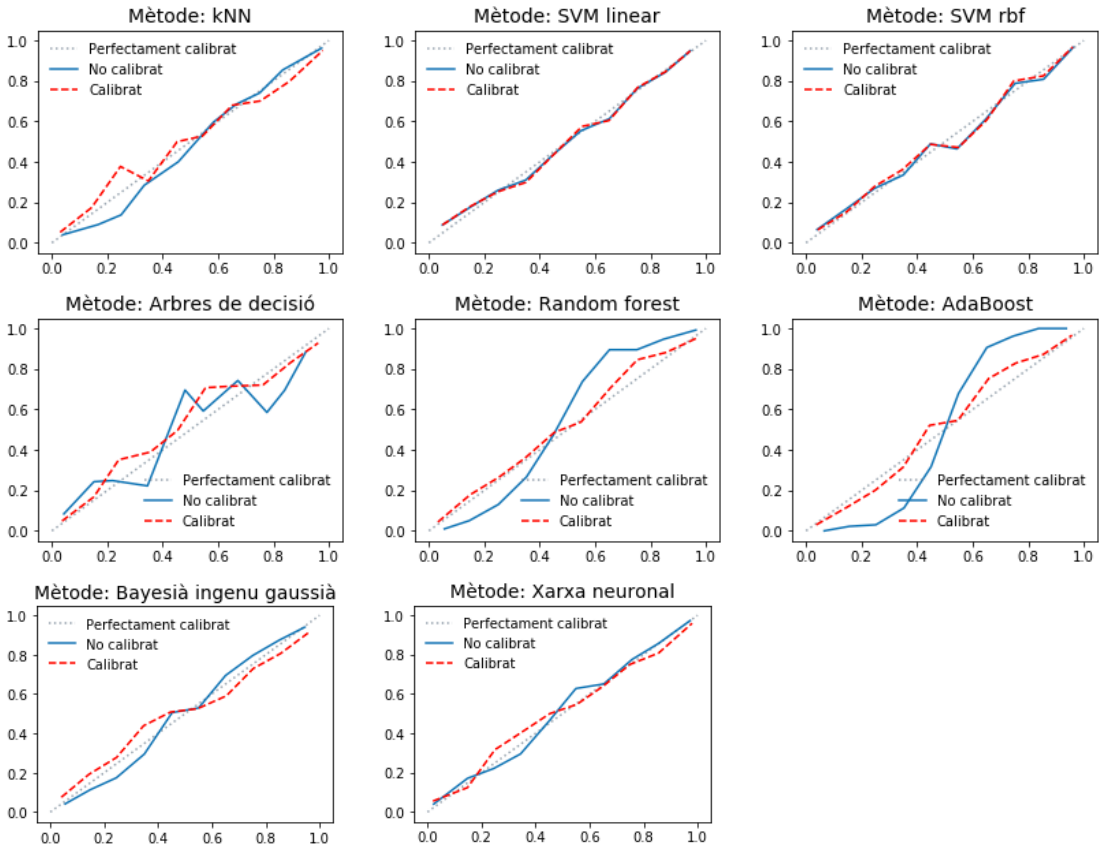
El darrer estudi dut a terme ha estat l'ús d'algorismes de classificació amb la finalitat d'estimar el risc d'incendi forestal. Els algorismes avaluats han estat: kNN, màquines de vectors de suport, arbres de decisió, *Random forest*, AdaBoost, bayesià ingenu gaussià i xarxes neuronals.

Com s'ha comentat en l'apartat 6.3.5, on s'ha aplicat un anàlisi de components principals, aquesta tècnica no sempre millora els resultats dels algorismes. És per això que s'ha analitzat l'efecte del mètode PCA en cadascun d'aquests. En el cas dels mètodes kNN, màquines de vectors de suport i la xarxa neuronal l'aplicació d'aquesta tècnica no ha modificat significativament els resultats obtinguts. Per altra banda, en els cas del mètode bayesià ingenu gaussià hi ha hagut una millora de les mètriques obtingudes. Així, per aquests algorismes s'ha emprat la tècnica PCA atès que o bé en milloren els resultats o bé el rendiment. Per contra, en el cas dels arbres de decisió, *Random forest* i AdaBoost, els resultat han empitjorat; de manera que no s'utilitzarà PCA en aquests algorismes.

Per altra banda, cal tenir en compte que per a obtenir l'estimació del risc d'incendi no s'ha utilitzat la classificació dels exemples en una de les dues classes: incendi i no incendi, sinó la probabilitat que l' exemple pertanyi a la classe incendis. Donat que la predicció de la probabilitat depèn molt dels algorismes utilitzats s'ha realitzat un anàlisi del seu comportament a fi d'ajustar

la distribució de les probabilitats amb la calibració de les prediccions. En primer lloc, s'han obtingut les corbes de calibració, també anomenades diagrames de fiabilitat, que mostren la freqüència relativa dels exemples en relació a la freqüència de la probabilitat predita. La separació de les corbes en relació a la diagonal mostra la bondat de la predicció, essent millors les prediccions ajustades a la diagonal.

La calibració s'ha dut a terme amb validació creuada utilitzant un conjunt de test en cada validació per a calibrar les probabilitats predites. Els següents diagrames de fiabilitat il·lustren el comportament dels algorismes abans i després de ser calibrats.



Imatge 26. Diagrames de fiabilitat dels diversos algorismes.

Com es pot observar, els classificadors que més es beneficien de la calibració de probabilitats son: els arbres de decisió, *Random forest* i AdaBoost. En els diversos experiments duts a terme s'ha observat que tan sols en el cas de la xarxa neuronal la calibració ha empitjorat la predicció inicial. Per tant, s'ha aplicat aquest mètode abans d'obtenir els mapes de risc d'incendi forestal en tots els algorismes excepte en la xarxa neuronal.

8 Avaluació dels models

Un cop s'han obtinguts els diversos models d'aprenentatge aquests han estat avaluats amb tres finalitats: estimar-ne l'adequació a la resolució dels diversos problemes plantejats, escollir-ne els més adients i plantejar possibles estratègies per a la seva optimització. Aquesta avaluació s'ha realitzat amb un conjunt de test format pel 30% del exemples originals i garantint en tot moment que la proporció de casos de cada classe fos similar a la del conjunt d'entrenament.

8.1 Avaluació dels algorismes d'agrupament de zones de risc d'incendi forestal segons les condicions meteorològiques

Per a avaluar els models d'agrupament obtinguts, s'han utilitzat tres mètriques diferents: el coeficient de silueta, l'índex Davies Bouldin i el temps necessari per a l'obtenció del model. Per altra banda, també s'ha tingut en compte la distribució d'exemples d'incendi i no incendi en cadascun dels grups així com la superfície total cremada en aquests.

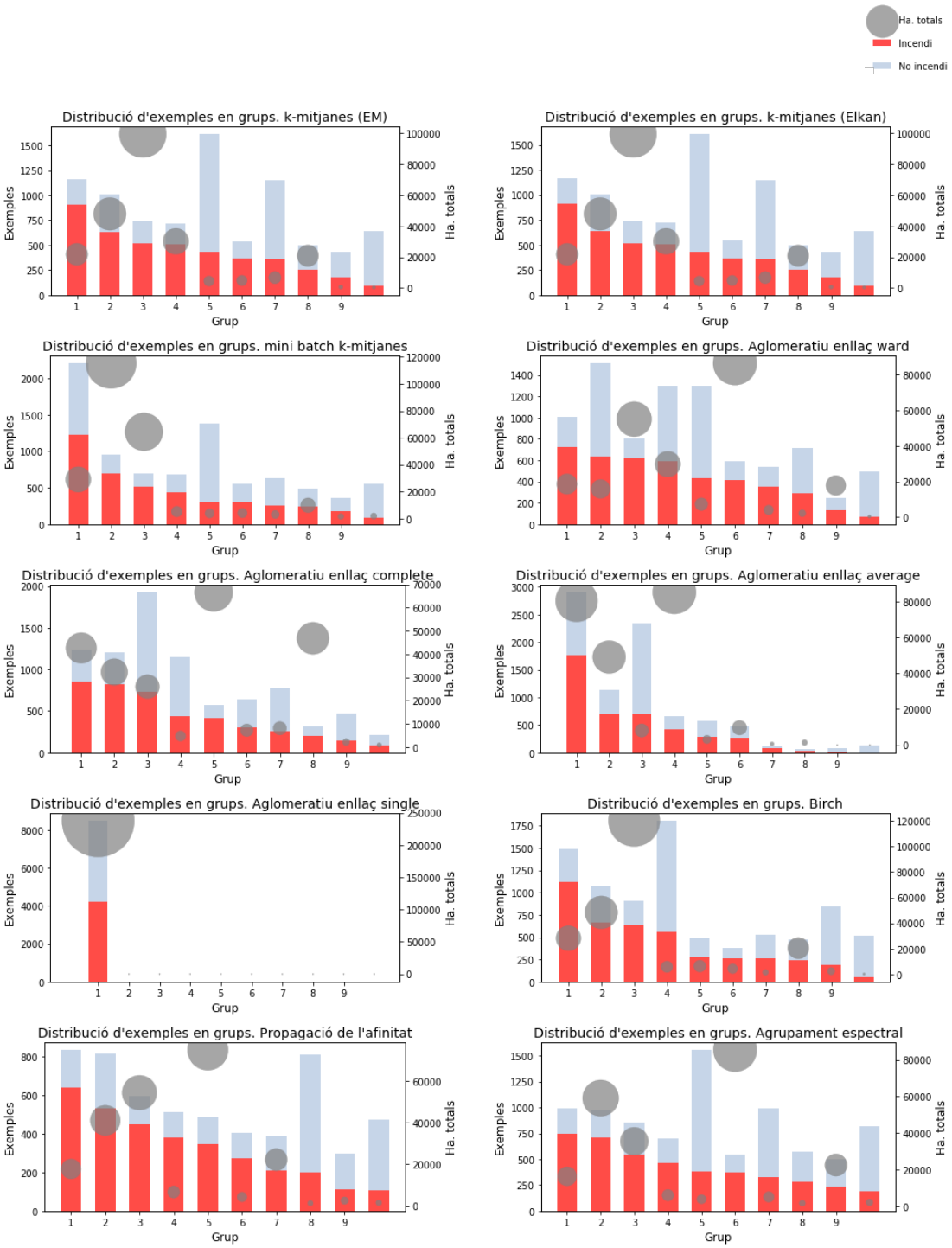
8.1.1 Basats en reanàlisis ERA5

La següent taula i els gràfics de distribució de casos en grups que l'acompanyen mostren els resultats obtinguts amb els mètodes d'agrupament basats en reanàlisis ERA5. Un cop analitzada la distribució de casos d'incendi i no incendi en els diversos clústers s'han descartat els resultats obtinguts amb el mètode d'agrupament jeràrquic utilitzant unió simple i mitjana, ja que el primer agrupa el 99,8% dels exemples en un sol clúster i el segon el 82,9% en tan sols quatre clústers. En canvi, en la resta de mètodes la gran majoria de grups mostren un predomini d'una de les dues classes: o bé incendis o bé no incendis.

Per altra banda, les millors mètriques s'han assolit amb els mètodes k-mitjanes i BIRCH. Els segueixen els mètode d'agrupament espectral i de propagació de l'afinitat; en aquest darrer cas, es descarta el mètode no només per haver generat uns grups amb una configuració lleugerament pitjor que els mètodes anteriors sinó pel seu elevat cost computacional. De fet, ha estat el mètode que ha requerit més temps per a l'agrupament dels exemples.

Mètode	Configuració	Coeficient de silueta	Índex Davies Bouldin	Temps *(segons)*
k-mitjanes	Inicialització *k-mitjanes ++*	0,415	1,136	2,30
k-mitjanes (Elkan)	Inicialització *k-mitjanes ++*	0,414	1,136	1,67
k-mitjanes (mini lots)	Inicialització *k-mitjanes ++*, *mida dels lots=20*	0,384	1,276	0,41
Agrupament jeràrquic aglomeratiu	Mètode d'unió: guarda	0,248	1,328	17,79
	Mètode d'unió: complet	0,184	1,585	5,96
	Mètode d'unió: mitjana	0,268	1,190	8,33
	Mètode d'unió: simple	-0,512	1,012	4,01
BIRCH	Factor de fuga=25, llindar=0,64	0,406	1,155	0,85
Propagació de l'afinitat	*Preference*=-70, *damping*=0.95	0,377	1,104	1.122,78
Agrupament espectral	Matrius d'afinitat: *kernel* gaussià	0,386	1,186	146,51

Taula 24. Mètriques dels diversos mètodes d'agrupament comparats.



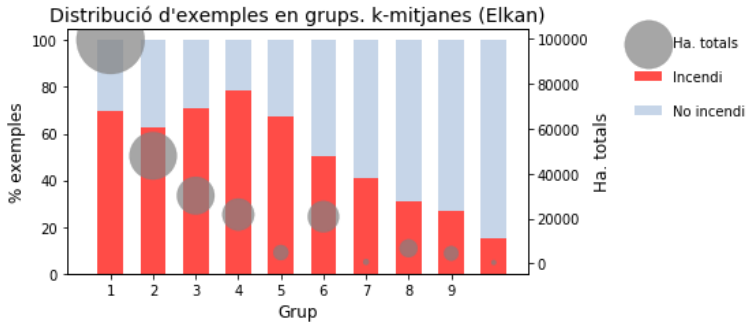
Imatge 27. Distribució dels exemples amb i sense incendi i superfície cremada en els grups obtinguts en els diversos mètodes.

Un cop analitzats els resultats s'ha optat pel mètode k-mitjanes ja que ha obtingut bones mètriques i al mateix temps distribueix els exemples en grups amb un clar predomini d'incendis o bé de no incendis i sense clústers massa petits.

Per a ordenar els grups s'ha aplicat un índex de risc on s'ha tingut en compte el percentatge d'exemples amb incendi de cada grup així com la superfície total cremada segons la següent fórmula:

$$\frac{\frac{n}{N} + \frac{\sum_{i=1}^{n} s}{\max\left(\sum_{i=1}^{n} s\right)}}{2}$$

on n és el nombre d'exemples d'incendi del grup, N el nombre total d'exemples del grup i s la superfície total cremada en cada exemple del clúster. El següent gràfic mostra els grups finals ordenats segons l'anterior índex i indicant el percentatge d'exemples d'incendi i no incendi en cada clúster.



Imatge 28. Percentatge d'exemples amb i sense incendi i superfície cremada en cada clúster (algorisme k-mitjanes).

La següent taula resumeix les principals característiques dels clústers obtinguts. Per una banda, les condicions meteorològiques mitjanes de cada clúster i, per l'altra, les característiques mitjanes dels exemples del grup: els casos amb incendi i sense incendis, la mitjana d'hectàrees forestals i no forestals cremades, l'orientació, pendent i altitud mitjana de l'indret d'ignició, el seu índex NDVI mitjà i la coberta del sòl predominant.

Clúster	1	2	3	4	5	6	7	8	9	10
Característiques meteorològiques										
Humitat relativa *(%, mitjana)*	34,4	48,3	47,6	55,7	42,2	57,5	64,8	77,5	70,2	78,0
Temperatura *(mitjana)*	28,1	16,3	24,1	26,2	17,5	19,0	15,2	22,9	10,2	14,1
Component *u* del vent *(mitjana)*	2,09	4,85	-2,00	0,49	0,16	0,31	0,83	-0,47	0,81	-2,07
Component *v* del vent *(mitjana)*	0,95	-2,22	2,42	3,22	0,39	-4,92	2,86	0,06	-0,92	-0,19
Característiques dels incendis										
Incendis	518	636	509	911	363	249	177	354	433	97
No incendis	227	374	210	251	177	245	254	797	1172	543
Forestal *(Ha, mitjana)*	145	62,6	44,2	20,8	12,6	71,5	3,9	17,5	10,1	4,6
No forestal *(Ha, mitjana)*	47,0	12,8	15,1	3,1	0,8	12,3	0,0	1,6	0,4	0,1
Orientació *(mitjana)*	181	179	178	178	177	175	171	172	165	175
Pendent *(mitjana)*	13	12	11	13	21	13	13	15	17	13
Altitud *(mitjana)*	499	313	322	279	982	250	356	326	747	306
NDVI *(mitjana)*	0,45	0,44	0,44	0,43	0,41	0,48	0,46	0,45	0,43	0,48
Coberta *(predominant)*	Matollars									

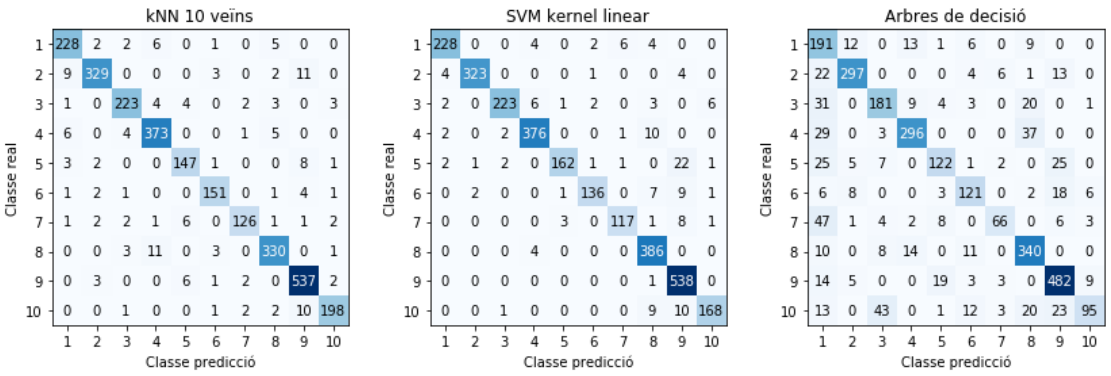
Taula 25. Principals característiques mitjanes dels clústers.

Aquestes dades permeten analitzar les característiques de cada grup. Per exemple, el grup 1 representa característiques meteorològiques de zones on es produeixen molts incendis, amb una superfície mitjana cremada elevada, en condicions de baixa humitat relativa, altes temperatures, vents mitjans de component est-nord-est en zones de matollar de poca pendent i altituds mitjanes. Per contra, el grup 2, tot i que també es correspon a zones i condicions meteorològiques amb un alt nombre d'incendis, en aquest cas les superfícies creades son menors, i es produeixen en condicions d'humitat relativa més alta, temperatures 12° més baixes, vents més forts i de component est-sud-est en cotes més baixes.

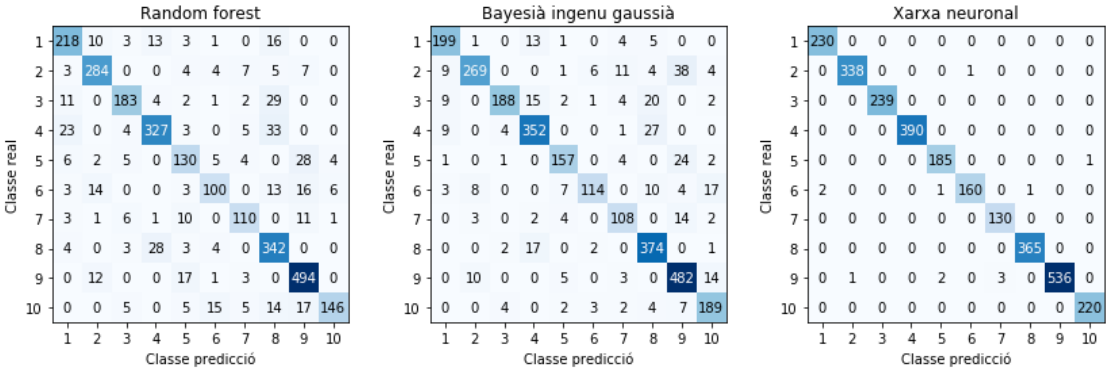
Un cop obtinguts els grups, s'han entrenat diversos classificadors per tal de classificar nous exemples en un dels clústers. Tot seguit es mostren els resultats obtinguts així com les matrius de confusió per a cadascun dels algorismes on es pot comprovar que els millors resultats s'assoleixen amb la xarxa neuronal. Així doncs, s'utilitzarà aquest mètode per a classificar els nous casos en un dels grups i, per tant, per a obtenir els mapes de risc.

Mètode	Configuració	Exactitud[33]	Precisió	Sensibilitat[34]	Mesura F1
kNN	Veïns=3	0,91	0,91	0,91	0,91
	Veïns=5	0,93	0,93	0,91	0,92
	Veïns=7	0,93	0,92	0,91	0,92
	Veïns=10	0,93	0,93	0,92	0,93
Màquines de vectors de suport	Kernel=lineal	0,94	0,96	0,93	0,95
	Kernel=*gaussià*	0,89	0,92	0,86	0,89
Arbres de decisió	Profunditat = 6	0,77	0,77	0,75	0,74
Random forest	Profunditat = 5	0,83	0,83	0,80	0,81
Bayesià ingenu gaussià		0,87	0,89	0,86	0,87
Xarxa neuronal	Capa oculta: 10 unitats	0,99	0,99	0,99	0,99

Taula 26. Mètriques dels diversos algorismes de classificació.



[33] En anglès, *accuracy*.
[34] En anglès, recall



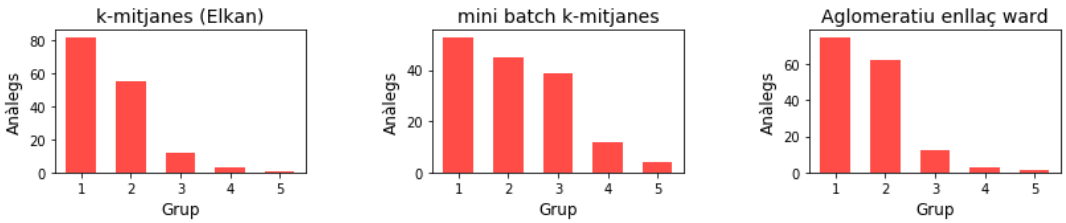
Imatge 29. Matrius de confusió.

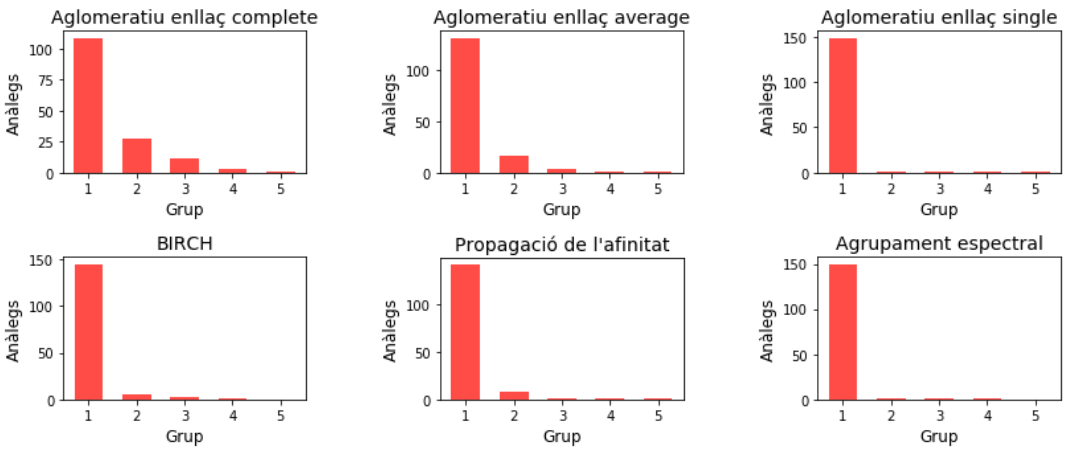
8.1.2 Basats en anàlegs

Tot seguit s'han avaluat els models d'agrupament basats en dades meteorològiques provinents dels anàlegs. Tot i que s'han obtingut bones mètriques per al mètode aglomeratiu en la seva versió d'unió mitjana i simple i també amb els mètodes BIRCH, propagació de l'afinitat i agrupament espectral, una anàlisi de la distribució dels anàlegs en cada clúster en desaconsellen l'ús ja que agrupen la major part de casos en un sol grup. De la resta de mètodes, k-mitjanes és el que ha assolit els millors resultats i, per tant, és el mètode utilitzat per a la categorització dels anàlegs. La següent taula i els gràfics de distribució d'exemples en clústers mostren aquests resultats.

Mètode	**Configuració**	**Coeficient de silueta**	**Índex Davies Bouldin**	**Temps** *(segons)*
k-mitjanes		0,519	0,662	0,06
k-mitjanes (mini lots)	*Mida dels lots=90*	0,522	0,708	0,05
Agrupament jeràrquic aglomeratiu	Mètode d'unió: guarda	0,483	0,730	0,002
	Mètode d'unió: complet	0,413	0,784	0,002
	Mètode d'unió: mitjana	0,699	0,494	0,002
	Mètode d'unió: simple	0,861	0,128	0,002
BIRCH	Factor de fuga=100, llindar=12	0,842	0,357	0,04
Propagació de l'afinitat	*Preference*=-70, *damping*=0.95	0,814	0,321	0,03
Agrupament espectral	Matrius d'afinitat: *kernel* gaussià	0,821	0,160	0,65

Taula 27. Mètriques dels mètodes d'agrupament per als anàlegs.



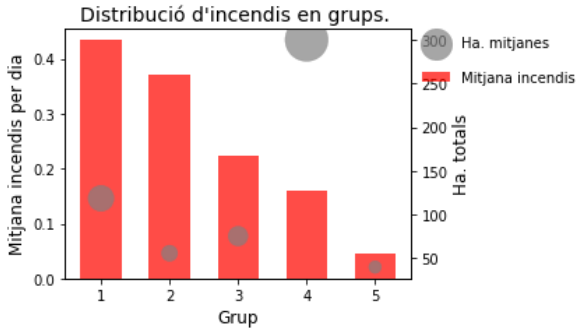


Imatge 30. Distribució dels anàlegs en els diversos clústers.

Per ordenar els grups en funció d'un índex de risc no s'ha emprat la fórmula proposada en l'apartat anterior per als models basats en reanàlisi degut a les diferències en els dos conjunts de dades meteorològiques. La nova fórmula utilitzada té en compte tant la mitjana d'incendis per dia com les hectàrees mitjanes cremades en els incendis produïts en els dies que formen cadascun dels clústers:

$$\frac{\frac{n}{N} + \frac{\bar{s}}{\max(\bar{s})}}{2}$$

on n és el nombre d'incendis produïts en els dies que formen el clúster, N el nombre total de dies del grup i $\bar{s}$ la superfície mitjana cremada. El següent gràfic mostra els grups finals ordenats segons aquest índex on s'indica la mitjana d'incendis per dia i la mitjana d'hectàrees cremades per a cada clúster.



Imatge 31. Mitjana d'incendis per dia i mitjana d'hectàrees cremades de cada clúster.

Tot seguit s'han analitzat tant les mitjanes de les principals característiques de cadascun dels cinc grups d'anàlegs com les característiques dels incendis produïts en els dies que formen cadascun dels clústers.

Clúster		1	2	3	4	5
Característiques dels anàlegs						
Núm. anàlegs		55	82	12	1	3
Vent		0,014	0,0147	0,0184	0,0433	0,0048
Component *u* del vent		6,98	6,6919	7,0445	6,2009	6,8667
Component *v* del vent		7,813	8,0942	7,925	6,1852	6,7222
Humitat relativa	*500 hPa*[35]	-0,827	-1,071	-1,946	-2,035	-1,522
	850 hPa[36]	-0,95	-1,079	-1,181	1,579	-0,008
Temperatura		10,825	14,372	11,54	19,881	18,511
Precipitació		27,688	36,532	55,409	145,78	93,526
Característiques dels incendis						
Incendis		855	906	181	11	9
Incendis per dia *(mitjana)*		0,43	0,37	0,22	0,16	0,05
Forestal *(Ha, mitjana)*		88,3	45,1	66,8	273,9	5,7
No forestal *(Ha, mitjana)*		30,2	10,5	8,5	26,5	34,1
Mes *(predominant)*		7	7	7	9	6
Hora inici *(mitjana)*		14:41	14:56	14:41	13:32	13:00
Humitat relativa *(%, mitjana)*	*1000 hPa*[37]	48,9	50,3	43,9	49,2	37,8
Temperatura *(mitjana)*		299,1	298,9	299,7	298,6	302,3
Component *u* del vent *(mitjana)*		0,44	0,13	0,51	-0,38	-0,23
Component *v* del vent *(mitjana)*		1,24	1,07	2,16	2,00	2,83
Orientació *(mitjana)*		172,8	179,3	181,9	168,4	235,5
Pendent *(mitjana)*		13,5	13,5	13,3	17,1	17,6
Altitud *(mitjana)*		325,0	343,7	313,5	372,3	410,7
NDVI *(mitjana)*		0,44	0,44	0,45	0,47	0,42
Coberta *(predominant)*		Matollar	Matollar	Matollar	Boscos aciculifolis	Conreus

Imatge 32. Principals característiques mitjanes dels anàlegs de cada clúster.

Les anteriors dades permeten identificar dies per la pertinença a un dels 5 grups. En el cas del grup 1, per exemple, aquest està format per dies amb vents febres i temperatures i precipitacions baixes, on s'ha produït una mitjana de 0,43 incendis per dia, amb 88,3 ha. forestals cremades de mitjana, predominantment el mes de juliol en zones de matollars de poca pendent i altituds baixes. Per contra, en el cas del grup 4, format per dies amb vents més forts i temperatures i precipitacions més altes, s'ha produït una mitjana de tan sols 0,16 incendis per dia tot i que amb una superfície cremada mitjana de 273,9 ha., predominantment el mes de setembre, en zones de bosc d'aciculifolis amb més pendent i més altitud.

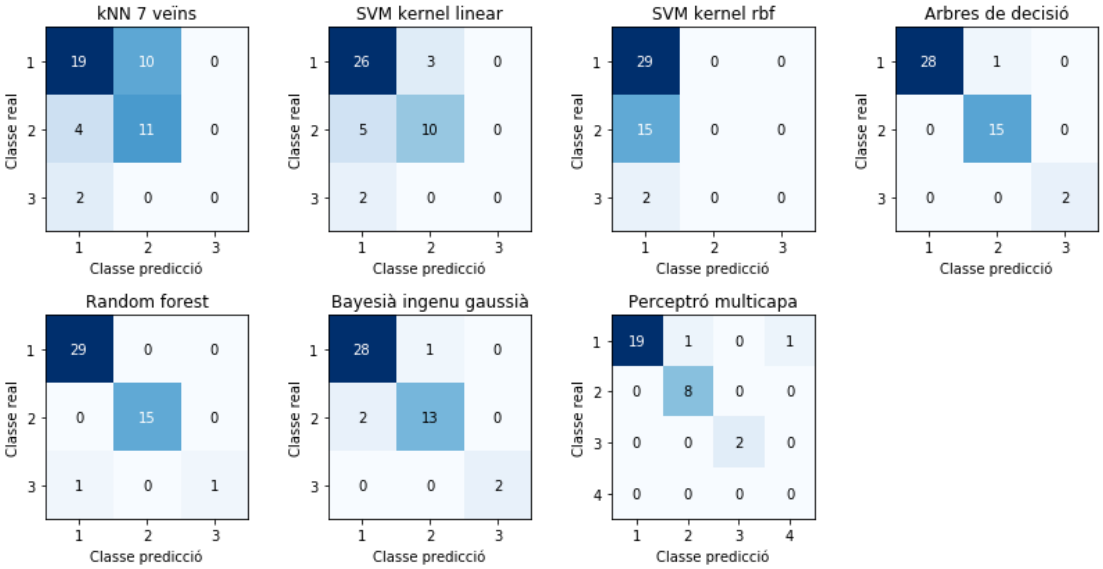
Un cop generats els clústers aquests s'han utilitzat per a classificar nous anàlegs entrenant diversos classificadors per avaluar-ne el comportament. La taula següent en mostra els resultats.

[35] Nivell de pressió atmosfèrica 500 hPa.
[36] Nivell de pressió atmosfèrica 850 hPa.
[37] Nivell de pressió atmosfèrica 1000 hPa.

Mètode	Configuració	Exactitud	Precisió	Sensibilitat	Mesura F1
KNN	Veïns=2	0,63	0,63	0,63	0,63
	Veïns=3	0,63	0,63	0,63	0,63
	Veïns=5	0,61	0,61	0,61	0,61
	Veïns=7	0,65	0,65	0,65	0,65
	Veïns=9	0,65	0,65	0,65	0,65
Màquines de vectors de suport	Kernel=lineal	0,78	0,78	0,78	0,78
	Kernel=*gaussià*	0,63	0,63	0,63	0,63
Arbres de decisió	Profunditat = 6	0,98	0,98	0,98	0,98
Random forest	Profunditat = 5	0,98	0,98	0,98	0,98
Bayesià ingenu gaussià		0,93	0,93	0,93	0,93
Xarxa neuronal	Capa oculta: 10 unitats	0,94	0,94	0,94	0,94

Taula 28. Mètriques dels algorismes de classificació.

En les següents matrius de confusió es pot observar que no hi ha prou exemples en els clústers 4 i 5 per a la validació creuada. Així és que només es mostra els resultat per als tres o quatre primers grups en funció del classificador. Per altra banda, els millors resultats s'han assolit amb els arbres de decisió i el mètode *Random forest* i els pitjors amb les màquines de vectors de suport amb *kernel* gaussià, on tots els exemples del conjunt de test han estat classificats en el mateix clúster.



Imatge 33. Matrius de confusió dels classificadors.

Tot i que ha estat possible entrenar classificadors robustos, els resultats obtinguts aconsellen l'ús de més dades d'anàlegs per a poder obtenir clústers més ben balancejats. Per tant, dels dos conjunts de dades meteorològiques analitzats s'opta pels reanàlisis ERA5 del Climate Data Store. Tot i així, els resultats assolits amb els anàlegs provinents del Servei Meteorològic de Catalunya en recomanen un futur anàlisi amb un conjunt de mida superior que cobreixin tant un període de temps major com tots els mesos de l'any.

8.2 Avaluació dels algorismes de regressió per a la predicció de la mida dels incendis forestals en el moment d'ignició

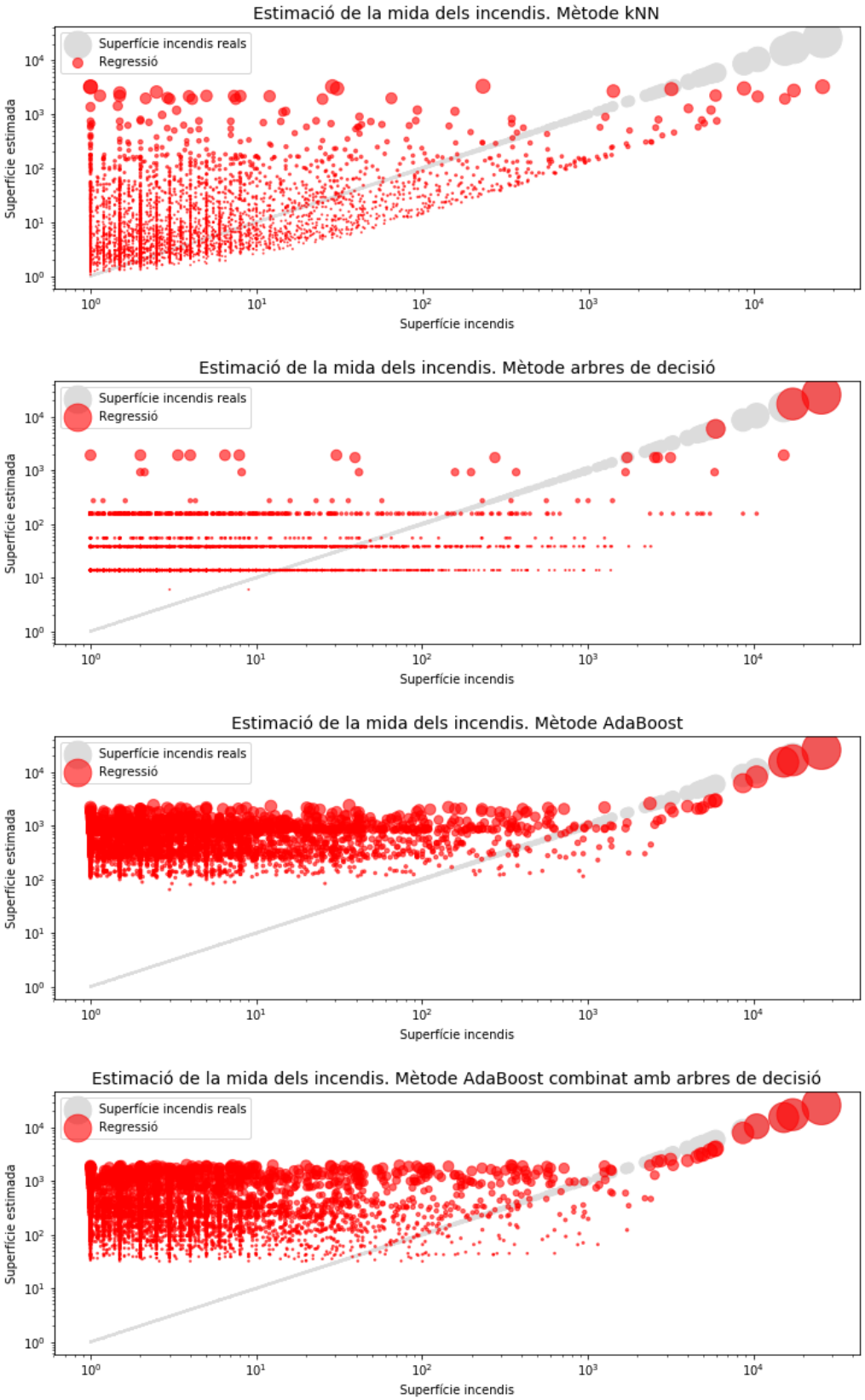
Per a avaluar els algorismes de regressió s'han utilitzat les següents mètriques: variància explicada, error absolut mitjà, error mitjà quadrat i coeficient de determinació. La variància explicada és una estimació de la bondat de l'ajust; quan més alt és aquest valor, que pot variar entre 0 i 1, millor és el grau d'ajust. Pel que fa a l'error absolut mitjà aquest mesura la distància mitjana entre cada punt de dues variables contínues. Per contra, l'error mitjà quadrat calcula la distància quadrada mitjana entre els valors estimats i els reals. Finalment, el coeficient de determinació R^2, la darrera mètrica utilitzada, mesura la bondat de l'ajust on els millors resultats son propers a 1.

Un cop analitzades aquestes mètriques resumides en la següent taula per als diversos algorismes emprats s'observa que el conjunt de dades disponibles no permeten predir la mida final dels incendis ja que mostren una predicció molt pobra.

Mètode	Configuració	Variància explicada	Error absolut mitjà (MAE)	Error mitjà quadrat (MSE)	Coeficient de determinació (R^2)
kNN	Veïns: 8	0,244	77,61	296.126,42	0,244
Màquines de vectors de suport	*Kernel*: lineal	0,0003	54,20	394.402,61	-0,007
	Kernel: gaussià	0,0002	54,20	394.485,14	-0,007
Arbres de decisió	Profunditat: 4	0,745	61,28	99.979,25	0,745
AdaBoost		0,511	617,53	541.695,68	-0,383
Arbre de decisió + AdaBoost		0,240	483,38	509.584,97	-0,301
Regressió logística		0,0001	54,65	394.708,00	-0,008
Kernel ridge	Alfa=1	0,017	109,65	385.238,19	0,017
Xarxes neuronals	Funció activació: sigmoide	0,012	87,54	387.239,39	0,011

Taula 29. Mètriques dels diversos algorismes de regressió.

Els següents gràfics mostren l'estimació de la mida dels incendis obtinguda tot comparant-la amb la mida real dels incendis per a quatre dels mètodes emprats. A causa de la baixa capacitat de predicció dels algorismes no s'ha utilitzat el conjunt de test sinó el propi conjunt d'entremanent per visualitzar el comportament dels models. Els algorismes que mostren cert aprenentatge son, per una banda kNN, amb una lleugera tendència en la predicció de la mida dels incendis tot i predir erròniament com a grans incendis part dels incendis petits, i, per l'altra, els arbres de decisió, AdaBoost i la combinació d'ambdós que han aconseguit la millor estimació de la mida dels incendis més grans.



Imatge 34. Comparació de la mida real i la mida estimada dels incendis.

Els resultats obtinguts aconsellen utilitzar un conjunt de característiques diferents per a l'estimació de la mida dels incendis en el moment d'ignició, concretament amb més atributs meteorològics.

8.3 Avaluació dels algorismes de classificació per a l'estimació del risc d'incendis forestals

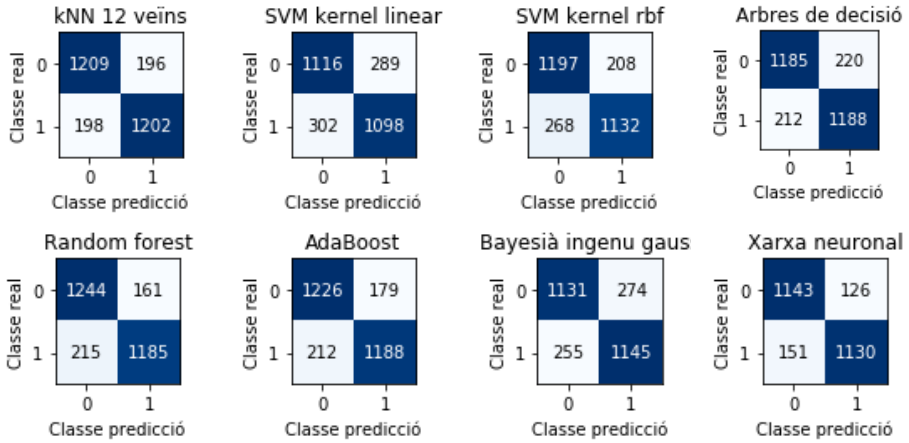
La següent taula mostra els resultats de diverses mètriques utilitzades per avaluar els algorismes supervisats de classificació implementats: kNN, màquines de vectors de suport, arbres de decisió, *Random forest*, AdaBoost, bayesià ingenu gaussià i xarxes neuronals. Per a obtenir aquestes mètriques i poder-les comparar amb la resta d'experiments duts a terme i tenint en compte que el conjunt d'entrenament no disposa d'una gran quantitat d'exemples, s'ha utilitzat validació creuada.

S'ha comprovat en els diversos experiments duts a terme que afegir la similitud de les condicions meteorològiques obtinguda prèviament augmenta entre un 3% i un 5% l'exactitud assolida per a la validació creuada.

Mètode		**Entrena-ment**	**Validació creuada**	**Test**			
	Configuració	Exactitud	Exactitud	Exactitud	Precisió	Sensibilitat	F1
kNN	Veïns=2	0,931	0,837	0,837	0,837	0,759	0,823
	Veïns=6	0,900	0,857	0,857	0,857	0,833	0,853
	Veïns=10	0,888	0,859	0,858	0,858	0,851	0,857
	Veïns=12	0,885	0,858	0,859	0,859	0,859	0,859
	Veïns=14	0,880	0,863	0,857	0,857	0,857	0,857
SVM	*Kernel* lineal	0,789	0,784	0,789	0,792	0,784	0,788
	Kernel gaussià	0,847	0,836	0,830	0,845	0,809	0,826
Arbres decisió	Profunditat:6	0,873	0,828	0,846	0,843	0,849	0,846
Random forest	Profunditat:5	0,888	0,869	0,866	0,880	0,846	0,863
AdaBoost		0,867	0,857	0,861	0,869	0,849	0,859
Bayesià ingenu gaussià		0,812	0,808	0,811	0,807	0,818	0,812
Xarxa neuronal	1 capa, 10 unitats	0,898	0,876	0,891	0,900	0,882	0,891

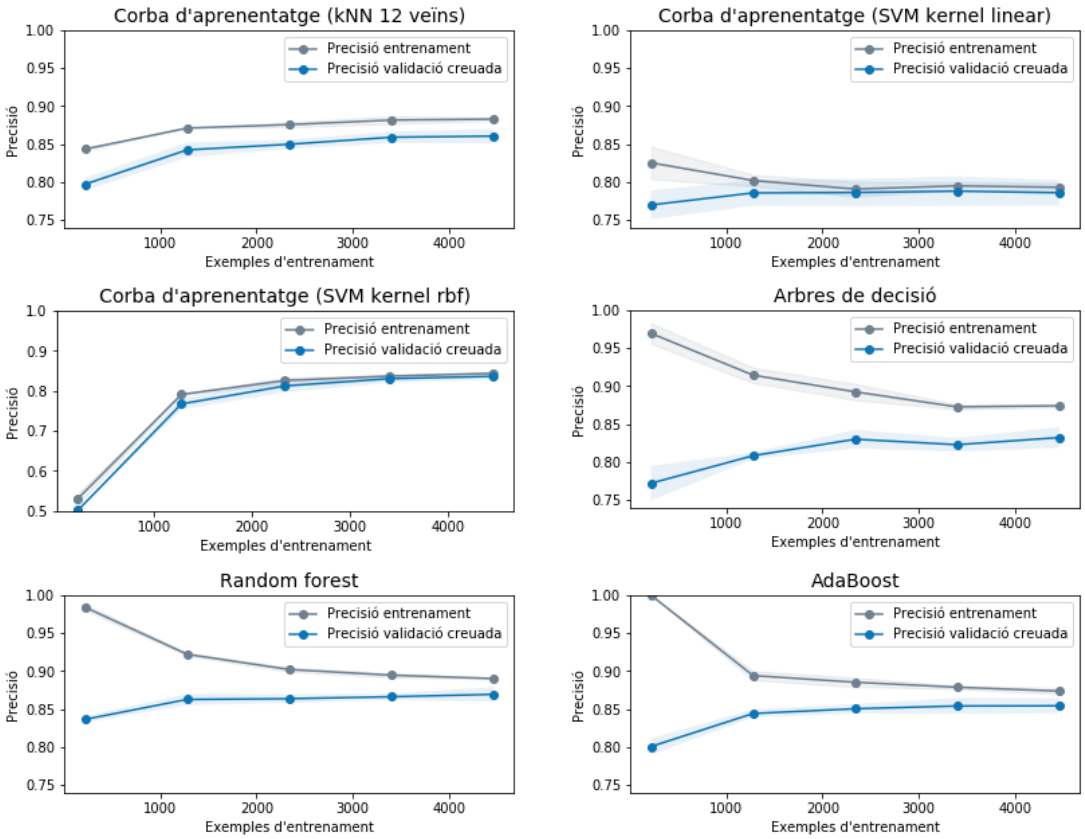
Taula 30. Mètriques dels diversos models de classificació.

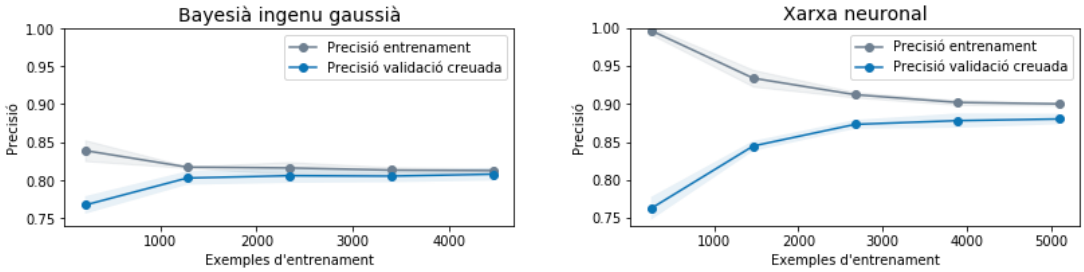
La següent imatge mostra les matrius de confusió dels divisos models. Per al cas del mètode kNN tan sols es mostra la matriu de confusió per a la configuració amb 12 veïns. Tots els algorismes de classificació utilitzats han obtingut exactituds en el conjunt de test iguals o superiors al 79%. Concretament, els algorismes que han superat el 85% han estat: kNN, *Random forest*, AdaBoost i la xarxa neuronal. D'aquests, els que han obtingut unes millors mètriques han estat l'algorisme *Random forest* i la xarxa neuronal.



Imatge 35. Matrius de confusió (classe 0: no incendis, classe 1: incendis).

Mes enllà de les mètriques anteriors, s'han analitzat les corbes d'aprenentatge dels models per tal de detectar possibles problemes de biaix o alta variància. Els següents gràfics en mostren els resultats.





Imatge 36. Corbes d'aprenentatge.

En el model kNN es manté la distància entre la corba d'aprenentatge dels conjunts d'entrenament i de validació indicant un possible problema de variància tot i que poc accentuat.

Per contra, en les màquines de vectors de suport tant amb *kernel* lineal com gaussià així com en el cas del mètode bayesià ingenu gaussià la variància és clarament baixa tan perquè la distància entre les dues cobres és molt petita com pel fet que l'error del conjunt d'entrenament es manté elevat. Per altra banda, l'elevat error d'entrenament indica un biaix elevat i, per tant, aquesta algorismes presenten un problema d'infraajust (en anglès, *underfitting*). Per a solucionar aquest problema no és útil augmentar el nombre d'exemples d'entrenament sinó que caldrà augmentar la complexitat del model afegint nous atributs. En aquest cas, es podrien recollir més atributs com, per exemple, atributs de vegetació, combustible, sequera, precipitacions, o bé utilitzar atributs polinòmics ja que d'aquesta forma es podria reduir el biaix del conjunt d'entrenament. Una segona solució és disminuir el paràmetre de regularització utilitzat ja que disminuint-lo el model s'ajustarà millor a les dades d'entrenament.

La resta de models assoleixen millors corbes d'aprenentatge. D'aquestes, les corresponents als arbres de decisió son les que presenten una major variància al mateix temps que un biaix inferior al dels casos anteriors. Tant l'alta variància com un error d'entrenament baix indiquen l'existència d'un problema de sobreentrenament (en anglès, *overfitting*). Aquest model pot ser millorat afegint més exemples tant d'incendis com de no incendis al conjunt d'entrenament o bé a partir de nous exemples recents o bé localitzant incendis actualment no registrats mitjançant imatges satèl·lit. També és possible reduir el nombre d'atributs utilitzats per a reduir la variància tot i que cal tenir en compte que pot augmentar el biaix.

Finalment, el models obtinguts amb els mètodes *Random forest*, AdaBoost i la xarxa neuronal presenten variància tot i que la tendència a convergir de les dues corbes d'aprenentatge i el baix error indiquen que augmentant el nombre d'exemples aquestes convergiran i, molt probablement, milloraria el model. En aquest cas, doncs, encara resta marge per a la millora del model a partir de l'ús de més exemples d'entrenament.

8.3.1 Optimització dels models de classificació

Un cop avaluats els models de classificació emprats, s'han analitzat diverses tècniques per tal d'optimitzar-ne el rendiment. Les tres aproximacions utilitzades han estat: la combinació de mètodes, l'avaluació de l'ús de *baggin*, *random subsapce methods* o bé *random patches* i, per acabar, l'ús d'algorismes genètics per a l'optimització d'hiperparàmetres. Els bons resultats obtinguts amb la xarxa neuronal converteixen aquest mètode en un bon candidat per a la darrera aproximació ja que permetrà optimitzar-lo i, així, intentar obtenir millors resultats en la classificació.

8.3.1.1 Combinació de mètodes

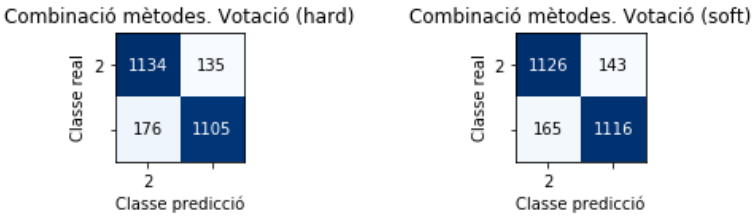
La combinació de mètodes (en anglès, *ensemble methods*) consisteix en l'obtenció de diversos models de classificació i la posterior combinació d'aquests per a millorar els seus resultats individuals. Esdevé útil tant per a obtenir classificadors més robustos com per a millorar la capacitat de generalització, ja que uns algorismes soluciones les dificultats que poden presentar uns altres. No obstant, acostuma a oferir millors precisions que els models individuals sempre i quan aquests no siguin massa similars ja que, en aquest cas, els seus errors no es compensaran amb la combinació sinó que se sumen.

S'han utilitzat set dels models obtinguts anteriorment: kNN amb 12 veïns, màquines de vectors de suport amb *kernel* gaussià, arbres de decisió, *Random forest*, AdaBoost, bayesià ingenu gaussià i xarxes neuronals amb una sola capa oculta de 10 unitats. El mètode emprat ha estat el de votació (en anglès, *voting*), consistent en obtenir la predicció de cadascun dels models inicials i assignar com a predicció final aquella que assoleix més vots. N'existeixen dues variant: majoria (en anglès, *majority* o *hard voting*) on es classifica d'acord amb l'etiqueta de la classe majoritària i probabilitats mitjanes ponderades (en anglès, *weighted average probabilites* o *soft voting*) on s'obté la classe a partir de la mitjana de les probabilitats calculades per a cada model inicial.

S'han obtingut resultats similars en ambdues variants del mètode augmentant en un 0,7% l'exactitud de la validació creuada del millor dels models inicials i un 3,6% pel que fa a l'exactitud mitjana dels models inicials. Per contra, l'exactitud obtinguda amb el conjunt de test és un 2,7% superior a la mitjana dels algorismes inicials i també superior a cadascun d'aquests excepte en el cas de la xarxa neuronal, que supera a l'exactitud de la combinació de mètodes en un 1,02%. La següent taula i les matrius de confusió que l'acompanyen mostren aquests resultats.

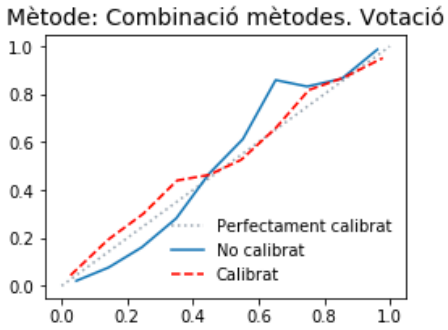
Mètode	Entrenament	Validació creuada	Test			
	Exactitud	Exactitud	Exactitud	Precisió	Sensibilitat	F1
Majoria	0,901	0,882	0,878	0,891	0,863	0,876
Prob. mitjana ponderada	0,900	0,883	0,879	0,886	0,871	0,879

Taula 31. Mètriques de la combinació de mètodes.



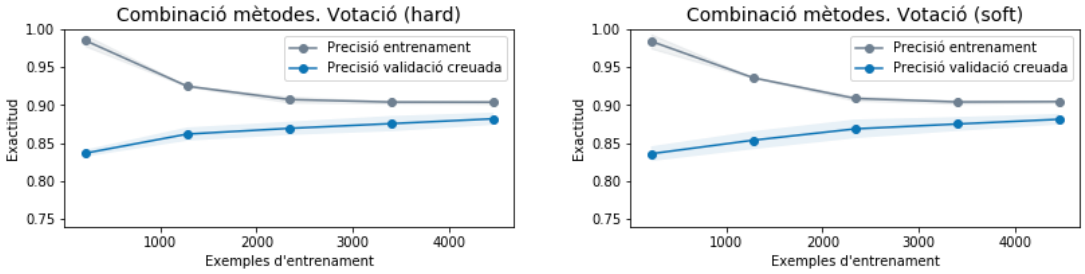
Imatge 37. Matrius de confusió.

Per altra banda, l'anàlisi del gràfic de calibració de la probabilitat del mètode de probabilitats mitjanes ponderades aconsella l'ús de la calibració. Per contra, aquesta no pot ser utilitzada per a la variant majoria del mètode ja que aquesta no utilitza probabilitats.



Imatge 38. Diagrama de fiabilitat de la combinació de mètodes (probabilitat mitjana ponderada).

Finalment, les corbes d'aprenentatge mostren un bon comportament de l'algorisme on no s'aprecia ni biaix ni sobreentrenament en cap de les dues variants. Tot i això, com ja s'havia observat en alguns dels mètodes inicials, és possible millorar el model ampliant el nombre d'exemples d'entrenament. S'utilitzarà doncs la variant *soft* del mètode per a la generació de mapes de predicció ja que aquesta permet l'obtenció de probabilitats.



Imatge 39. Corbes d'aprenentatge de les dues variants de la combinació de mètodes.

8.3.1.2 Bagging, random subsapce methods, random patches

El mètode *bagging* consisteix en dividir el conjunt d'exemples en subconjunts de casos i entrenar un classificador amb cadascun d'aquests. Com en el cas de la combinació de mètodes, un cop entrenats els diversos classificadors aquests poden ser utilitzats per a la classificació de noves instàncies escollint la classe que rep més vots o bé a partir de la mitjana de les probabilitats calculades per a cada model.

S'ha utilitzat aquest mètode ja que pot ser útil en el cas que el conjunt inicial contingui casos mal etiquetats, és a dir, pertanyents a una classe errònia. En aquest cas, aquesta mostra errònia només afectarà a un conjunt dels classificadors i no n'afectarà d'aquests. Cal tenir en compte que en generar el conjunt d'exemples negatius, tot i haver utilitzat mesures de similitud per eliminar el 20% dels casos més propers al conjunt d'incendis, existeix la possibilitat que una part dels casos de no incendi tinguin atributs més propis d'incendis. És per aquest motiu que aplicar *bagging* pot ajudar a aïllar aquests exemples en la classificació.

Per altra banda, l'algorisme *random subsapce methods* (RSM) divideix el conjunt de dades en subconjunts d'atributs eliminant de forma aleatòria alguns d'aquests atributs. Finalment, l'algorisme *random patches* divideix el conjunt original tant en l'espai d'atributs com de mostres.

La següent taula mostra els resultats obtinguts on s'aprecia una disminució en l'exactidud excepte en el cas dels arbres de decisió. Aquesta disminució es dona en major mesura en el cas de disminuir l'espai dels atributs com succeeix amb els mètodes RSM i *Random patches*, atès que la dimensionalitat del conjunt de mostres és baixa i no es dona alta redundància d'atributs.

En relació als resultats dels arbres de decisió, cal tenir en compte que l'algorisme *Random forest* es basa en l'ús d'arbres de decisió i el mètode *bagging*; això explica que s'hagin obtingut resultats similars amb el mètode *Random forest* que utilitzant arbres de decisió i *bagging*.

Mètode	**Configuració**	**Exactitud** (validació creuada)			
		Original	*Bagging*	RSM	*Random patches*
kNN	Veïns=10	0,858	0,841	0,840	0,839
SVM	*Kernel gaussià*	0,836	0,815	0,805	0,809
Arbres de decisió	Profunditat màxima 6	0,828	0,860	0,863	0, 863
Bayesià ingenu gaussià		0,808	0,789	0,786	0,792
Xarxa neuronal	1 capa, 10 unitats	0,876	0,848	0,851	0,838

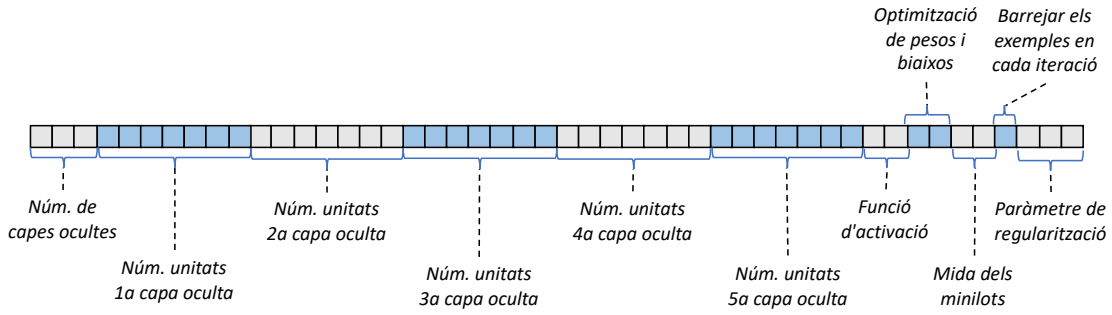
Taula 32. Exactitud dels classificadors.

8.3.1.3 Optimització del perceptró multicapa mitjançant algorismes genètics.

Un dels millors classificadors obtinguts ha estat la xarxa neuronal. Això ens duu a cercar una millor arquitectura d'aquesta utilitzant un dels mètodes principals d'optimització, els algorismes genètics. Per tant, s'utilitzarà aprenentatge profund, concretament el perceptró multicapa i, per altra banda, algorismes genètics que permetin optimitzar-ne l'arquitectura.

Amb aquesta finalitat s'ha implementat una algorisme genètic on s'ha emprat el mètode de creuament d'individus en dos punts amb una probabilitat de creuament d'entre un 10 i un 15% i mutació consistent en la inversió dels valors de determinats atributs de l'individu i amb una probabilitat de mutació dels atributs d'entre un 25 i un 30%. Aquesta mutació ha permès explorar noves zones de l'espai de solucions. Per altra banda, s'han realitzat diversos experiments d'entre 12 i 70 generacions i amb poblacions d'entre 50 i 100 individus.

Cada individu de la població representa una configuració d'un perceptró multicapa definit per un conjunt d'atributs, al seu torn formats per un conjunt de gens. La següent imatge mostra els gens de cada individu.



Taula 33. Gens d'un individu de la població.

S'ha utilitzat una funció decoradora per afegir les següents restriccions a la solució per tal d'afitar l'espai de solucions a aquelles acceptables, com es mostra en la següent taula.

Atribut	Valors acceptables
Núm. de capes ocultes	2 a 5
Núm. Unitats capa oculta 1	2 a 127
Núm. Unitats capa oculta 2	2 a 127
Núm. Unitats capa oculta 3	2 a 127
Núm. Unitats capa oculta 4	2 a 127
Núm. Unitats capa oculta 5	2 a 127
Funció d'activació	lineal, tangent hiperbòlica, ReLU[38]
Optimització de pesos i biaixos	mètode quasi-Newton, descens de gradient estocàstic
Mida dels minilots	32, 64, 128, tots els exemples
Barrejar els elements en cada iteració	sí, no
Paràmetre de regularització	0 a 0,1 (escala logarítmica)

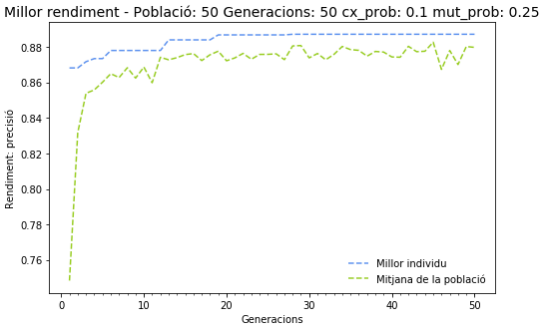
Taula 34. Espai de solucions.

La següent taula mostra els resultats obtingut en alguns dels experiments duts a terme on s'ha assolit un augment de l'exactitud en relació a la xarxa neuronal inicial. Les corbes que acompanyen la taula resumeixen el procés d'optimització per a cadascun del experiments.

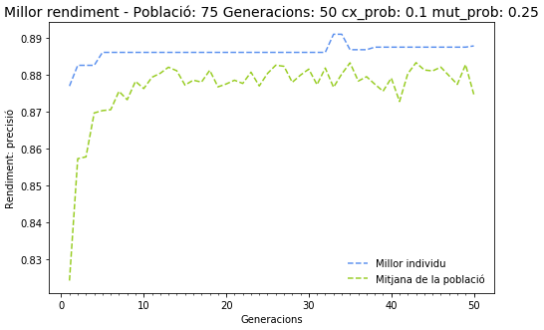
Experiment	Capes ocultes	Conjunt	Població	Generacions	Probabilitat creuament	Probabilitat mutació	Temps entrenament	**Exactitud** (validació creuada)		
								Xarxa neuronal original	Perceptró multicapa	Millora
1	1 a 5	50%	50	50	10%	25%	22h58'15''	0,876	0,887	+0,011
2	1 a 5	50%	75	50	10%	30%	21h33'36''		0,891	+0,015
3	1 a 5	50%	70	50	10%	30%	19h35'31''		0,882	+0,006
4	1 a 5	50%	100	12	10%	30%	8h15'		0,890	+0,014
5	1 a 5	100%	70	70	15%	25%	32h09'05''		0,896	+0,020

Taula 35. Resultats de la optimització del perceptró multicapa.

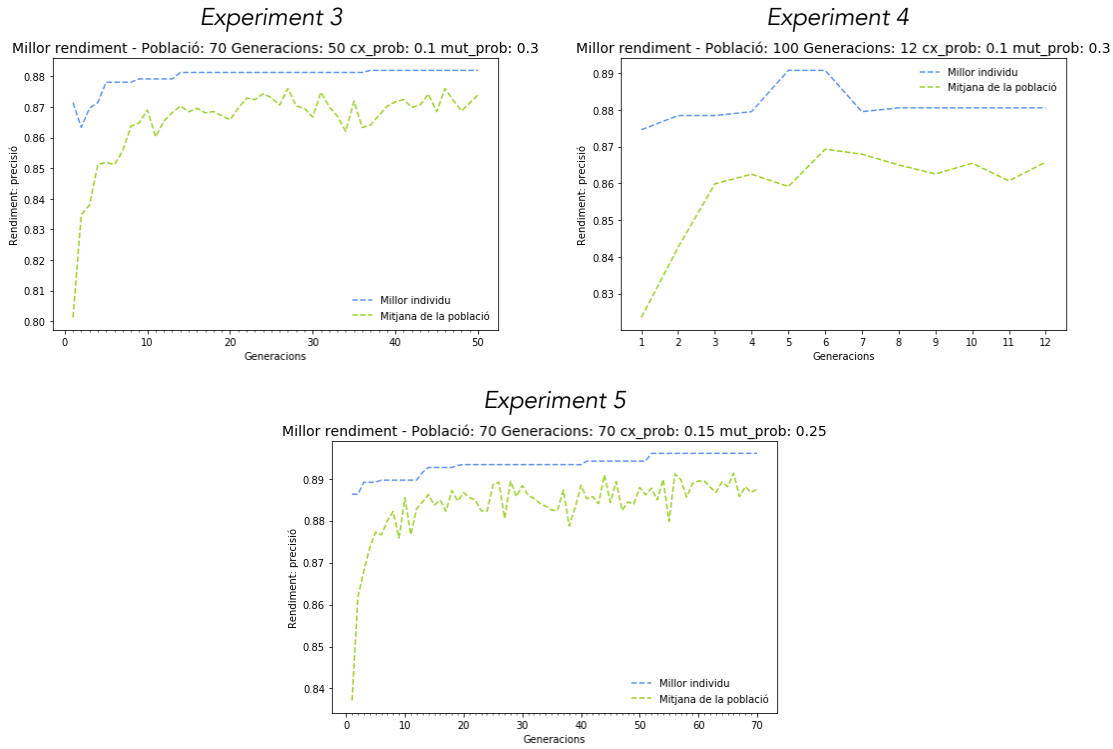
Experiment 1



Experiment 2

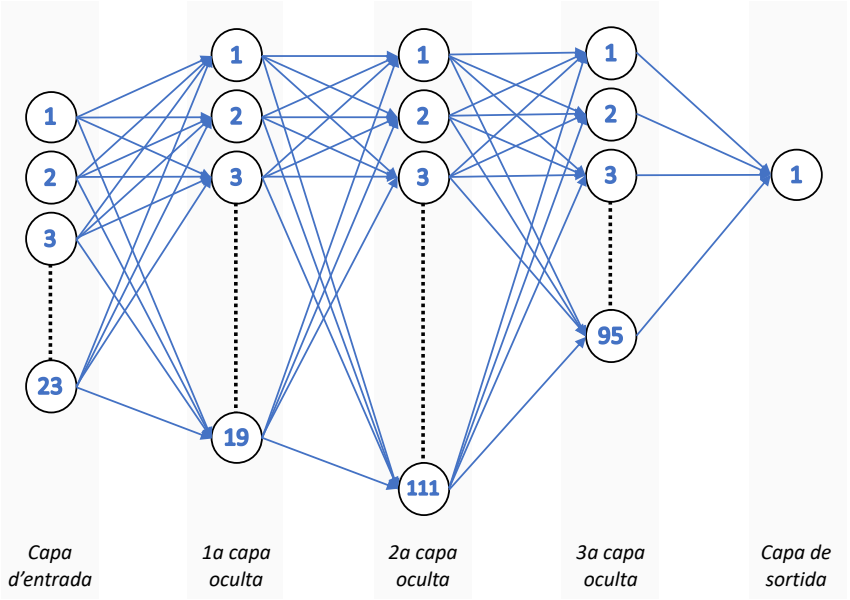


[38] Rectified Linear Unit.



Imatge 40. Corbes d'optimització.

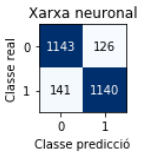
Tot i que no es pugui assegurar que l'algorisme genètic hagi assolit un òptim global, la millor solució és l'obtinguda en l'experiment 5 amb un augment de la precisió de la validació creuada del perceptró multicapa respecte a la xarxa neuronal inicial de dos punts. En aquest experiment s'ha utilitzat el 100% del conjunt d'exemples, una població de 70 individus, 70 generacions, una probabilitat de creuament del 15% i una probabilitat de mutació del 25%. Tot seguit es mostra l'estructura del perceptró obtingut format per tres capes ocultes de 19, 111 i 95 unitats.



Imatge 41. Estructura final del perceptró multicapa.

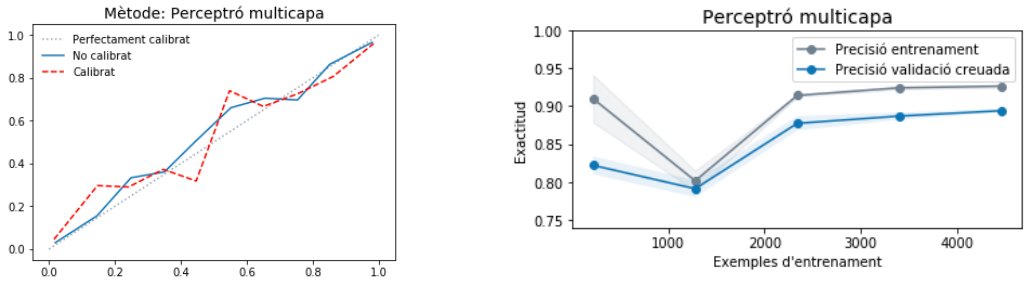
Amb l'optimització del perceptró multicapa mitjançant algorismes genètics s'han assolit els millors resultats amb un 94,6% d'exactitud en l'entrenament, un 89,6% en la validació creuada i un 89,5% en el conjunt de test.

Entrenament	Validació creuada	Test			
Exactitud	Exactitud	Exactitud	Precisió	Sensibilitat	F1
0,946	0,896	0,895	0,900	0,900	0,895



Taula 36. Mètriques del perceptró multicapa optimitzat genèticament. Matriu de confusió.

L'anàlisi del gràfic de calibració indica que no és necessària la calibració de les probabilitats per a l'estimació del risc d'incendi. Per altra banda, el gràfic d'aprenentatge mostra un baix biaix i una alta variància ja que la distància entre les dues corbes és significativa. Això és degut a un problema de sobreaprenentatge del model.



Imatge 42. Gràfic de calibració. Corba d'aprenentatge.

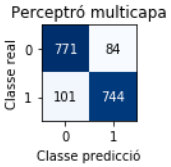
Per a reduir la variància i, per tant, el sobreaprenentatge es pot optar per tres estratègies: augmentar la regularització, augmentar el nombre d'exemples d'entrenament o bé reduir el nombre d'atributs per tal de reduir la complexitat del model i, per tant, la variància tot i que augmenti el biaix.

S'ha realitzat un conjunt d'experiments al voltant de les tres estratègies anteriors. En els primers s'ha augmentat la mida del conjunt d'entrenament utilitzant el 80% del conjunt global en comptes del 70%. Això ha comportat la disminució del conjunt de test que ha passat del 30 al 20% dels exemples. Els resultat obtinguts han reduït la variància.

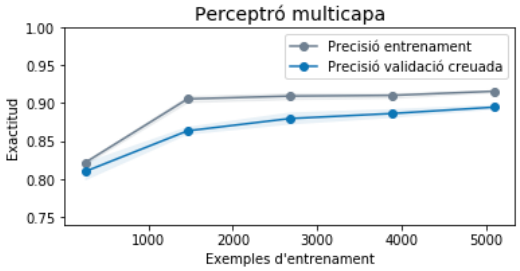
Pel que fa a al segona estratègia, l'augment del paràmetre de regularització del perceptró multicapa ha disminuït la variància. El millor equilibri entre variància i precisió del model s'ha obtingut amb un valor de 0,4 per al paràmetre de regularització. Finalment, s'ha reduït la dimensionalitat passant d'una variabilitat mínima del 96,13% al 95%. Aquests experiments no han assolit bons resultats ja que en tots s'ha mantingut la variància inicial.

Així doncs, s'ha optat per mantenir els mateixos atributs, augmentar el paràmetre de regularització així com la mida del conjunt d'entrenament. Tot seguit es mostren els principals estadístics de la nova configuració del perceptró multicapa i la nova corba d'aprenentatge on s'aprecia que tot i una lleugera disminució de l'exactitud del model s'ha reduït la variància.

Entrenament	Validació creuada	Test			
Exactitud	Exactitud	Exactitud	Precisió	Sensibilitat	F1
0,917	0,895	0,891	0,899	0,880	0,889



Taula 37. Principals mètriques del perceptró multicapa un cop reduïda la variància. Matriu de confusió.

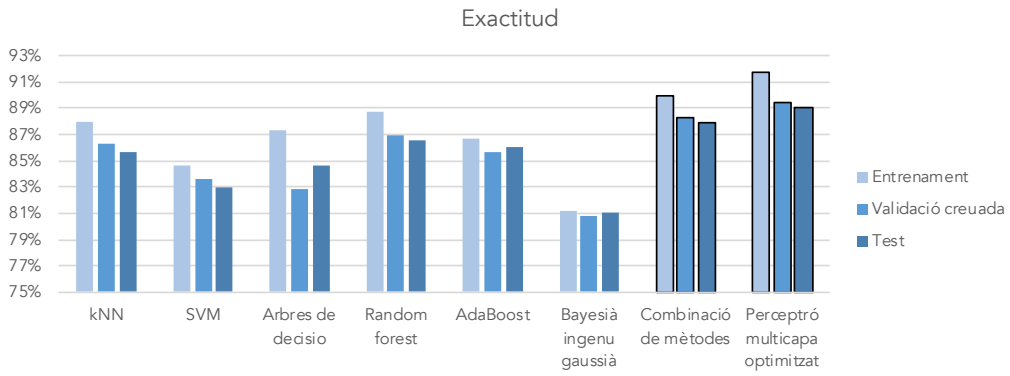


Imatge 43. Corba d'aprenentatge del perceptró multicapa un cop reduïda la variància.

Una de les conclusions d'aquests experiments és que disposar de més exemples d'entrenament milloraria els resultats del perceptró multicapa ja que molt probablement en reduiria encara més la variància.

8.3.2 Comparació dels models finals

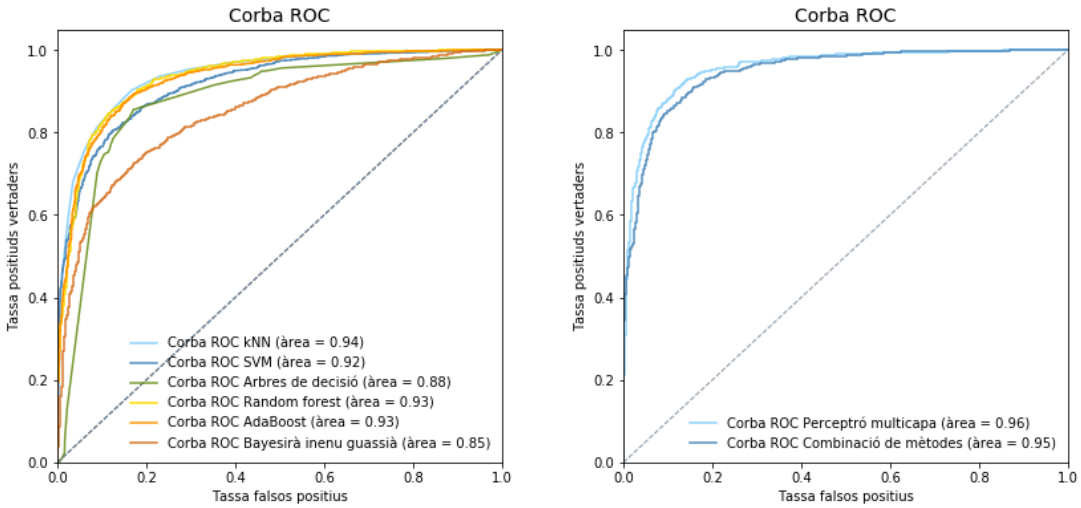
El següent gràfic mostra de forma resumida l'exactitud de l'entrenament, la validació i el test dels models de classificació obtinguts on es pot apreciar com les optimitzacions dutes ha terme han millorat de forma significativa els resultats, sobretot en el cas de l'optimització del perceptró multicapa mitjançant algorismes genètics.



Imatge 44. Resum dels models de classificació obtinguts.

Per altra banda, s'ha utilitzat la mètrica *Receiver Operating Characteristic* (ROC) per tal de comprovar el comportament dels diversos classificadors. Els següents gràfics mostren la corba ROC per a la classe incendis, és a dir, els casos positius, tant per als models originals com per a

les optimitzacions. Aquelles corbes amb una major àrea per sota es corresponen als models amb una major capacitat de diferenciar les dues classes: incendis i no incendis. Pel que fa als models originals els millors son: veïns més propers, *Random forest* i AdaBoost. Pel que fa als algorismes optimitzats, ambdós mostren un comportament millor que els models anteriors. D'aquests dos, el perceptró multicapa optimitzat genèticament assoleix una major capacitat de classificació.



Imatge 45. Corbes ROC dels diversos models de classificació.

8.3.3 Anàlisi de la importància dels atributs en l'estimació del risc d'incendi

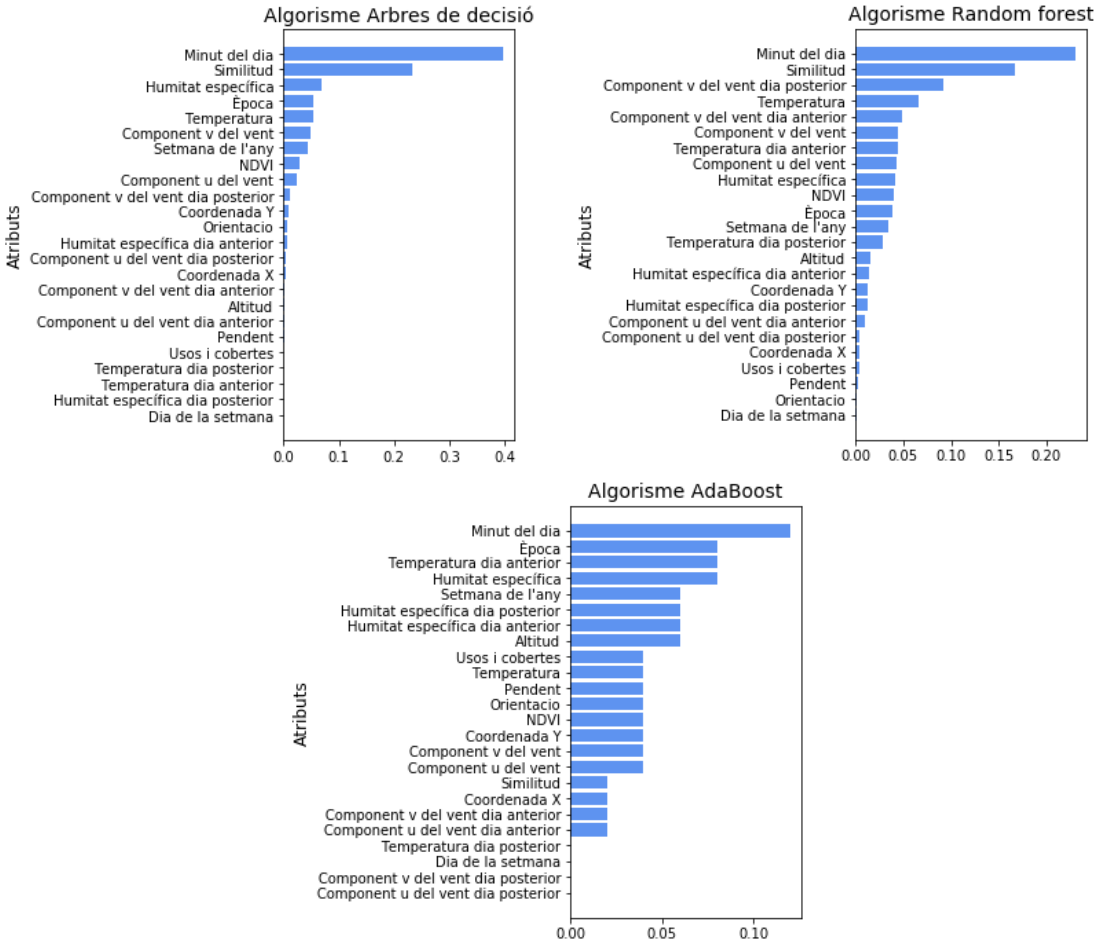
No és possible conèixer amb detall el pes de cada atribut en la classificació els models. És per això que s'ha centrat aquesta anàlisi en els algorismes implementats que ho permeten: els arbres de decisió, *Random forest* i AdaBoost.

Pel que fa als arbres de decisió, l'atribut més important és el moment del dia, seguit de: la similitud de les condicions meteorològiques, la humitat específica, la temperatura, l'època, la component *v* del vent i la setmana de l'any. Aquest model no ha utilitzat 7 dels atributs per a la classificació dels incendis: la component *u* del vent del dia anterior, els usos i cobertes del sòl, la temperatura tant del dia anterior com posterior a l'incendi, la pendent, la humitat específica del dia posterior i el dia de la setmana.

Com en el cas anterior, el model *Random forest* també ha donat més importància al moment del dia i a la similitud de les condicions meteorològiques que a la resta d'atributs. Tot seguit ha donat importància a la component *v* del vent (tant del dia actual com anterior i posterior), la temperatura tant del mateix dia com del dia anterior, la component *u* del vent, la humitat específica i l'índex NDVI. Els atributs amb menys importància han estat: la pendent, l'orientació, i el dia de la setmana.

En darrer lloc, en el model AdaBoost, com en els casos anteriors, l'atribut amb més pes ha estat el moment del dia. Per contra, el pes de la similitud de les condicions meteorològiques ha estat molt inferior. També ha donat importància a: l'època, la temperatura del dia anterior, la humitat específica tant del dia actual com de l'anterior i el posterior, la setmana de l'any i l'altitud.

Un dels objectius del projecte era valorar la utilitat de l'ús de dades meteorològiques del dia previ i posterior als incendis en els models d'aprenentatge atès que son dades que ajuden a modelitzar l'evolució de les condicions meteorològiques més enllà del moment d'ignició. L'anàlisi del pes dels atributs indica que aquests tenen importància en els diversos models. Concretament, el 3r i 5è atributs amb més pes per al model *Random forest* son les components *v* del vent del dia previ i posterior. Per altra banda, el model AdaBoost dona molt de pes a la temperatura del dia anterior i també a la humitat específica dels dies anterior i posterior.



Imatge 46. Importància dels atributs. Algorismes arbres de decisió, Random forest i AdaBoost.

9 Implementació. Generació dels mapes de risc

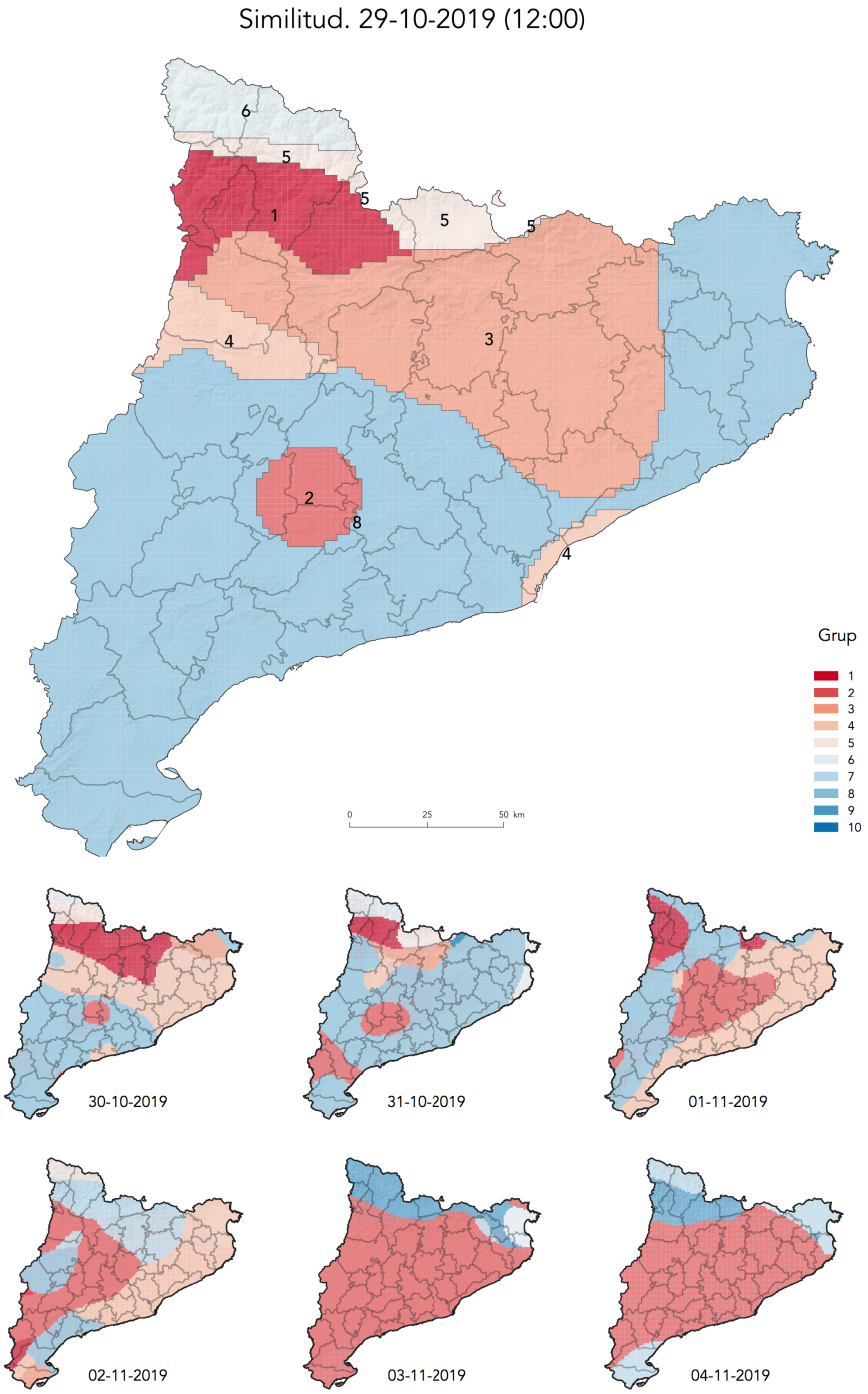
Un cop entrenats els models d'aprenentatge, aquests han estat utilitzat per a la generació de diversos mapes: mapes de ris d'incendi forestal segons les condicions meteorològiques i mapes d'estimació del risc d'incendi forestal.

9.1 Generació de mapes de zones de risc d'incendi forestal segons les condicions meteorològiques

Per a implementar els models obtinguts en l'apartat 8.1 per a la categorització de les zones segons les condicions meteorològiques s'han obtingut noves previsions meteorològiques del model TIGGE de l'ECMWF. L'ECMWF disposa de previsions des del 2006 fins a l'actualitat. S'han utilitzat les previsions corresponents a les 12:00 hores dels dies 29-10-2019 i els sis dies posteriors i del 27-07-2017 i els sis dies posteriors amb una resolució de 0,1°.

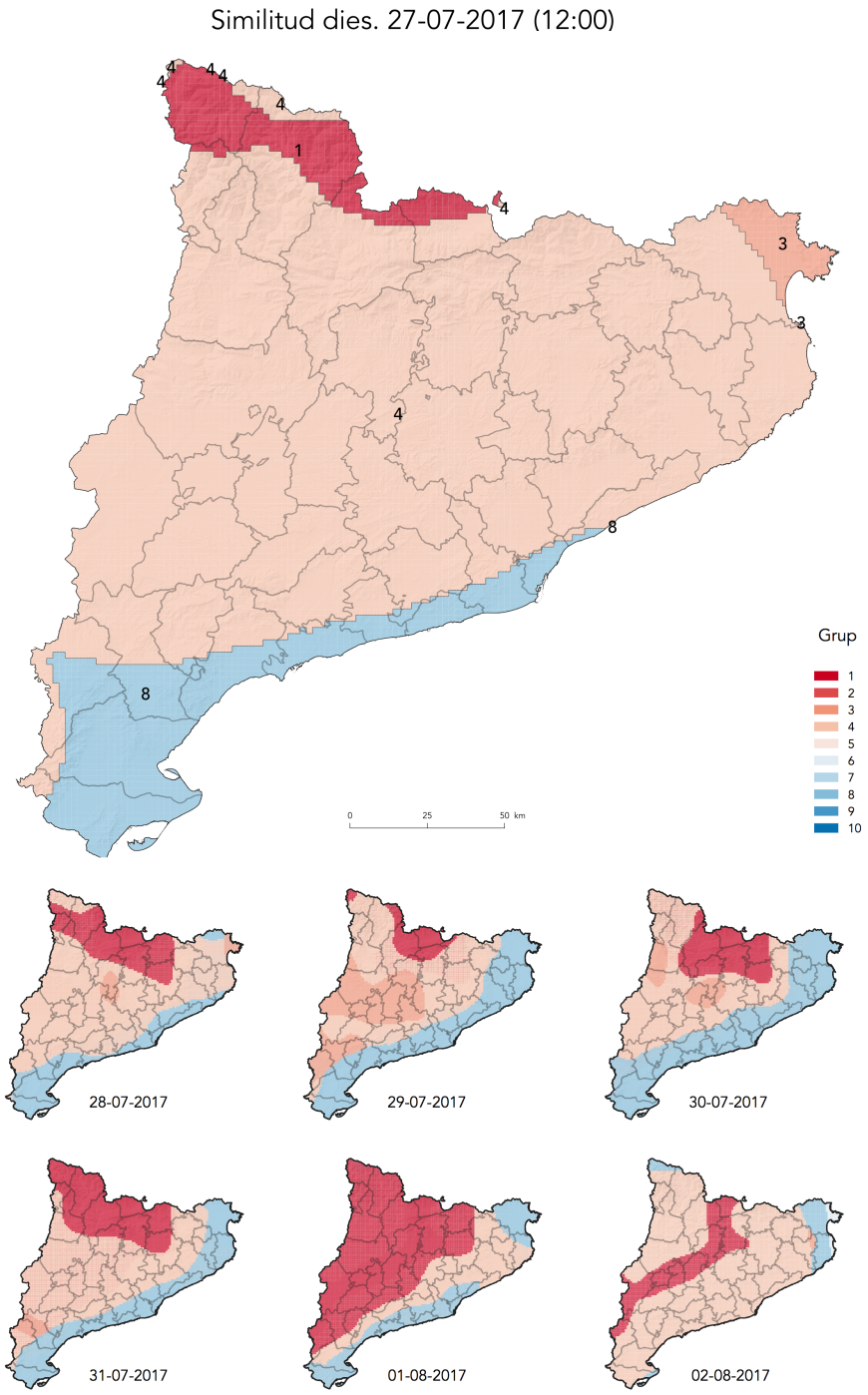
En primer lloc, s'han remostrejat les dades meteorològiques a una resolució de 0,01°. Tot seguit, s'han obtingut els mapes de risc classificant les diverses zones de l'àrea d'estudi en un dels 10 clústers obtinguts anteriorment amb la xarxa neuronal. Així, els següents mapes mostren els grups en què es classifiquen les diverses zones segons la similitud de les seves condicions meteorològiques respecte a algun dels 10 grups generats en l'apartat 8.1.

El primer dels mapes mostra les prediccions per al dia 29-10-2019 i els 6 dies posteriors, és a dir, una setmana de tardor potencialment amb un risc baix d'incendi. En aquesta primera predicció del model, la major part del territori queda agrupada en el clúster 8 amb un dels índexs de risc definits anteriorment més baix. Els resultats obtinguts son coherents amb les condicions meteorològiques i també amb el fet que en aquell dia no es va produir cap incendi.



Imatge 47. Mapa de similitud de les condicions meteorològiques del dia 29-10-2019 i sis dies posteriors. Estimacions a les 12:00.

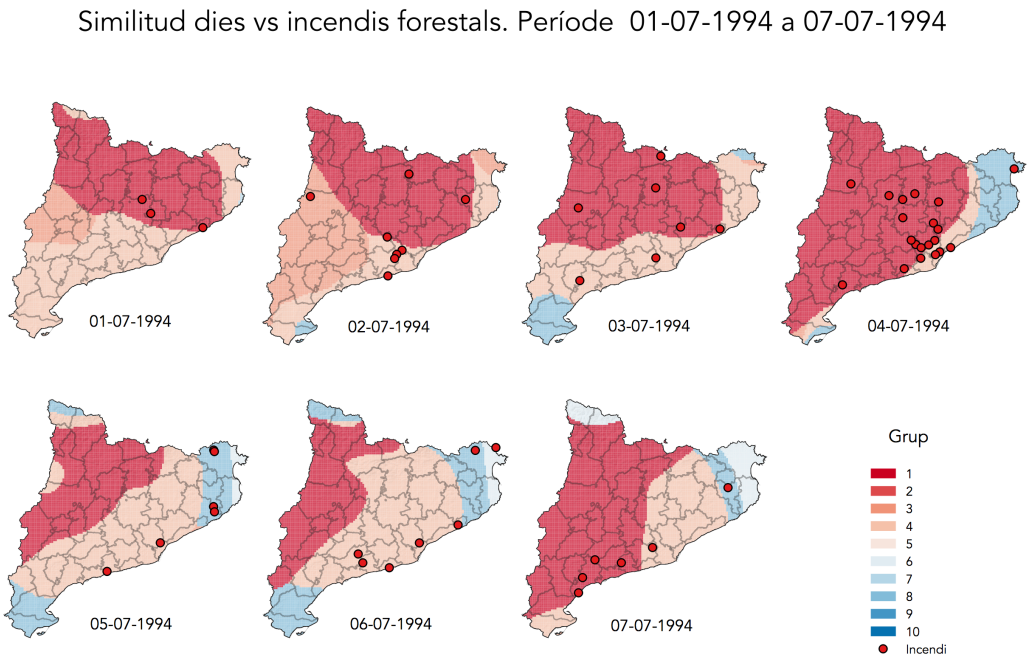
Per contra, la predicció per al dia 27-07-2017 i els sis dies posteriors mostra un predomini dels clústers 1, 2, 3 i 4, tots ells caracteritzats per un alt nombre d'incendis forestals i una quantitat de superfície cremada també elevada. Aquests resultats també son coherents amb una de les èpoques de l'any amb més incendis forestals a Catalunya com son els mesos de juliol i agost.



Imatge 48. Mapa de similitud de les condicions meteorològiques per a les 12:00 del dia 27-07-2017 i els sis dies posteriors.

En darrer lloc, s'han obtingut els mapes corresponents als dies de la primera onada d'incendis de l'estiu del 1994[39] on s'ha afegit la localització dels incendis iniciats en cadascun dels dies del període. S'ha escollit aquesta setmana atès que és una de les etapes del període d'estudi amb més concentració d'incendis.

[39] https://ca.wikipedia.org/wiki/Incendis_forestals_de_Catalunya_de_1994



Imatge 49. Mapa de similitud de les condicions meteorològiques per a les 12:00 del dia 01-07-1994 i els sis dies posteriors.

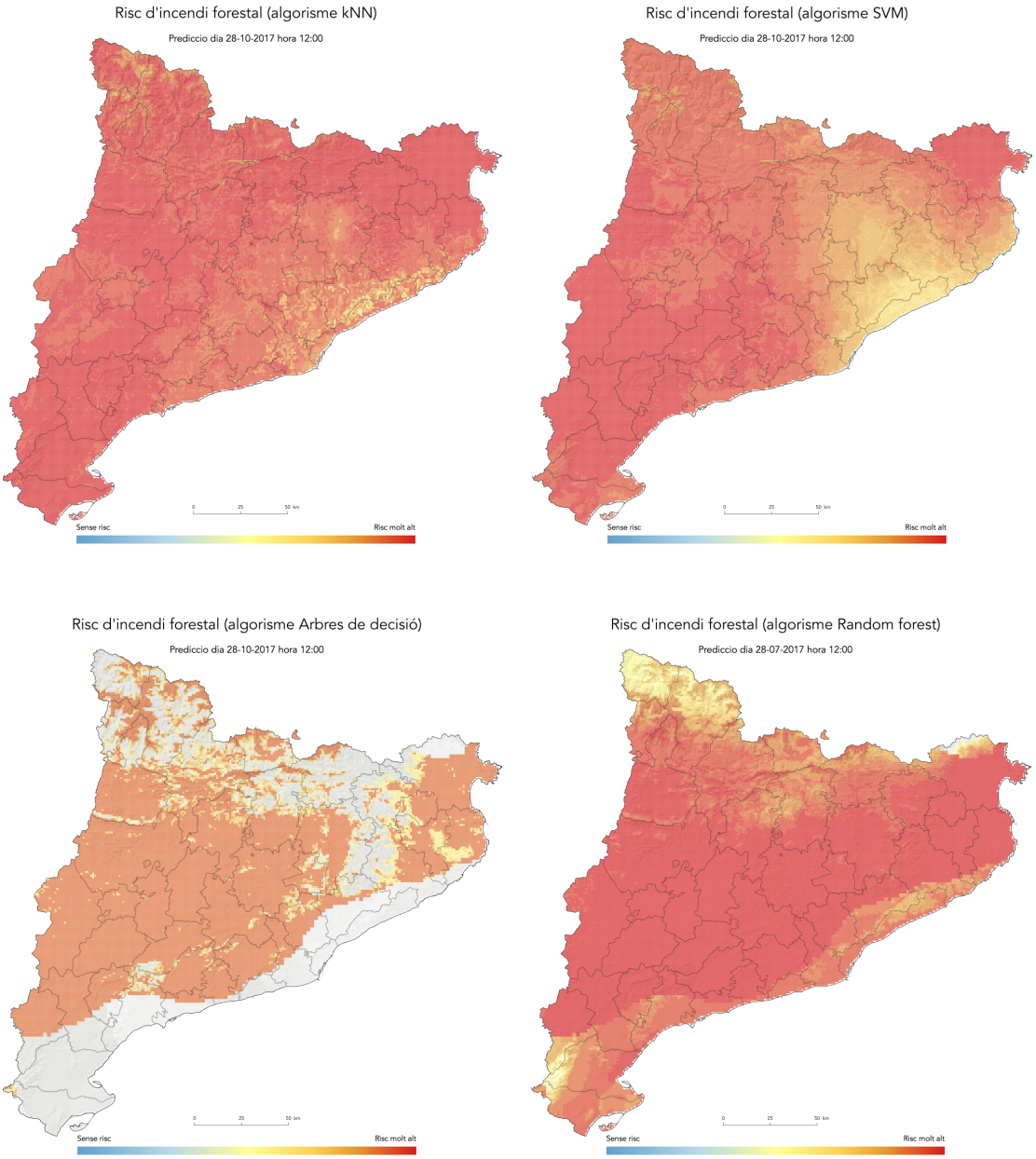
9.2 Generació de mapes de risc d'incendi forestal

Un cop entrenats els diversos classificadors, aquests han estat utilitzats per a la generació de mapes d'estimacio del risc d'incendi forestal per a les dues zones d'estudi: el conjunt del territori de Catalunya i el Parc Natural de Sant Llorenç del Munt i l'Obac. S'han obtingut mapes per a tres dates diferents amb la finalitat de poder-ne comparar els resultats.

En el cas dels arbres de decisió, tot i que la calibració de la probabilitats duta a terme ha millorat els resultats obtinguts, ja que s'ha passat d'un total de tan sols 6 probabilitats diferents a 21, aquests continuen essent menys detallats que en la resta de models. Això és degut al fet que aquest algorisme obté la probabilitat de què un exemple del conjunt de test pertanyi a una determinada classe com al nombre d'exemples seleccionats en una determinada fulla dividit pel total d'exemples seleccionats per aquesta mateixa fulla de l'arbre. Aquest comportament ha imposat dur a terme un conjunt d'experiments amb la profunditat màxima de l'arbre ja que valors baixos disminuïen la precisió obtinguda i valors alts augmentaven el nombre de fulles i, per tant, reduïen el nombre d'exemples en aquestes, limitant la capacitat de precisió en la probabilitat obtinguda.

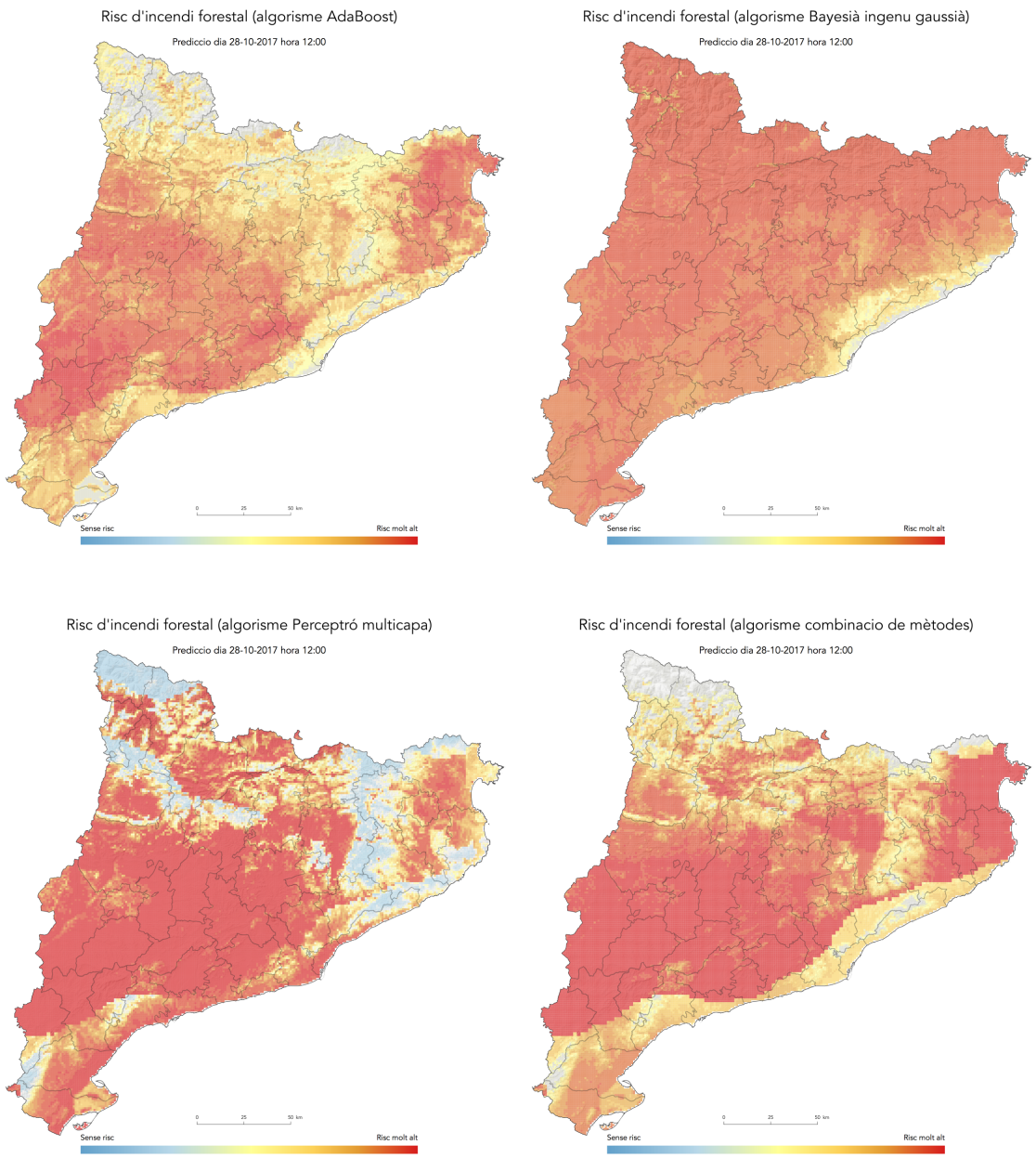
L'anterior limitació pot ser superada utilitzant el mètode *Random forest* ja que aquest obté diversos arbres i calcula la probabilitat com a la porció d'arbres que voten per una determinada classe. Això permet obtenir uns resultats més robustos.

Els següents mapes mostren el comportament dels models kNN, SVM, arbres de decisió i *Random forest* per a un dia de juliol del 2017, és a dir, un període de l'any on el risc d'incendi és potencialment superior.



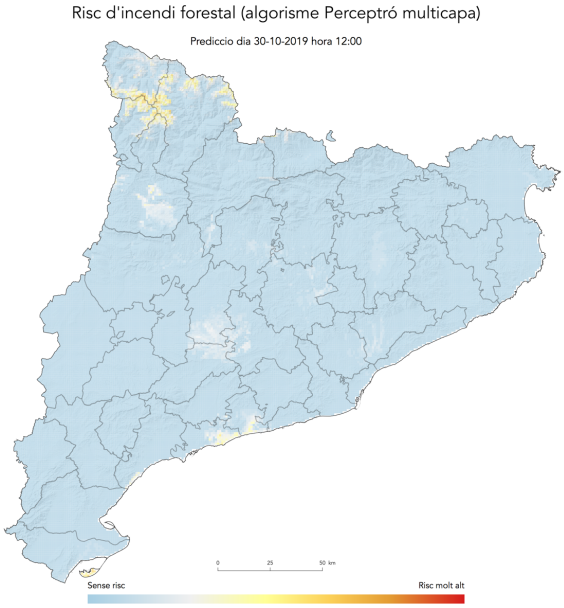
Imatge 50. Mapes d'estimació del risc d'incendi forestal. Models: kNN, SVM, arbres de decisió i Random forest.

Els següents mapes mostren l'estimació del risc d'incendi per a la resta de models on es pot comprovar que el comportament alhora d'estimar la probabilitat de la classificació varia en cadascun dels algorismes. Per altra banda, també s'aprecia un pes diferent dels diversos atributs com ara el atributs meteorològics donant lloc a estimacions diferents.



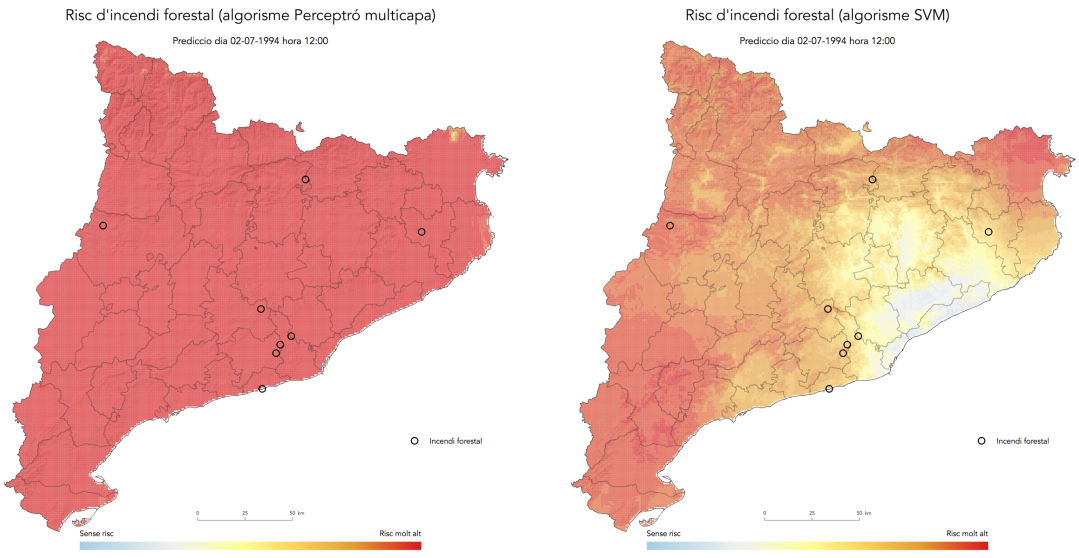
Imatge 51. Mapes d'estimació del risc d'incendi forestal. Models: AdaBoost, bayesià ingenu gaussià, perceptró multicapa i combinació de mètodes.

Després d'haver obtingut els mapes de predicció del risc d'incendi forestal per a un dia amb un risc alt s'han generat mapes per a un dia de tardor amb risc baix i en què no s'ha produit cap incendi forestal per a poder-ne comparar els resultats. El següent és un dels mapes obtinguts, concretament amb el perceptró multicapa optimitzat genèticament, on no s'assoleix un risc d'incendi superior al 50% en la major part del territori.



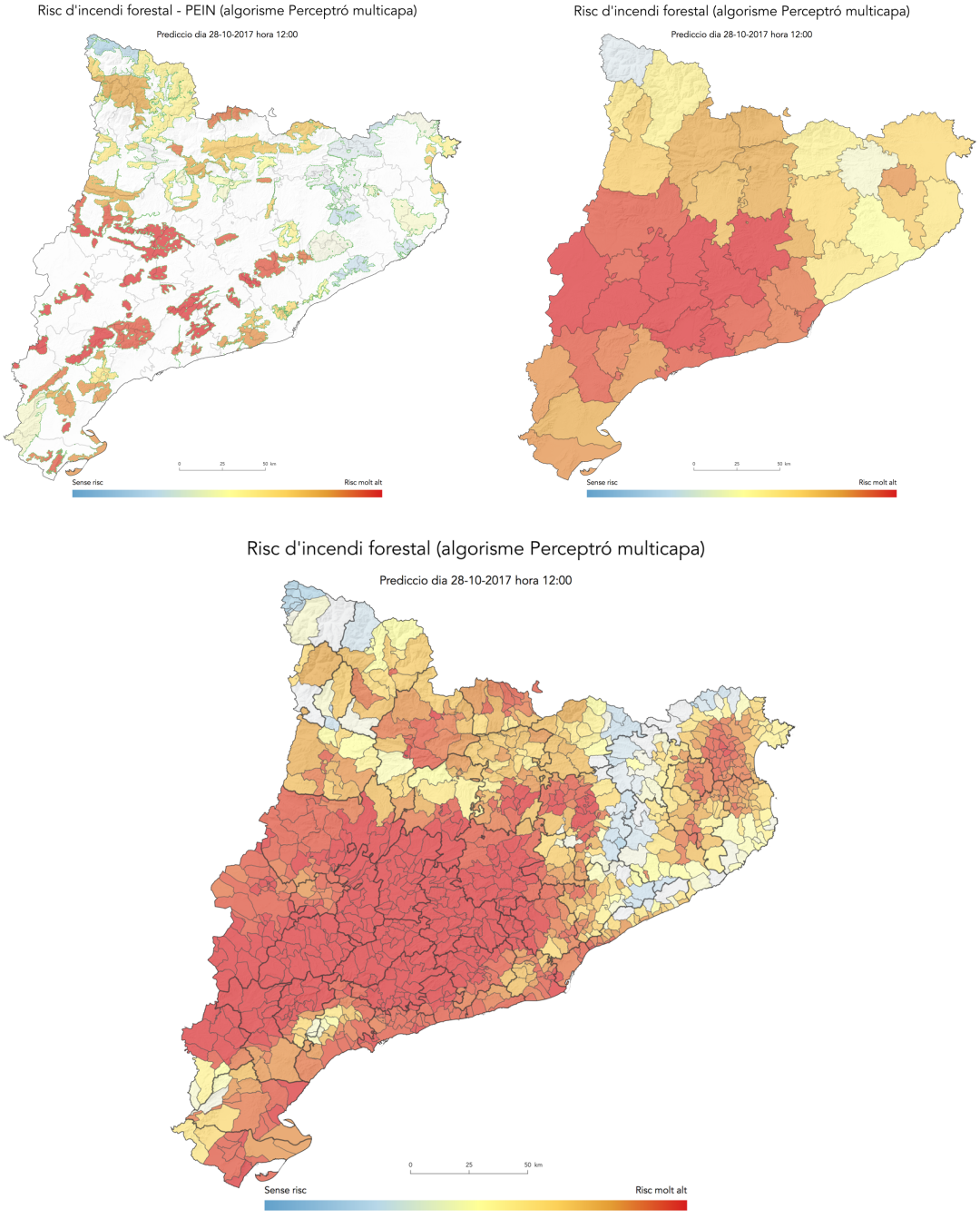
Imatge 52. Mapa d'estimació del risc d'incendi forestal obtingut amb el perceptró multicapa.

Finalment, s'ha obtingut el mapa de risc d'incendi forestal per a un dels dies de la onada d'incendis de l'estiu de l'any 1994. S'ha escollit un dia on es varen iniciar un elevat nombre d'incendis forestals per a poder comparar la seva localització amb la predicció de l'algorisme. Com es pot apreciar en el següent mapa, la predicció obtinguda per al dia 02-7-1994 és de risc molt alt d'incendi en la major part de l'àrea d'estudi i la localització dels punts d'inici d'incendi coincideixen amb zones amb estimacions de risc d'incendi molt altes.



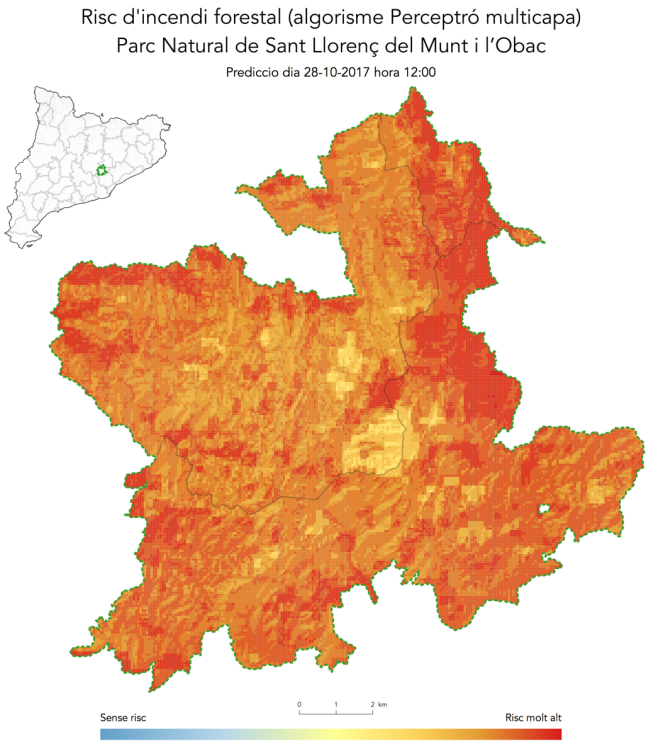
Imatge 53. Estimació del risc d'incendi forestal. Models: perceptró multicapa i SVM.

El detall assolit amb els anteriors models permeten l'obtenció d'estimacions per a diversos àmbits com, per exemple: municipal, comarcal o bé per al principals espais naturals.

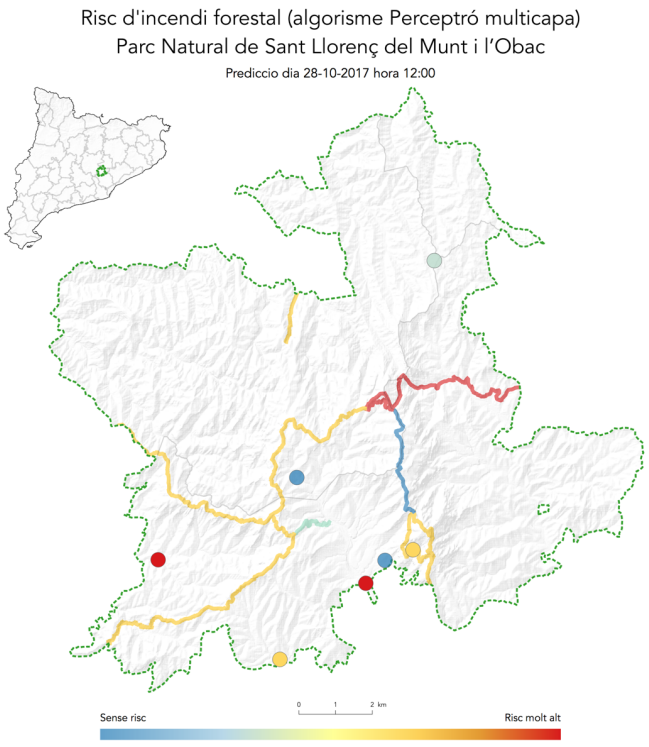


Imatge 54. Risc mitjà per espais naturals, comarques i municipis.

Un cop obtingudes les estimacions per a la primera zona d'estudi s'han realitzat proves per a la segona zona d'estudi, el Parc Natural de Sant Llorenç del Munt i l'Obac amb la finalitat d'avaluar el nivell de resolució espacial que ofereixen els algorismes emprats. Els resultats aconseguits permeten obtenir l'estimació del risc d'incendi per a les diverses zones del parc i fins i tot en determinats indrets sensibles com, per exemple, els principals senders i les zones d'estacionament de vehicles. Aquests mapes permetrien gestionar l'accés a determinades zones del parc natural en funció de l'estimació del risc d'incendi.



Imatge 55. Mapa del risc d'incendi forestal del Parc Natural de Sant Llorenç del Munt.



Imatge 56. Mapa del risc d'incendi forestal dels principals senders i zones d'estacionament del Parc Natural de Sant Llorenç del Munt.

10 Conclusions

La principal conclusió que s'ha extret del projecte és la idoneïtat de les tècniques d'aprenentatge automàtic alhora de predir incendis forestals. En contraposició a altres mètodes amb un major cost computacional com ara les simulacions i els models matemàtics, els mètodes utilitzats han permès obtenir models i prediccions de forma molt ràpida i sense la necessitat de dades complexes o de difícil obtenció.

Per altra banda, cal dir que una part important del temps dedicat a aquest tipus de projectes se centra en l'anàlisi i la preparació de les dades. Atès l'alt volum de dades, que han superat els 12 GB, és necessari automatitzar al màxim els processos d'extracció i transformació d'aquestes així com l'ús de tecnologia *big data*.

Alhora de generar els conjunts de dades per a l'entrenament dels models, un dels principals problemes ha estat l'obtenció d'exemples negatius que permetessin als algorismes aprendre a diferenciar entre incendis i no incendis. La dificultat rau, per una banda, en obtenir conjunts ben balancejats i, de l'altra, en el fet que tots els indrets i moments en què no s'han produït incendis no sempre es corresponen a situacions de baix risc d'incendi. Per a resoldre aquests inconvenients s'ha proposat en aquest projecte l'ús de similitud de Pearson per eliminar els exemples negatius excessivament similars als incendis i la utilització del mètode *bagging* per a minimitzar l'efecte dels exemples mal etiquetats en els models de classificació.

Un cop obtingudes les dades per a l'entrenament ha estat necessari adaptar les diverses tècniques d'aprenentatge als problemes que calia resoldre. En el cas de l'obtenció de zones de risc d'incendi segons les condicions meteorològiques ha calgut estimar el nombre de clústers en els quals s'agrupen aquestes condicions ja que aquesta dada no es coneixia d'entrada. El fet que el mètode del colze no oferís resultats concloents ha requerit l'ús d'algorismes que permetessin obtenir aquest valor. En aquest projecte s'ha proposat la utilització conjunta de diversos d'aquests mètodes per a validar el nombre final de clústers com son: DBSCAN, OPTICS, propagació de l'afinitat i BIRCH.

Pel que fa als dos conjunts de dades meteorològiques utilitzades, ambdos han estat vàlids alhora d'obtenir grups d'exemples en funció de les condicions meteorològiques. Per una banda, els reanàlisis ERA5, a diferència dels anàlegs, permeten obtenir models amb detall suficient per a establir zones de risc en l'àrea d'estudi. Per altra banda, els reanàlisis han donat bons resultats amb les dades utilitzades. Cal tenir en compte que només s'ha disposat de dades de quatre mesos de l'any: juny, juliol, agost i setembre i per això, es recomana un futur anàlisi amb un conjunt superior que cobreixin tant un període d'estudi major com tots els mesos de l'any.

Pel que fa als models de regressió, aquests han obtingut una predicció pobra de la mida dels incendis en el moment d'ignició. Podem concloure que el conjunt de dades disponibles no son suficients per a resoldre aquest problema i és per això que es proposa utilitzar un conjunt de característiques diferents per a l'estimació de la mida dels incendis, concretament amb més atributs meteorològics. Cal dir que l'ampliació del conjunt de dades original amb més dades orogràfiques com ara la rugositat i dades d'altitud, pendent, i orientació a diversos radis del punt d'ignició han millorat els resultats però no han estat suficients per a obtenir prediccions vàlides.

Finalment, quant a l'estimació del risc d'incendi forestal a través d'algorismes supervisats de classificació, els primers models obtinguts han mostrat la necessitat de realitzar un anàlisi de la calibració de les probabilitats predites per tal d'obtenir resultats comparables entre els diversos models.

Respecte a l'optimització dels algorismes de classificació s'han obtinguts bons resultats tant amb l'ús de la combinació de mètodes com amb l'optimització mitjançant algorismes genètics de l'arquitectura d'un perceptró multicapa per a l'estimació del risc d'incendi. En aquest darrer cas, s'han assolit els millors resultats amb precisions del 90%. Tot i l'alt cost computacional d'aquest mètode, es recomana el seu ús ja que ha permès obtenir models de predicció robustos. Per altra banda, en alguns dels models utilitzats s'ha identificat la necessitat d'augmentar el nombre d'exemples d'entrenament per a obtenir millors resultats, concretament en els models: perceptró multicapa, combinació de mètodes, Random forest i AdaBost.

Finalment, la implementació dels diversos models obtinguts han permès obtenir dos tipus de mapes de risc diferents. Per una banda, mapes de zones de risc d'incendi segons la meteorologia que permeten identificar zones del territori amb unes determinades condicions meteorològiques i conèixer les principals característiques dels incendis potencials en aquestes zones. Aquesta informació pot ser útil en la prevenció i la gestió dels incendis forestals. Per altra banda, s'han obtingut mapes d'estimació del risc d'incendi amb suficient detall en les prediccions per a diversos àmbits (comarques, municipis, espais naturals, etc.) i fins i tot vàlids per zones concretes com la segona zona d'estudi, el Parc Natural de Sant Llorenç del Munt i l'Obac on els resultats assolits han permès obtenir l'estimació del risc d'incendi per a les diverses zones del parc i, fins i tot, en determinats indrets sensibles com, per exemple, els principals senders i les zones d'estacionament de vehicles.

Es pot afirmar que la major part dels experiments han assolit precisions i models robustos. Per un costat, els models basats en les condicions meteorològiques permeten classificar nous exemples de prediccions meteorològiques en un dels 10 grups establerts amb un 99% de precisió, i els models basats en anàlegs en un dels 5 grups d'anàlegs identificats amb un 98% de precisió. Per altra banda, els models de classificació utilitzats per a obtenir les prediccions de risc d'incendi forestal han assolit el 90% de precisió.

El projecte també ha permès analitzar el pes dels diversos atributs alhora de realitzar prediccions del risc d'incendi. Concretament, s'ha confirmat la utilitat de l'ús de dades meteorològiques del dia previ i posterior als incendis en els models d'aprenentatge, l'ús de la similitud de les condicions meteorològiques obtingudes prèviament mitjançant els mètodes d'agrupament i el pes d'alguns atributs com: el moment del dia, la humitat específica, l'època, la temperatura i la component nord del vent.

Cal dir que les dades utilitzades i els models obtinguts, i més concretament aquells basats en reanàlisi ERA5 poden ser utilitzats per a qualsevol zona i no només per a les zones d'estudi d'aquest projecte. Així és que, els resultats d'aquest projecte es poden reproduir en altres zones.

En conjunt, s'han assolit tots els objectius tant generals com específics definits a l'inici del projecte. Concretament, s'han tractat dades espai-temporals amb tecnologia geoespacial per al posterior entrenament d'algorismes d'aprenentatge automàtic i també per a la interpretació dels resultats dels models obtinguts i s'ha enriquit els models tant amb dades orogràfiques com meteorològiques prèvies i posteriors al moment d'ignició. Per altra banda, s'han implementat models en l'entorn *big data* Apache Spark per a facilitar el tractament de gran volums de dades. Finalment, s'han emprat tant algorismes de categorització per agrupar les zones de risc segons les condicions meteorològiques, algorismes de regressió per a l'estimació de la mida dels incendis en el moment de la ignició i algorismes de classificació per a estimar el risc d'incendi forestal. Tot això ha permès l'obtenció de models predictius de risc per a les zones d'estudi.

11 Glossari

AdaBoost *m* mètode d'aprenentatge supervisat basat en un conjunt elevat de regles senzilles per a crear classificadors robustos.

agrupament jeràrquic *m* mètode de categorització basat en l'aglomeració o divisió de les dades.

algorismes genètics *m* algorismes d'optimització que simulen la selecció natural sobre un conjunt d'individus per a cercar la millor solució a un problema.

anàlisi de components principals *m* tècnica per a la representació de dades en un espai de dimensionalitat inferior.
en **principal component analysis**.
sigla **PCA**.

anàlisi geoespacial *m* anàlisi de dades centrat en elements amb localització geogràfica.
en **geospatial analyst**.

Apache Flink *m* entorn de treball de codi obert de processament de fluxos de dades amb la possibilitat de treballar en mode per lots com a un cas especial de processament de fluxos.

Apache Hadoop *m* entorn de treball de codi obert que fa possible el processament distribuït de grans volums de dades mitjançant un clúster d'ordinadors.

Apache Mahout *m* biblioteca de codi obert d'algorismes d'aprenentatge automàtic escalables escrita en Java que pot ser utilitzada quan el volum de dades que cal processar és molt gran.

Apache Flink *m* entorn de treball de processament de fluxos que utilitza el sistema de missatges Apache Kafka per a garantir la tolerància a les fallades, memòria intermèdia i emmagatzematge d'estat.

Apache Spark *m* entorn de treball de codi obert per al processament distribuït en clústers d'ordinadors de grans volums de dades.

Apache Storm *m* entorn de treball de codi obert de processament de fluxos de grans volums de dades amb una latència molt baixa i, per tant, indicat per al processament proper al temps real.

aprenentatge automàtic *m* camp de la intel·ligència artificial que estudia tècniques per a proveir a les màquines de la capacitat d'aprendre.
en **machine learning**.

aprenentatge no supervisat *m* tipus d'aprenentatge on no es disposa de cap mena d'informació sobre les sortides.

aprenentatge profund m aprenentatge basat en xarxes neuronals amb diverses capes ocultes.

aprenentatge supervisat m tipus d'aprenentatge on es coneix quina és la resposta del sistema.

arbres de decisió *m* mètode d'aprenentatge supervisat basat en regles.

bagging *m* mètode de conjunts de classificadors consistent en dividir el conjunt d'exemples en subconjunts de casos i entrenar un classificador amb cadascun d'aquests subconjunts.

categorització *m* mètodes d'aprenentatge no supervisat capaços d'obtenir un conjunt de categories a partir d'un conjunt d'objectes inicials.
en **clustering**.
sinònim **agrupament**.

classificació *f* mètodes d'aprenentatge supervisat on es disposa d'un conjunt d'objectes dels que es coneix la classe de sortida.
en **classification**.

Climate Data Sotere *m* organisme europeu encarregat de proveir informació climàtica: observacions satèl·lit, mesures, projeccions de models climàtics i previsions estacionals.

coeficient de correlació de Pearson *m* mesura de correlació lineal entre dues variables.

coeficient de silueta *m* mètode utilitzat per a l'avaluació de la consistència dels agrupaments de dades.

dades massives *f* conjunt de tecnologies per al tractament de grans volums de dades de fonts diverses i on la velocitat de processament és important.
en **big data**

Density-based spatial clustring of Applications with noise *m* mètode de categorització basat en àrees d'alta densitat.
sigla **DBSCAN**.

EarthData *m* servei del la NASA que té la finalitat de posar a disposició del la comunicat científica i la societat en general les dades relacionades amb ciències de la Terra de la NASA .

ERA5 *m* producte de l'ECMWF format per dades climàtiques fruit de la combinació d'observacions històriques i models.

European Centre for Medium-Range Weather Forecast *m* institució europea encarregada d'elaborar prediccions meteorològiques.
sigla **ECMWF**.

H_2O *m* eina de codi obert per a l'aprenentatge automàtic amb interfície gràfica d'usuari.

índex Davies Bouldin *m* mètode utilitzat per a l'avaluació de la consistència dels agrupaments de dades.

informàtica en núvol *f* model que permet l'accés en xarxa a recursos informàtics compartits sota demanda i des de qualsevol indret.
en **cloud computing**.

Institut Cartogràfic i Geològic de Catalunya *m* entitat responsable de desenvolupar les tasques d'informació cartogràfica i geològiques competència de la Generalitat de Catalunya.
sigla **ICGC**.

k-mitjanes *m* mètode de categorització basat en l'assignació dels exemples al centroide més proper.
en **k-means**.

MapReduce *m* model de programació per al processament de grans volums de dades que utilitza computació paral·lela i distribuïda.

màquines de vectors de suport *m* mètode de classificació que permet classificar objectes definits per atributs numèrics en dues classes.
en **suport vector machines**.
sigla **SVM**.

MLlib *m* biblioteca per a l'aprenentatge automàtic escalable sobre l'entorn Apache Spark. Suporta diversos llenguatges de programació com: Java, Scala, Python i R.

Naïves Bayes *m* mètode d'aprenentatge supervisat fonamentat en models probabilístics.

normalized difference vegetation index *m* índex de vegetació que permet estimar la qualitat, quantitat i el desenvolupament de la vegetació a partir de dades de la intensitat de radiació de diverses bandes de l'espectre electromagnètic.
sigla **NDVI**.

perceptró multicapa *m* tipus de xarxes neuronals on cada capa aprèn característiques més complexes que parteixen precisament de l'aprenentatge de les capes anteriors.

processament de fluxos *m* paradigma de programació que permet a les aplicacions el processament en paral·lel.
en **stream processing**.

processament per lots *m* processament on les dades rebudes no es processen immediatament ja que el temps de resposta no és rellevant.
en **batch processing**.

propagació de l'afinitat *m* mètode de categorització basat en l'enviament de missatges.

Python *m* llenguatge de programació interpretat de propòsit general i d'alt nivell.

QGIS *m* programari SIG de codi lliure i multiplataforma.

Random forest *m* variant del mètode de classificació arbres de decisió que utilitza l'agregació de bootstrap per a millorar l'estabilitat i la precisió de la classificació tot evitant el sobreentrenament.

random patches *m* mètode de conjunts de classificadors consistent en dividir el conjunt original tant en l'espai d'atributs com en l'espai de mostres.

random subspace methods *m* mètode de conjunts de classificadors consistent en dividir el conjunt d'atributs eliminant de forma aleatòria alguns d'aquets atributs.
sigla **RSM**.

regressió *f* mètodes d'aprenentatge supervisat on es disposa d'un conjunt d'objectes dels que es coneix el valor de sortida.
en **regression**.

SAMOA *f* plataforma dissenyada per a l'aprenentatge sobre dades distribuïdes en temps real.

Servei de Prevenció d'Incendis Forestals *m* organisme de la Generalitat de Catalunya responsable, entre d'altres, dels plans de prevenció d'incendis forestals i de la definició de les zones de risc d'incendi i els índex de perill.

sistema d'informació geogràfica *m* sistema d'informació que permet la recopilació, emmagatzematge, gestió, anàlisi, visualització, consulta i presentació de tot tipus d'informació geoespacial.
en **geographic information system**.
sigla **SIG**.

veí més proper *m* mètode de classificació basat en els algorismes de categorització que classifica un objecte a partir de l'objecte, o objectes, més proper.
en **k-nearest neighbour**.
sigla **k-NN**.

xarxes neuronals *f* mètode de classificació que simula les propietats dels sistemes neuronals dels éssers vius mitjançant model matemàtics.

12 Bibliografia

ALKATHERI, Safaa; ABBAS, Samah Anwar; SIDDIQUI, Miazzam Ahmed (2019). "A Comparative Study of Big Data Frameworks". *International Journal of Computer Science and Information Security* (vol. 17, núm 1, pàg. 66-73).

CALBO, Josep; CUNILLERA, Jordi; LLASAT, Carme *i altres* (2019). "Projeccions climàtiques per a Catalunya". *Aigua i canvi climàtic. Diagnosi dels impactes previstos a Catalunya* (2009). Barcelona: Agència Catalana de l'Aigua - Generalitat de Catalunya.

CASTELLI, Mauro; VANNESCHI, Leonardo; POPOVIŠ, Aleš (2019). "Predicting Burned Areas of Forest Fires: an Artificial Intelligence Approach". *Fire Ecology* (núm 11, pàg 106-118).

COFFIELD, Shane R.; GRAFF, Casey A.; CHEN, Yang *i altres* (2019). "Machine learning to predict final fire size at the time of ignition". *International Journal of Wildland Fire*.

ELLINGWOOD, Justin (2016). *Hadoop, Storm, Samza, Spark and Flink: Big Data Frameworks Compared*. Digital Ocean. [Data de consulta: 27 de desembre de 2019] <https://www.digitalocean.com/community/tutorials/hadoop-storm-samza-spark-and-flink-big-data-frameworks-compared#apache-hadoop>

GHORBANZADEH, Omid; KAMRAN, Khalil Valizadeh; BLASCHKE, Thomas *i altres* (2019). "Spatial Prediction of Wildfire Susceptibility Using Field Survey GPS Data and Machine Learning Approaches". *Fire* (núm 2).

GHORBANZADEH, Omid; ROSTAMZADEH, Hashem; BLASCHKE, Thomas *i altres* (2018). "A new GIS-based data mining technique using an adaptive neuro-fuzzy inference system (ANFIS) and k-fold cross-validation approach for land subsidence susceptibility mapping". *Natural Hazards* (núm 94, pàg 497-517).

INTERGOVERNMENTAL PANEL ON CLIMATE CHANGE (2014). *Climate Change 2015, Impact, Adaptation, and Vulnerability*: Cambridge University.

JAAFARI, Abolfazl; ZENNER, Eric K.; PANAHI, Mahdi *i altres* (2019). "Hybrid artificial intelligence models based on a neuro-fuzzy system and metaheuristic optimization algorithms for spatial prediction of wildfire probability". *Agricultural and Forest meteorology*.

KIN, Sea Jin; LIMS, Chul-Hee; KIM, Gang Sun *i altres* (2019). "Multi-Temporal Analysis of Forest Fire Probability Using Socio-Economic and Environmental Variables". *Remote Sensing* (núm 11).

LANDSET, Sara; KHOSHGOFTAAR, Taghi M.; RICHTER, Aaron N.; HASANIN, Tawfiq (2015). "A survey of open source tools for Machine learning with Big Data in the Hadoop ecosystem". *Journal of Big Data* (núm. 2:24).

LANEY, Doublas (2001, febrer). "3D Data management: Controlling Data Volume, Velocity and Variety". *Gartner*.

MELL, Peter; GRANCE, Timothy (2011). *The NIST Definition of Cloud Computing*. Gaithersbur: National Institute os Standards and Technology. [Data de consulta: 27 de desembre de 2019] <http://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf>

OUSOOUS, Ahmed; BENJELLOUN, Fatima-Zahra; LAHCEN, Ayoub Ait *i altres* (2017). "Big Data tecnologies: A survey". *Computer and Information Sciences* (núm 30, pàg 431-448).

POURGHASEMI, Hamid Reza; BEHESHTIRAD, Masood; PRADHAN, Biswajeet (2014). "A comparative assessment of prediction capabilities of modified analytical hierarchy process M-AHP and Mamdani fuzzy logic

models using Netcad GIS for forest fire susceptibility mapping". *Geomatics, Natural Hazards and Risk* (núm 7, pàg 861-885).

RADKE, David; HESSLER, Anna; ELLSWORTH, Dan (2019). "FireCast: Leveraging Deep Learning to Predict Wildfire Spread". [ponència]. A: *Twenty-Eighth International Joint Conference on Artificial Intelligence.* Macao.

RAMON REVILLA, Anna (2019). *Catalunya s'haurà d'adaptar i aprendre a conviure amb el foc* [Data de consulta: 27 de desembre de 2019] <http://blog.creaf.cat/noticies/catalunya-shaura-dadaptar-aprendre-conviure-amb-el-foc/>

ROLNICK, David; DONTI, Priya L.; KAACK, Lynn H. *i altres* (2019). "Tackling Climate Change with Machine Learning". Cornell University.

SAFI, Youssef; BOUROUMI, Abdelaziz (2013). "Prediction of forest fires using Artificial Neural Networks". *Applied Mathematical Sciences* (núm 6, pàg 271-286).

SAKELLARIUS, Stavros; TAMPEKIS, Stergios; SAMARA, Fani *i altres* (2017). "Review of State-of-the-art decision suport Systems (DSSs) for prevention and suppression of forest fires". *Journal of Forestry Research* (núm 28, pàg 1107-1117).

SAYAD, Younes Oulad; MOUSANNIF, Hajar; MOATASSIME, Hassan Al (2019). "Predictive modeling of wildfires. A new dataset and machine learning approach". *Fire Safety Journal* (núm 104, pàg 130-146).

SU, Zhangwen; HU, Haiqing; WANG, Guangyu *i altres* (2018). "Using GIS and Random Forests to identify fire drivers in a forest city Yichun China". *Geomatics, Natural Hazards and Risk* (núm 9, pàg 1207-1229).

SUBRAMANIAN, Sriram Ganapathi; CROWLEY, Mark (2018). "Using Spatial Reinforcement Learning to Build Forest Wildfire Dynamics Models From Satellite Images". *Frontiers in ICT* (núm 5).

Tercer informe sobre el canvi climàtic a Catalunya (2016). Barcelona: Generalitat de Catalunya / Institut d'Estudis Catalans.

TURNER, Vernon (2014). *The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things*. EMC. [Data de consulta: 22 de maig de 2017] <https://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm>

ZHANG, Guoli; WANG, Ming; LIU, Kai (2019). "Forest Fire Susceptibility Modeling Using a Convolutional Neural Network for Yunnan Privince of China". *International Journal of Disaster Risk Science* (núm 10).