Universitat Autònoma de Barcelona

Anàlisi de la Probabilitat d'Incendi en Habitatges

Treball de Final de Grau

Autora: Laia Escursell Rof

Professorat tutor: Gemma Sanjuan Gomez

20 de gener de 2025

Repositori GitHub

Matemàtica Computacional i Analítica de Dades

Índex

Capítol 1

Introducció

El 2023, a Espanya hi va haver 249 morts per incendis. La xifra més elevada des que es va començar l'estudi el 2010[8]. La gran part d'aquestes morts es van produir en habitatges. I als Estats Units, el 2022, hi va haver 374.300 incendis, 2.720 morts, 10.250 ferits i 10.821.300.000$ de pèrdues[1]. Això ens indica que els incendis representen un problema greu de seguretat tant per les pèrdues humanes com materials.

Per aquest motiu, aquest projecte té com a objectiu analitzar quins factors augmenten la perillositat dels incendis en habitatges utilitzant dades de la base *National Fire Incident Reporting System (NFIRS)*. És a dir, mirar quins factors poden fer que l'incendi estigui limitat a l'objecte incendiat, a l'habitació o a l'edifici.

En l'última dècada, els estudis centrats a aquest tema han començat a tenir una mica més d'importància. A Espanya han sortit diverses estadístiques sobre incendis, però la majoria centrades en incendis forestals[4][3].

També hi ha algun estudi que se centren a estimar el risc de mort que hi ha quan es produeix un incendi en un habitatge[10]. I d'altres que busquen poder predir la probabilitat anual mitjana de tenir una experiència d'incendis residencials al llarg tota la vida d'una persona[9].

Capítol 2

Objectiu Principal

Com que no hi ha estudis on la finalitat sigui mirar, en el cas de patir un incendi, si aquest seria més greu o menys, l'objectiu principal d'aquest treball és crear un model predictiu que permeti determinar la gravetat dels incendis en habitatges a partir de les dades recollides pel sistema *NFIRS*. Aquest model ens ha de donar informació útil per identificar els factors claus que influeixen en la gravetat dels incendis. D'aquesta manera, es podran millorar les estratègies de prevenció i estratègia per aquest tipus d'incendis.

Per tal de poder complir l'objectiu principal del treball, necessitarem complir-ne d'altres. Necessitarem fer una exploració i comprensió de la Base de Dades per a seleccionar les variables més rellevants pel nostre treball. També analitzarem com els factors socioeconòmics poden influir en la severitat de l'incendi. I mentre fem aquesta anàlisi també mirarem quins altres factors són determinants. Com el fet de tenir instal·lats detectors de fum a casa i quins acostumen a ser els factors que originen l'incendi. Són per falta de conscienciació de la gent, per falta de manteniment o són esdeveniments aleatoris?

Capítol 3

Recopilació i Comprensió de Dades

Comencem el treball explicant d'on hem recopilat les dades per fer el nostre estudi. I fent una comprensió d'aquestes per a determinar quines són les més necessàries per poder aconseguir el nostre objectiu. Explicarem el procediment que hem seguit per seleccionar la informació rellevant i quines eines i tècniques hem utilitzat per comprendre el significat de cada variable. L'objectiu és garantir la fiabilitat de la Base de Dades per poder fer posteriorment una correcta interpretació.

3.1 La Base NFIRS

El Sistema Nacional d'Informes d'Incidents de Focs (*NFIRS*, les sigles en anglès)[5] recopila els informes que fan de forma voluntària els parcs de bombers dels Estats Units, on els informes tenen un registre de la totalitat de les seves activitats. El rang de les activitats és variat, des de la resposta a incendis fins a serveis mèdics d'emergència i desastres naturals.

La *NFIRS* és la base de dades nacional anual més gran del món sobre incendis. L'any 2021, uns 22.300 parcs de bombers van reportar-hi dades, fent que es registressin més de 29 milions d'informes, dels quals 1,1 milions són respostes a incendis.

Tot i que la participació en el *NFIRS* és voluntària, el sistema recull aproximadament el 70% de tots els incidents d'incendis que es produeixen anualment als Estats Units. Pel que fa a la *NFIRS* sigui una eina important a l'hora d'analitzar la seguretat contra incendis i la resposta a emergències. Gràcies a aquesta recopilació de dades es poden identificar tendències i millorar la protecció de les comunitats i la seguretat pública.

Per tal de poder dur a terme el nostre projecte, hem agafat el dataset més actual de la *NFIRS* que és el que conté tots els incidents del 2023.

3.2 Estructura de la NFIRS

La *NFIRS* és un sistema modular que es compon d'11 mòduls[7][6]. El mòdul bàsic, d'incendi, incendi en estructures, víctimes civils i víctimes del cos de Bombers són obligatoris per a tots els incidents, mentre que la resta són opcionals i s'omplen en funció del tipus d'incident. Cada mòdul té taules de dades interrelacionades que s'uneixen mitjançant

identificadors únics. A continuació es descriuran tots els mòduls que conformen la Base de Dades.

Nom Mòdul	**Descripció**	**Arxius que conté**
Bàsic	conté tota la informació general de cada incident	basicaid.txt basicincident.txt incidentaddress.txt fdheader.txt
Incendi	Descriu els incidents de tipus incendis. En cas d'incendis forestals, es pot utilitzar un mòdul més especialitzat.	fireincident.txt
Incendis en Estructures	Descriu tots els incendis en estructures. Aquest mòdul és usat juntament amb el mòdul Incendi.	No conté cap fitxer directe
Víctimes Civils	Reporta les lesions i morts dels civils o altres persones del personal d'emergències (com podrien ser policies o personal d'EMS) que estan relacionats amb l'incident.	civiliancasualty.txt
Víctimes de Cos de Bombers	Reporta les lesions i morts del personal del cos de Bombers. Aquest mòdul també pot ser utilitzat per reportar l'exposició de bombers a components químics o biològics en un incident.	ffcasualty.txt ffequipfail.txt
EMS (Serveis Mèdics d'Emergències)	Conté les dades dels pacients i els serveis prestats per part del personal d'urgència mèdica.	ems.txt
Materials perillosos	Per incidents on hi ha hagut vessaments o emissions de 55 galons o més de materials perillosos. Quan és necessari s'usa juntament amb el mòdul Incendi.	hazchem.txt hazmat.txt hazmatequipinvolved.txt hazmobprop.txt
Incendis forestals	Reporta incendis forestals.	wildlands.txt
Aparells o Recursos	Reporta les dades específiques de cada aparell que respon a un incident. Aquest mòdul no s'utilitza quan es fa servir el mòdul Personal.	No conté cap fitxer directe
Personal	Agrupa la mateixa informació que el mòdul Aparells o Recursos, però també inclou el personal associat a l'aparell.	No conté cap fitxer directe

Incendis Intencionats	Reporta la informació addicional de tots aquells incendis que han estat classificats com ïncendis intencionats"o amb l'opció de ser-ho. El departament d'Arson és qui s'encarrega d'estudiar aquests casos.	arson.txt arsonagencyreferral.txt arsonjuvsub.txt

Hi ha un fitxer, `codelookup.txt`, que proporciona els codis utilitzats a la *NFIRS*, amb les seves definicions per facilitar la interpretació de les dades.

3.3 Fitxers necessaris i els seus Atributs més importants

Com ja hem vist, la base *NFIRS* té molta informació repartida en diferents fitxers. Per tal de poder dur a terme el nostre objectiu, no necessitem tots els fitxers. Anem a veure quins fitxers necessitarem pel nostre treball.

`basicincident.txt`: Conté informació important sobre el tipus de propietat que ens permetrà filtrar per quedar-nos amb els habitatges. Té els següents atributs:

- `INCIDENT_KEY` (str, clau primària): Identificador únic de l'incident.
- `STATE` (str, clau primària): Estat dels Estats Units on es va reportar l'incident.
- `FDID` (str, clau primària): Identificador del parc de Bombers.
- `INC_DATE` (int, clau primària): Data de l'incident.
- `INC_NO` (str, clau primària): Número d'incident.
- `EXP_NO` (int, clau primària): Número d'exposició. L'exposició es defineix com un incendi resultant d'un altre incendi fora del lloc d'origen. Per exemple, si el foc de l'edifici encén un camió aparcat a l'exterior, el foc del camió és un incendi d'exposició.
- `INC_TYPE` (category): Número que descriu el tipus d'incident que s'ha atès. Pot ser incendi, rescat, materials perillosos, servei, bones intensions, alarma falsa, desastre natural o incident especial. Dintre de cada una d'aquestes categories s'explica amb més detall l'incident. Per exemple, els incidents de tipus incendis són els números entre 100 i 200. Un incendi que està relacionat amb cuinar és el número específic 113.
- `PROP_LOSS` (float): Estimació de la pèrdua total en dòlars de la propietat. A nivell d'assegurances es coneix com a pèrdua de continent.
- `CONT_LOSS` (float): Estimació de la pèrdua total en dòlars del contingut.

- PROP_USE (category): Indica l'ús específic de cada propietat. Igual que a INC_TYPE té varies grans categories, però també s'especifica amb més detall el tipus en concret.
- CENSUS (str): Número de sis dígits que identifica una àrea de terra dins dels Estats Units. No totes les jurisdiccions tenen números de seccions censals.
- DET_ALERT (category): Indica si els detectors instal·lats van avisar als ocupants de la propietat o no.
- HAZ_REL (category): Indica si en l'incident hi va haver alliberament de material perillós. I en cas afirmatiu de quin material: gas natural, gasolina, propà, etc.
- MIXED_USE (category): En el cas que la propietat tingui més d'un ús, aquest paràmetre s'utilitza per especificar els diferents usos.

incidentaddress.txt ens permet analitzar patrons de geografia. Ens funcionarà com a enllaç per afegir factors socioeconòmics. Té els següents atributs:

- INCIDENT_KEY, STATE, FDID, INC_DATE, INC_NO i EXP_NO: Funcionen com a clau primària.
- LOC_TYPE (category): Diu on s'ha ocasionat l'incident. Pot ser un carrer, una intersecció, una direcció o altres.
- CITY (str): Ciutat on es va produir l'incident.
- ZIP5 (str): El codi numèric assignat pel servei postal dels Estats Units a totes les jurisdiccions dels EUA.
- NUM_MILE (str): Número de la ubicació concreta on s'ha produït l'incident.
- STREETNAME (str): El nom del carrer on es va produir l'incident.
- STREETTYPE (category): Descriu el tipus de carrer que apareix després d'un nom de carrer. Pot ser travessa, camp, pont, avinguda, etc.
- STREET_PRE (category): Descriptor direccional que apareix abans d'un nom de carrer. Pot ser Est, Nord, Nordest, etc.
- STREETSUF (category): Descriptor direccional que apareix després d'un nom de carrer. Pot ser Est, Nord, Nordest, etc.
- APT_NO (str): El número de l'apartament en què s'ha produït l'incident.

També necessitarem el fitxer fireincident.txt perquè és crucial per acabar classificant els incendis en funció de les causes. Com ara fallades elèctriques i altres. Els paràmetres que té són els següents:

- INCIDENT_KEY, STATE, FDID, INC_DATE, INC_NO i EXP_NO: Funcionen com a clau primària.

- CAUSE_IGN (category): El factor causal general que va provocar que una font de calor encenés un material combustible. La causa podria ser el resultat d'un acte de delinqüència, una fallada mecànica o un acte natural.

- FACT_IGN_1 (category): Factor contribuent que va permetre que la font de calor i el material combustible es combinessin per encendre el foc. Igual que en altres paràmetres, té diverses grans categories, però també especifica amb més detall el factor contribuent.

- FACT_IGN_2 (category): Un altre factor contribuent que va permetre que la font de calor i el material combustible es combinessin per encendre el foc.

- HEAT_SOURC (category): La font específica de calor que va provocar el foc. Igual que en altres paràmetres, té diverses grans categories, però també especifica amb més detall el factor contribuent.

- FIRST_IGN (category): Primer objecte que s'ha encès degut a la font de calor. Variable categòrica.

- TYPE_MAT (category): La composició del material del primer objecte en incendiar-se. Pot ser un gas, un líquid inflamable, un producte químic, un plàstic, etc.

- ITEM_SPRD (category): L'objecte que més contribueix a la propagació del foc, si és diferent de l'objecte que es va encendre primer.

- MAT_SPRD (category): Tipus de material que més contribueix a la propagació de la flama, si es diferent del tipus de material que es va encendre inicialment.

- FIRE_SPRD (category): Fins a quin punt es va arribar a estendre els danys de la flama del foc. És a dir, fins a on realment es va cremar. Pot ser que l'incendi es quedés limitat a l'objecte inicialment encès, a l'habitació, la planta, tot l'edifici o més.

- STRUC_TYPE (category): Identifica una estructura com a tipus de propietat específic.

- DETECTOR (category): Ens indica l'existència d'equips de detecció d'incendis.

- DET_TYPE (category): Identifica el tipus de sistema de detecció d'incendis qui hi havia a la zona on s'ha originat l'incendi.

- DET_OPERAT (category): Ens indica el funcionament del detector en relació amb l'incendi.

- DET_EFFECT (category): L'eficiència dels equips de detecció d'incendis per alertar els ocupants i si aquests van respondre o no.

- DET_FAIL (category): El motiu pel qual el detector no va funcionar o no va funcionar correctament.

- AES_PRES (category): Ens indica l'existència d'un sistema d'extinció automàtica (AES).

- AES_OPER (category): Ens indica el funcionament del sistema d'extinció automàtica (AES) en relació amb l'incendi.
- AES_FAIL (category): El motiu pel qual el sistema d'extinció automàtica (AES) no va funcionar o no va funcionar correctament.
- BLDG_INVOL (float): Número total d'edificis involucrats en l'incendi.
- AREA_ORIG (category): L'ús principal de la zona on es va iniciar el foc. Pot ser una habitació, un vehicle, etc.Igual que en altres paràmetres, té diverses grans categories, però també especifica amb més detall on es va originar l'incendi.
- HUM_FAC_1 (category): Variable categòrica que indica la situació humana que va permetre que la font de calor i el material combustible es combinessin per iniciar l'incendi. Pot ser perquè estigues dormit, sota els afectes de l'alcohol, persona mentalment inestable, etc.
- TOT_SQ_FT (float): La mida de la planta principal en peus quadrats. És una estimació.

Per tal de poder classificar els riscos d'incendis avaluant l'impacte humà en vers a víctimes i el grau de les lesions necessitem el fitxer civiliancasualty.txt. Aquests són els atributs que té:

- INCIDENT_KEY, STATE, FDID, INC_DATE, INC_NO i EXP_NO: Funcionen com a clau primària.
- GENDER (category): El gènere de la víctima.
- AGE (int): L'edat de la víctima en anys o, si la víctima és un nadó, en mesos.
- RACE (category): Identificació de la raça de la víctima. Pot ser blanc, negre o afroamericà, indis americans o nadius d'Alaska, asiàtic, nadius hawaians, altres o indeterminat.
- ETHNICITY (category): Identifica l'ètnia de la víctima. L'ètnia designa un subgrup de la població que té un patrimoni cultural comú.
- SEV (category): Indica la gravetat de la lesió en una escala de menys greu a més greu fins a arribar a la mort.
- CAUSE_INJ (category): L'esdeveniment físic que va causa la lesió.
- PRIM_SYMP (category): La lesió aparent més greu de la víctima.
- BODY_PART (category): Indica la part del cos que ha patit la lesió més greu.
- HUM_FACT1 (category): L'estat físic o mental de la persona abans de convertir-se en víctima. Pot ser adormida, sota els afectes de l'alcohol, discapacitat física o mental, etc.

- `FACT_INJ1` (category): Els valors més significatius que contribueixen a la lesió de la víctima.
- `LOC_INC` (category): Localització de la víctima en relació a la zona d'origen del foc en el moment de l'inici del foc.
- `GEN_LOC_IN` (category): Localització de la víctima en relació a la zona d'origen del foc en el moment en què es va produir la lesió.
- `STORY_INC` (float): Identifica l'apartament on es trobava la víctima a l'inici de l'incident.
- `STORY_INJ` (flaot): Identifica l'apartament on es trobava la víctima quan es va produir la lesió.
- `SPC_LOC_IN` (category): Identifica la ubicació específica de la víctima en el moment de la lesió.

Finalment, necessitarem el fitxer `ACSST5Y2023.S1901-Data.csv`. Aquest fitxer no depèn de la base de dades *NFIRS* sinó de la *Census Bureau*[2]. Conté informació relacionada amb els ingressos en dòlars del 2023 ajustats a la inflació. Aquest fitxer té molts paràmetres, però només en necessitem dos pel que volem:

- `GEO_ID` o `Geography` (str): Codi numèric que identifica de manera única totes les àrees geogràfiques. Aquest número inclou el codi postal *ZIP5*.
- `S1901_C01_012E` o `Estimate!!Households!!Median income (dollars)` (float): Mediana de l'estimació dels ingressos anuals en dòlars de cada `GEO_ID`.

Capítol 4

Pre-Processament de Dades

4.1 Neteja Bàsica

Abans de poder entrenar cap model és necessari fer una manipulació del conjunt de dades perquè permeti i faciliti tots els processos que li aplicarem. Comencem analitzant els paràmetres de `basicincident.txt`.

Com que la nostra Bade de Dades inclou altres incidents que no són incendis, el primer que hem fet ha sigut eliminar tots aquests incidents que no són incendis. Els incidents de tipus incendi són tots aquells on `INC_TYPE` pren valors entre el 100 i 199.

Seleccionem les propietats que siguin habitatges. Això vol dir que ens hem de quedar amb els valors on `PROP_USE` sigui 400, 419, 429, 439, 449, 459, 460, 462 o 464.

Ara que ja tenim la nostra base de dades amb els requisits que necessitem, mirem com tenim repartits els valors NULLs. Hi ha algun valor NULL al paràmetre `STATE`. Com que aquest paràmetre és una clau primària, eliminem els registres amb valor NULL. Hi ha més variables que tenen valors NULLs, però els gestionarem més endavant.

Passem a analitzar el fitxer `incidentaddress.txt`. Mirem els valors NULLs i, igual que abans, esborrem els registres on `STATE` és NULL. Fem el mateix per `CITY` i `ZIP5`.

Repetim el mateix procés pel fitxer `fireincident.txt`.

Finalment, mirem el fitxer `civiliancasualty.txt`. D'aquest fitxer ens interessa saber les víctimes totals que hi ha a cada incendi i quantes pels diferents graus de lesió. Per tal d'aconseguir la informació que necessitem haurem d'agrupar les dades per les claus primàries i contar el nombre de persones per cada lesió.

Per últim, ajuntem les dades dels quatre fitxers amb l'ajut de les claus primàries.

4.2 Gestió de Valors NULLs

Una vegada que ja tenim totes les dades juntes, podem mirar com gestionar els valors NULLs. Abans però, esborrem els registres duplicats. Per gestionar els valors NULLs, mirem el percentatge dels valors NULLs per cada variable.

Tenim forces valors NULLs amb les paràmetres relacionats amb el número de víctimes i lesionats. Ho omplim amb el valor 0 perquè les dades que ens falten són perquè no es van agafar en no haver-hi cap víctima.

Igual que abans, esborrem els registres amb valors NULLs dels paràmetres `CITY`, `ZIP5`, `FIRE_SPRD`, `PROP_LOSS` i `CONT_LOSS`.

El percentatge de valors NULLs pels paràmetres `HAZ_REL` i `MIXED_USE` és molt alt. Eliminem aquestes variables.

Veiem que tenim dos paràmetres amb informació molt similar: `DETECTOR` i `DET_ALERT`. Aquestes dues variables prenen com a valors afirmatiu, negatiu o indeterminat. Per tal de decidir amb quina variable ens quedem, mirem quina conté més informació qualitativa sumant la quantitat de dades que tenen amb valor afirmatiu i negatiu. Fent això veiem que `DETECTOR` té 94.264 registres i `DET_ALERT` 32.530. Així que ens quedem amb la primera variable i eliminem la segona.

Veiem que tenim molts paràmetres que tenen 'Indeterminat' com a valor. Aquesta dada no ens dona gaire informació així que eliminem els registres amb valors NULLs i 'Indeterminat' de les variables `FACT_IGN_1` i `AES_PRES`.

Pel paràmetre `TYPE_MAT`, substituïm els valors NULLs per la categoria *UU*.

Mirant amb detall els valors de `ZIP5` veiem que hi ha registres amb valors *0* i *00000*. Aquests valors no pertanyen a cap codi postal, així que optem per eliminar els registres amb aquests valors.

4.3 Gestió de Valors Atípics

Les variables `PROP_LOSS` i `CONT_LOSS` tenen varis valors atípics. Aquests valors els tractarem tenint en compte el teorema de Chebyshev. Calculem un llindar superior i inferior: $mean \pm 3 \cdot std$. Els valors atípics són els que queden fora d'aquests llindars i els eliminarem.

4.4 Conversió de Dades

Ara que ja tenim la gestió dels valors NULLs fets, passem a fer la conversió de les dades. La variable `INC_DATE` la convertirem a *datetime*. Els paràmetres `DETECTOR` i `AES_PRES` les convertirem a booleanes. La resta de variables són categòriques i les hem de tractar de manera específica.

Com hem dit anteriorment, moltes de les variables categòriques que tenim són números on està especificat ja sigui el lloc on s'ha produït l'incendi, l'objecte en incendiar-se primer o altres. Però aquestes especificacions es poden ajuntar per categories. Així que `AREA_ORIG`, `HEAT_SOURC`, `FIRST_IGN`, `TYPE_MAT` i `FACT_IGN_1` agruparem els seus valors segons les categories de cadascun d'elles.

4.5 Creació de Noves Variables

Creem una nova variable que ens indiqui el mes en què es va generar l'incendi per veure si està relacionat amb el grau de severitat d'aquest. Però la simplificarem convertint-la en una variable categòrica que indiqui l'estació en què es va produir l'incendi. Hivern, primavera, estiu o tardor.

També crearem una altra variable que sigui la suma de `PROP_LOSS` i `CONT_LOSS`. D'aquesta manera podrem obtenir les pèrdues totals que ha causat l'incendi.

En el cas de la variable `ZIP5`, és molt difícil tractar-la en un model perquè tot i ser de tipus numèrics no es poden tractar com a tal perquè no hi ha relació entre ells. Per aquest motiu, i també amb l'objectiu de veure quin efecte socioeconòmic hi ha, hem decidit agafar la mediana de l'estimació dels ingressos anuals corresponents a cada codi postal. Aquesta nova variable que associa el `ZIP5` amb la mediana d'ingressos anual li hem dit `MEAN_INCOME`.

4.6 Normalització de les Dades

Hem fet una transformació logarítmica a les variables numèriques `TOTAL_LOSS` i `MEAN_INCOME` per tal d'aconseguir que segueixin una distribució Normal.

Ara ja tindríem la base de les nostres dades per entrenar-les més tard. La Base de Dades que ens ha quedat té 38 paràmetres. Dels quals 2 són booleans, 12 floats, 8 ints i 16 categòriques.

4.7 One-Hot Encoding

La majoria dels models, no saben gestionar de manera correcta les variables categòriques, per aquest motiu necessitem convertir-les en booleanes. Per fer això, hem de crear una columna nova per cada categoria de cada variable. Com que hi ha alguns paràmetres amb moltes categories i això faria que acabéssim tenint una base de dades amb un número molt elevat de columnes, hem optat per mirar quants registres té cada categoria de cada paràmetre, quedar-nos amb les que tenen més pes i les altres agrupar-les en una única categoria *Altres*.

Per la variable `AREA_ORIG_CATEGORY` ens hem quedat amb *Function Areas*, *Strucutral Areas*, *Storage Areas*, *Assembly*, *Outside Areas*, *UU* (Indeterminat) i *Other*.

Pel paràmetre `HEAT_SOURC_CATEGORY` amb *Operating Equipment*, *Open Flame or Smoking*, *Hot Objects*, *UU* i *Other*.

Per `FIRST_IGN_CATEGORY` amb *Structural*, *General Materials*, *Organic Materials*, *Furniture*, *Soft Goods and Wearing Apparel*, *UU* i *Other*.

Per `TYPE_MAT_CATEGORY` amb *Wood*, *Textiles*, *Plastics*, *Combustible Liquid*, *UU* i *Other*.

Per `FACT_IGN_1_CATEGORY` amb *Misue Material*, *Electrical Failure*, *Operational Deficiency*, *NN* (cap) i *Other*.

Per la variable `STRUC_TYPE` ens hem quedat amb les de tipus *1* i *2* i les altres les hem posat de tipus *0*.

Per `PROP_USE` ens hem quedat amb les propietat que són *419*, *429*, *439*, *449*, *459*, *460* i les altres les hem etiquetat a *400*.

Per últim, de `INC_TYPE` ens hem quedat amb *111*, *113*, *114*, *116*, *118*, *120*, *121*, *122*, *123* i les altres són *100*.

Ara ja podem convertir les variables a booleanes amb la funció `get_dummies`.

4.8 Eliminació de Variables

Ara que ja tenim una nova columna per cada una de les categories amb les quals ens hem quedat per cada variable, ja podem esborrar la categòrica original. A més, també podem eliminar altres variables com claus primàries que ja no ens aporten informació o altres variables que les hem gestionat d'una altra manera. Finalment, ens queda una base de dades amb 66 variables.

4.9 Gestió del desequilibri de les classes

Ara que ja tenim la nostra base de dades mirem com ens queden repartides les dades segons la nostra variable objectiu `FIRE_SPRD`. Aquesta variable consta de 5 classes: la classe *1* ens indica que l'incendi es va quedar limitat a l'objecte inicial de l'incendi, la classe *2* que es va quedar limitat a l'habitació, la *3* a la planta, la *4* a l'edifici i la *5* més enllà de l'edifici. El motiu pel qual hem escollit aquesta variable com l'objectiu és perquè ens permet veure fàcilment la severitat de l'incendi.

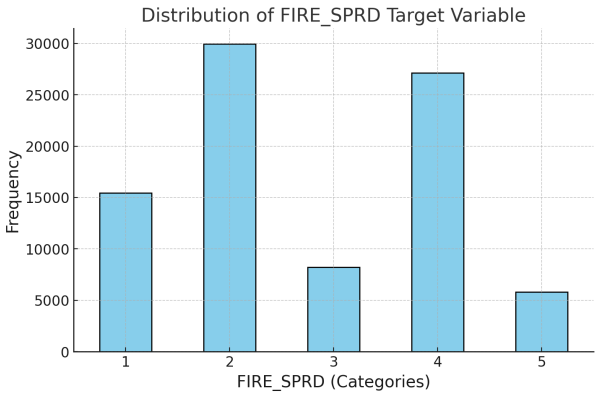

Figura 4.1: Distribució Variable Objectiu `FIRE_SPRD`

A partir de la imatge 4.1 podem veure que la distribució de les classes estan molt desequilibrades. Tenim moltes poques dades de la classe 5. Com que tant la classe 4 com la 5 indiquen que la severitat de l'incendi és elevada, podem ajuntar aquestes dues classes. De la classe 3 també tenim molt poques dades així que optem per eliminar-la.

Un altre canvi que fem és modificar el nom de les etiquetes de les classes perquè comenci per 0. Això ens evitarà problemes futurs a l'hora d'entrenar amb certs algoritmes. En la Taula 4.1 podem veure com ha quedat finalment la nostre variable objectiu `FIRE_SPRD`.

Classe final del dataset	Classe original del dataset	Significat
0	*1*	Incendi confinat a l'objecte
1	*2*	Incendi confinat a l'habitació
2	*4* i *5*	Incendi a tot l'edifici o més

Taula 4.1: Resum de les classes finals de la variable objectiu `FIRE_SPRD`

La variable de sortida dels nostres models serà *0*, *1* o *2*. Classificar el grau de gravetat de l'incendi.

Tot i aquests canvis, les classes segueixen sense estar equilibrades. Per tal d'aconseguir-ho farem un *undersampling*. Consisteix en quedar-nos amb el nombre de registres que té la classe amb menys dades. La resta de les classes s'agafa de forma aleatòria els registres.

Finalment, tenim una base de dades amb 38.877 files i 66 columnes.

Capítol 5

Anàlisi numèric

En aquest apartat, anem a analitzar la Base de Dades un cop realitzat el pre-processament explicat en el capítol 3. Intentarem comprendre el nostre conjunt de dades mitjançant representacions gràfiques. Mirarem les distribucions de les diferents variables i aspectes més específics com on és més comú que comenci l'incendi, quins factors són més contribuents als incendis i altres. També examinarem la relació entre els factors socioeconòmics i la gravetat de l'incendi.

Comencem analitzant quina importància pot tenir el fet de tenir detectors de fum a casa. En la Imatge 5.1 podem veure com aquesta variable no està equilibrada perquè tenim gairebé el doble de mostres on si hi han detectors de fum.

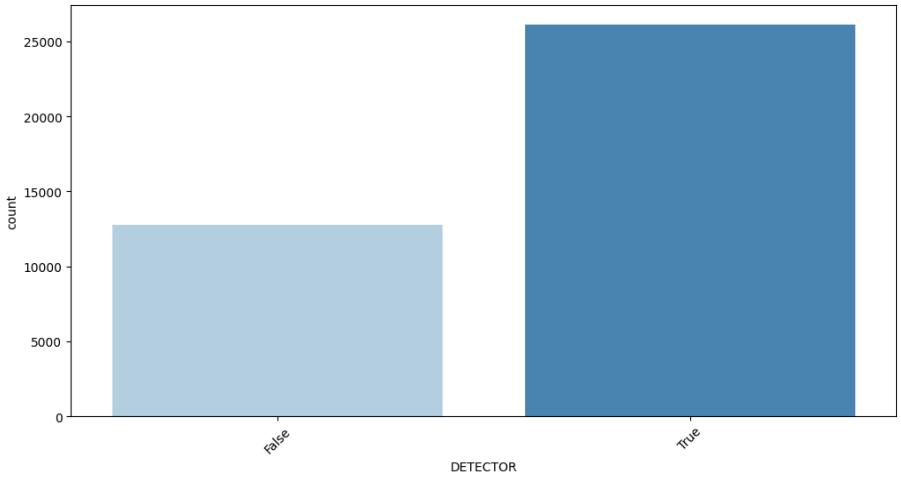

Figura 5.1: Distribució de les mostres de la variable `DETECTOR`

Com que la variable `DETECTOR` no està equilibrada, cal analitzar les mostres de forma independent. Quan la presència de detectors de fum és certa, a partir de la Imatge 5.2, podem veure com per la classe 0 i 1 hi ha gairebé les mateixa quantitat de mostres però que per la classe 2 disminueix notablement. En canvi, quan no hi ha presència de detectors de fum, el comportament de la variable és completament l'oposat. Això ens pot portar a pensar que el fet de tenir detectors de fum instal·lats a casa pot disminuir el risc de tenir un incendi amb el grau de severitat alt, mentre que el fet de no tenir-ne augmenta aquest risc. Això podria ser perquè el fet de no tenir detectors de fum fa que la presència de

l'incendi es notifiqui més tard i com a conseqüent s'acabi alimentant més.

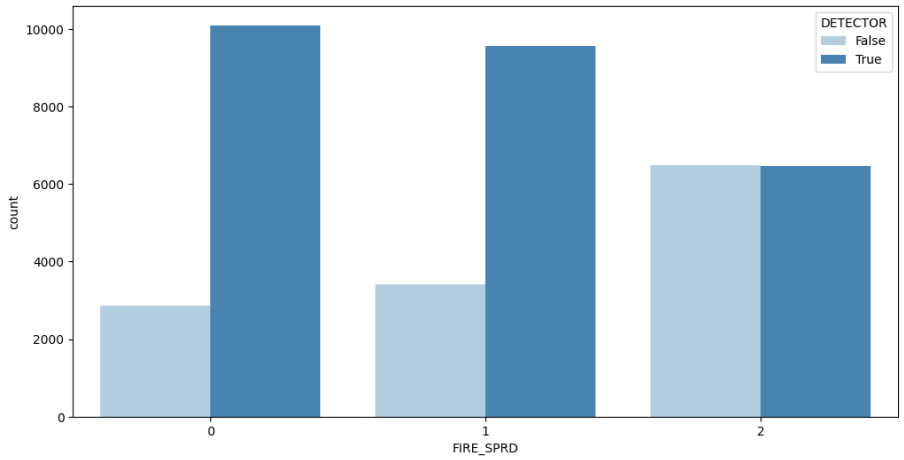

Figura 5.2: Distribució de la presència o no presència de detectors de fum per cada grau de perillositat de l'incendi

Passem a analitzar com afecta la presència de sistemes d'extinció automàtics. Ràpidament, mirant la Imatge 5.3 veiem que el nostre conjunt de dades té moltes més mostres sense presència de AES que amb si. Això és normal ja aquests mecanismes no són gaire habituals tenir-los en habitatges. Com que tornem a tenir una variable desequilibrada, tornem a analitzar les mostres de manera independent. Veiem com el número de mostres segons la categoria no varia quan no tenim AES. Però en canvi, quan si hi ha AES, veiem que hi ha molts menys incendis que acabin evolucionant fins el grau de gravetat més alt.

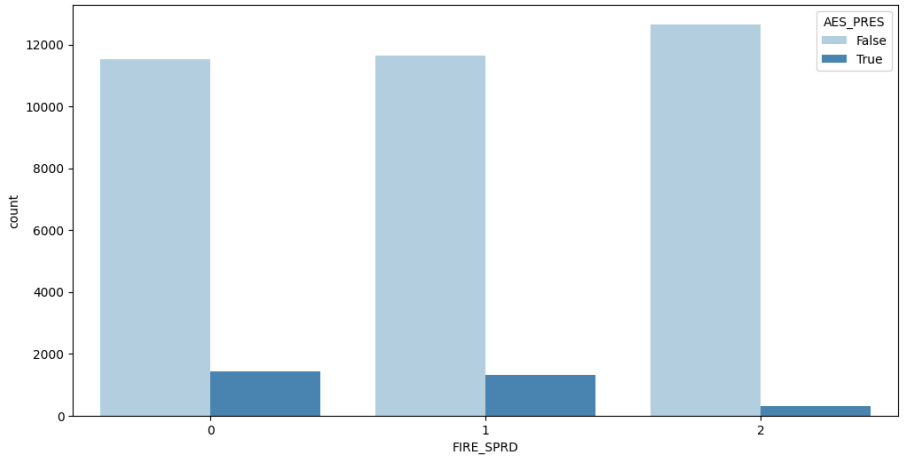

Figura 5.3: Distribució de la presència o no presència de AES per cada grau de perillositat de l'incendi

Durant el pre-processament, hem retallat la Base de dades per només quedar-nos amb els incidents de tipus incendi. Anem a veure com aquestes estan repartides. Mirant la Imatge 5.4 veiem que gairebé tots els incendis són de tipus 111, que són incendis d'edifici i el segon tipus, el 113, són incendis de cuina on el foc s'ha quedat al recipient cremat.

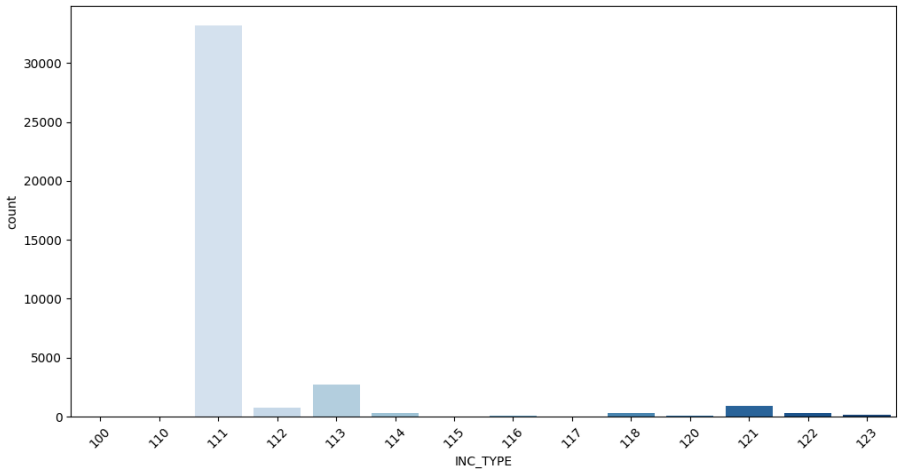

Figura 5.4: Distribució de la variable `INC_TYPE`

Ara passem a analitzar les característiques més centrades amb informació respecte a l'incendi. Examinem a quin lloc es va iniciar l'incendi. A la Imatge 5.5 podem observar que la categoria més dominant és *Function Areas*. Aquesta categoria fa referència a les zones més utilitzades dels habitatges. Com el dormitori, el menjador, la cuina i l'estudi. El lavabo i el safareig també s'inclourien en aquesta categoria. La segona categoria amb més pes és *Structural Areas* que seria la paret, taulada, balcó, etc. I la tercera són llocs d'emmagatzematge com un armari o el garatge.

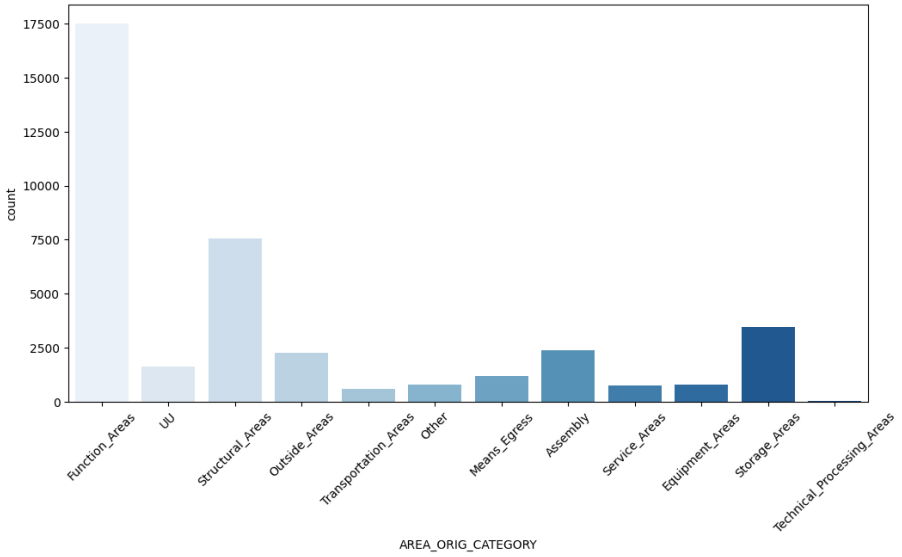

Figura 5.5: Distribució de la variable `AREA_ORIG`

Al mirar quin impacte tenen les categories per cada grau d'extensió del foc, Imatge 5.6, veiem que els incendis d'estructura acostumen a comportar un grau de perillositat alt. Això és a causa de la dificultat d'extinció que sovint pot haver-hi en incendis que es produeixen entre parets o sostres.

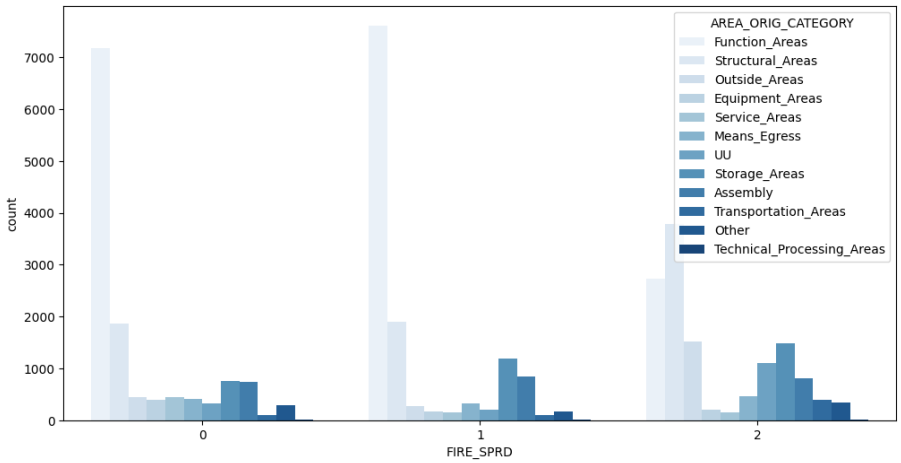

Figura 5.6: Distribució de les diferents àrees de l'origen del foc segons la gravetat de l'incendi

A partir de la Imatge 5.8, podem veure com la majoria dels incendis de la nostra Base de Dades són causats degut a una espurna expulsada per algun aparell electrodomèstic. Altres causes freqüents són objectes calents com la brasa o cendra calenta, materials amb flama oberta com un cigar o espelmes.

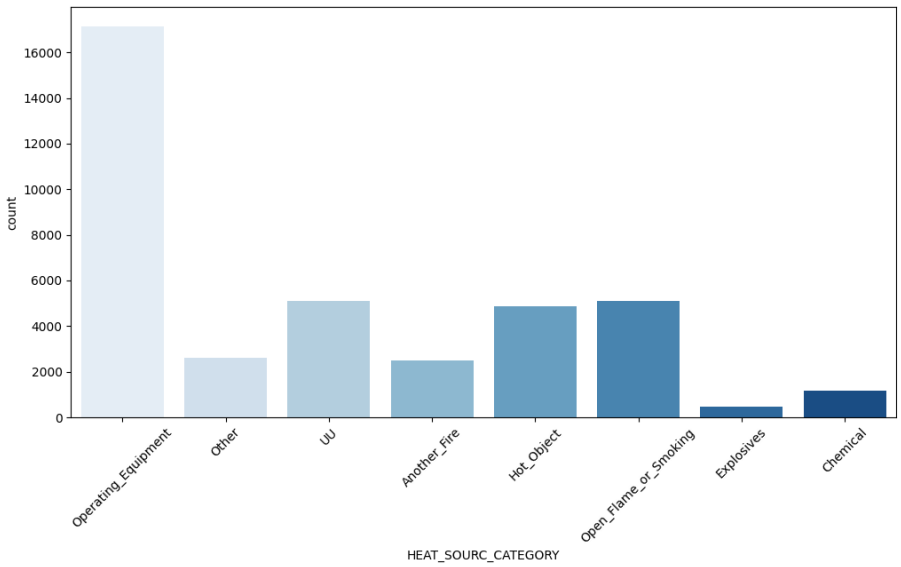

Figura 5.7: Distribució de la variable `HEAT_SOURC`

De la Imatge 5.8, podem extreure que en general la causa de l'incendi no determina la gravetat d'aquest.

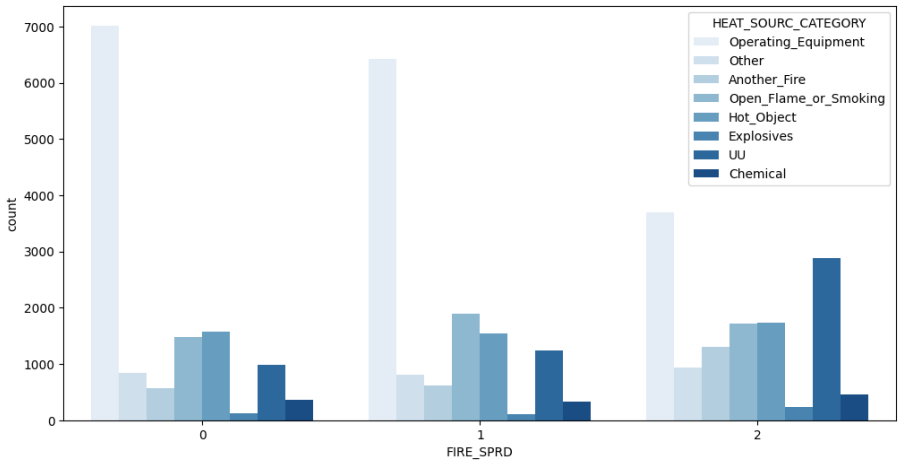

Figura 5.8: Distribució de les diferents fonts de calor causants de l'incendi foc segons la gravetat del mateix

Anem a mirar si el primer objecte en incendiar-se té alguna relació amb la gravetat de l'incendi. A través de la Imatge 5.9, veiem que els objectes que s'incendien els primers són de les categories estructura, materials orgànics com podria ser el menjar, mobles i materials generals com llibres o cables elèctrics.

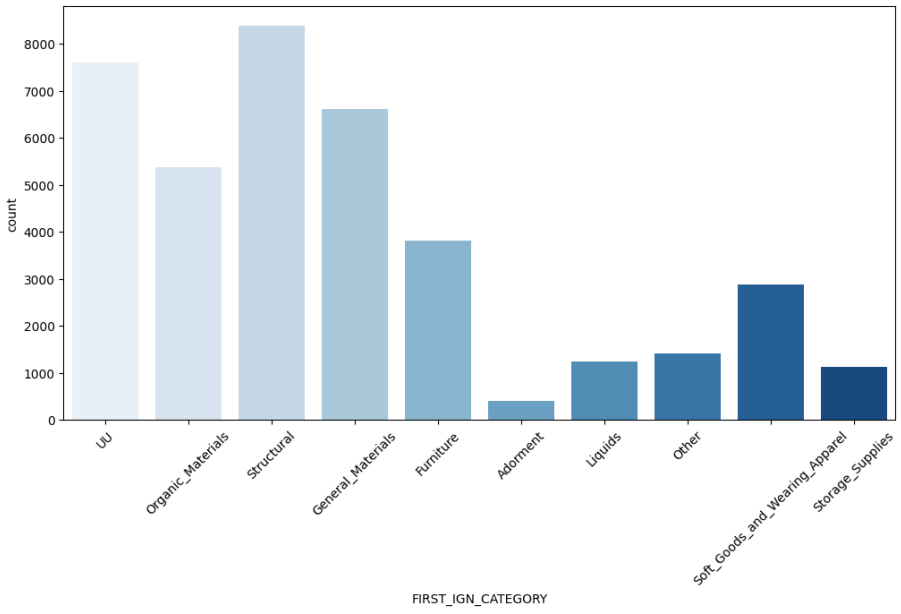

Figura 5.9: Distribució de la variable `FIRST_IGN`

La Imatge 5.10 ens remarca igual que la variable `AREA_ORIG` que els incendis estructurals, tenen un grau de perillositat molt elevat.

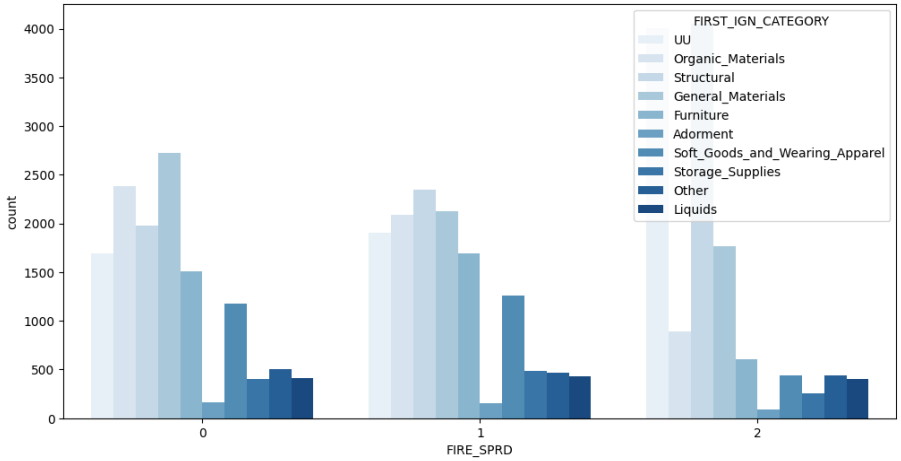

Figura 5.10: Distribució dels diferents objectes que van ser els primers en incendiar-se segons la gravetat de l'incendi

Ara que ja sabem quins objectes acostumen a ser els primers en incendiar-se, vegem quins són els materials més comuns d'aquests. Amb la Imatge 5.11, veiem que tenim molts objectes dels quals no en sabem el material. Tot i així, anem a mirar quina importància tenen la resta dels materials que coneixem. Els materials que s'han encès més inicialment són de fusta, tèxtils o plàstics.

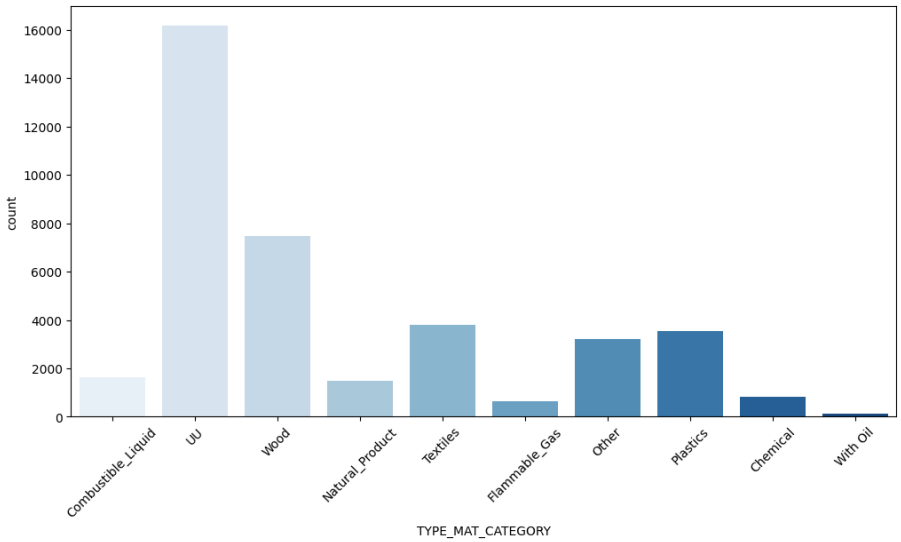

Figura 5.11: Distribució de la variable `TYPE_MAT`

En la Imatge 5.12, veiem com la fusta pot provocar incendis més greus. Això pot ser degut a que la fusta és un bon combustible.

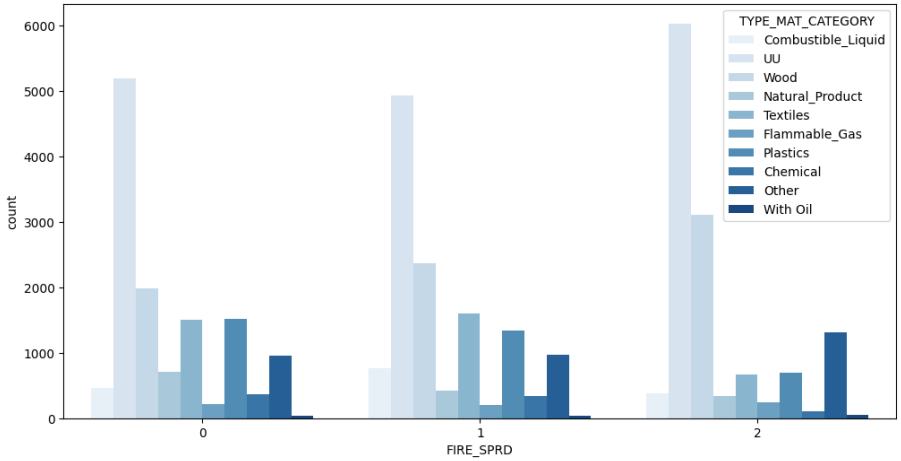

Figura 5.12: Distribució dels diferents materials objectes que van ser els primers en incendiar-se segons la gravetat de l'incendi

Passem a analitzar quins factors van permetre que s'originés l'incendi. A partir de la Imatge 5.13 extraiem que gran part dels incendis són causats com a conseqüència d'un mal ús de material. Ja sigui per material oblidat com una espelma, perquè la font de calor estava massa prop d'un possible combustible o perquè s'estava jugant amb foc.

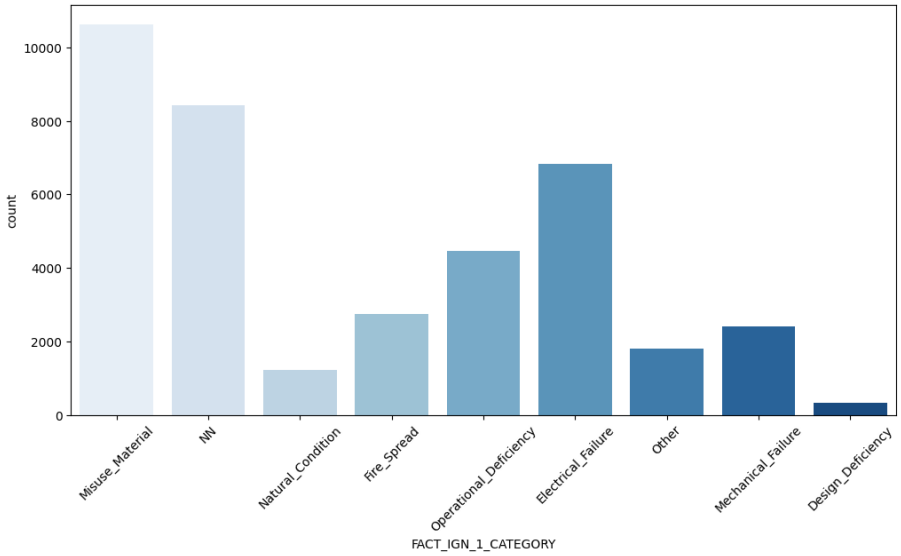

Figura 5.13: Distribució de la variable `FACT_IGN`

Mirant la Imatge 5.14, podem arribar a la conclusió que la causa de l'incendi no està relacionada amb el comportament d'aquest.

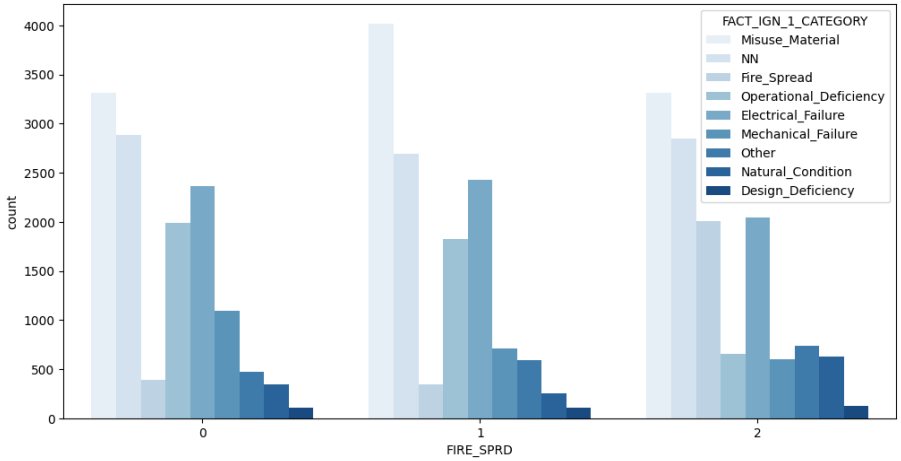

Figura 5.14: Distribució dels factors contribuents a originar l'incendi segons la gravetat d'aquest

Un cop fet l'anàlisi de la majoria de les variables categòriques, passem a les numèriques. Analitzem quina relació hi ha entre les pèrdues que generen els incendis i la severitat d'aquest. Mirant la Imatge 5.15 veiem que quan més gran és el grau de perillositat de l'incendi, més alt són les pèrdues. Això pot ser perquè quan més gran és la zona afectada per l'incendi, més zones cremades hi ha i com a conseqüent hi ha més pèrdues.

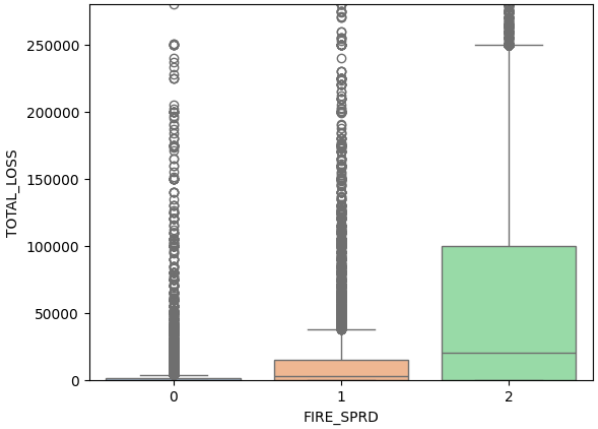

Figura 5.15: Distribució de les pèrdues totals segons la gravetat de l'incendi

A partir de la Imatge 5.16, veiem que a mesura que la severitat de l'incendi augmenta, la mediana dels ingressos anuals es va reduint. Això pot ser per diversos motius. Un motiu podria ser que els materials que hi ha a les cases amb ingressos anuals més baixos són menys resistents al foc i per això l'incendi acaba sent de gravetat elevada.

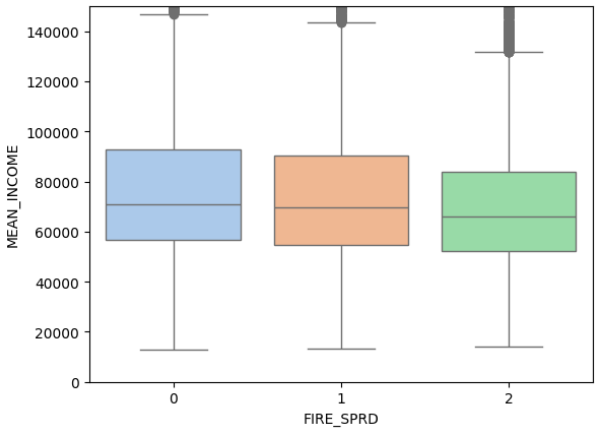

Figura 5.16: Distribució dels ingressos anual segons la gravetat de l'incendi

A la Imatge 5.17, veiem com en incendis de grau lleu i mitja el número de mostres entre ingressos baixos i elevats estan igualats. En canvi, hi ha molts més incendis de grau alt per ingressos baixos.

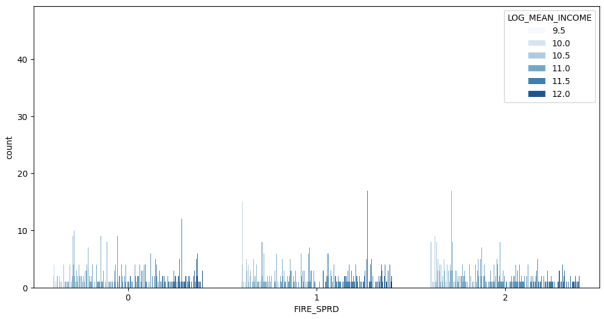

Figura 5.17: Distribució de la mediana dels ingressos anual segons la gravetat de l'incendi

Mirant la Imatge 5.18, veiem que la mitjana per els diferents graus de perillositat d'un incendi la trobem a 0. Això és perquè en la majoria d'incendis no hi ha hagut ferits. Però per altre banda, observem que a mesura que augmenta la severitat de l'incendi també ho fan el nombre de víctimes en casos on si que hi han hagut lesionats.

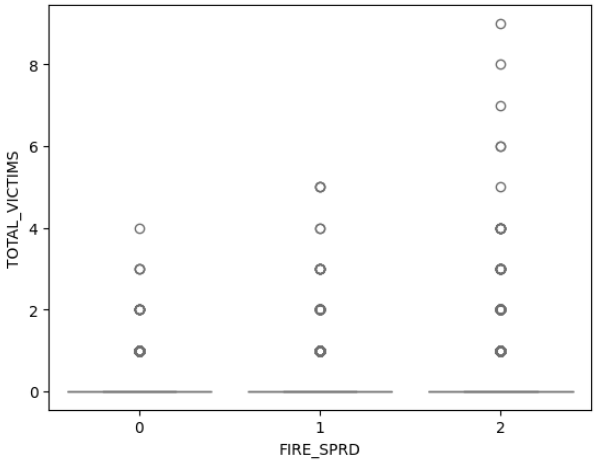

Figura 5.18: Distribució de les víctimes totals en l'incendi segons la gravetat d'aquest

Per últim, mirem quina correlació hi ha entre les diferents variables. A partir de la Imatge 5.19, veiem que hi ha correlacions molt elevades prop de la diagonal. Això és a causa de les variables categòriques. Les diferents variables que hem agrupat per categories estan relacionades amb les altres categories que tenia el paràmetre principal. Si ens fixem amb la variable objectiu `FIRE_SPRD`, veiem que les variables més relacionades amb questa són `DETECTOR` amb una correlació de -0.24, `TOTAL_LOSS` amb 0.34, `AREA_ORIG_Funtion_Areas` amb -0.28, `HEAT_SOURC_Operating_Equipment` amb -0.21, `INC_TYPE_111` amb 0.22 i `INC_TYPE_113` amb -0.26.

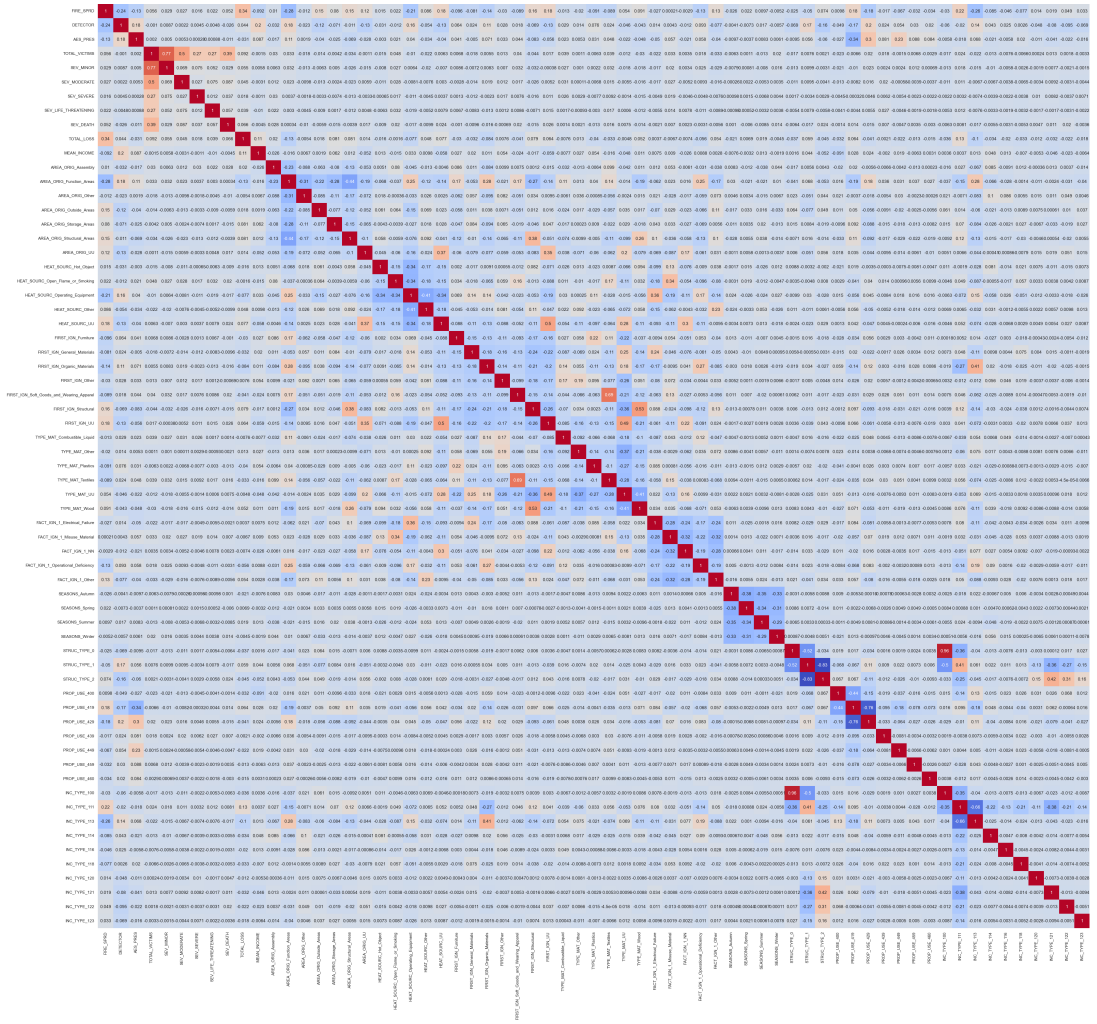

Figura 5.19: Matriu de Correlació de les diferents variables

Capítol 6

Entrenament de Models

6.1 Separació de les conjunt de dades a entrenar i per fer el test

Després d'haver equilibrat les dades de la variable objectiu, dividim el nostre dataset entre els subconjunts d'entrenament i de validació. La proporció d'aquests subconjunts és del 70-30 respectivament. La separació del dataset l'hem fet de manera aleatòria utilitzant estratificació per mantenir l'equilibri de les classes en els dos subconjunts. El subconjunt d'entrenament s'utilitzarà per desenvolupar els models, mentre que el subconjunt de validació es reserva per avaluar el rendiment dels models en dades no vistes. D'aquesta manera reduïm el risc de sobreajustament dels models.

6.2 Entrenament de Models

Per tal de trobar el model amb una millor predicció, provarem diversos algorismes. Aquests algorismes els hem seleccionat en funció de la seva eficàcia a les tasques de classificació i per la capacitat per capturar relacions no lineals. Pel nostre cas, necessitem models de classificació multiclasse i que no requereixin de problemes lineals.

6.2.1 Radom Forest (RF)

Un dels models seleccionats ha sigut Random Forest ja que és un model bastant sòlid però al mateix temps ràpid. A més, és útils per datasets amb variables categòriques i numèriques barrejades, com en el nostre cas. També ens va bé perquè gestiona bé problemes no lineals i té mecanismes per evitar el sobreajustament. El problema que podem tenir és que en teniu un dataset força gran pot ser menys eficient.

El primer mètode que hem utilitzat a sigut utilitzant paràmetres predeterminats per fer una primera estimació bàsica. I un cop hem tingut aquesta primera estimació, hem intentat millorar el model. El primer intent de millora ha sigut buscar els hiperparàmetres que optimitzen el model. Per fer-ho hem utilitzat el `GridSearchCV`. Trobem que els millors resultats s'obtenen quan la profunditat màxima dels arbres no està limitada, el nombre màxim de característiques considerades per a dividir en cada node es fa amb *sqrt*, el

nombre mínim de mostres requerides per a una fulla és 4, el nombre mínim de mostres requerides per dividir un node és 2 i el nombre de branques en el bosc és 200.

Una altre prova per veure si aconseguíem millorar el model ha sigut mirant quins paràmetres tenen més importància a l'hora d'entrenar el model i quedar-nos amb els més importats amb l'objectiu de reduir el nombre de característiques de dataset. Les variables més importants són `TOTAL_LOSS` i `MEAN_INCOME` amb una importància del 0.22 i 0.19 respectivament. Amb aquest mètode, hem entrenat dos models: un amb les 31 característiques més importants i l'altre amb les 56 millors.

6.2.2 Gradient Boosting

Tot seguit hem entrenat models de Gradient Boosting ja que normalment són més precisos que el *Random Forest*. Va bé per datasets amb moltes dades i característiques. Cal ajustar els hiperparàmetres per tal d'optimitzar el model i són més costos que *Random Forest*. En concret hem provat XGBoost per la capacitat d'ajustament i perquè és molt robust, LightGBM perquè gestiona eficientment columnes amb molts valors zero i CatBoost per la bona gestió que fa amb les variables booleanes. Els dos primers cal ajustar molts paràmetres, mentre que el tercer és més fàcil d'ajustar.

XGBoost (XGB)

Hem realitzat la mateixa estratègia que hem seguit a l'hora d'entrenar models de *Random Forest*. Primer hem entrenat un model genèric i posteriorment l'hem intentat millorar amb tècniques per trobar els hiperparàmetres que ens donen els resultats més òptims.

Aquest algorisme és més costós que el *Random Forest*. Per aquest motiu primer buscarem els hiperparàmetres amb valors més separats i després, en funció dels hiperparàmetres trobats, entrenarem dos models més on reduint els intervals dels paràmetres per buscar la màxima precisió dels hiperparàmetres. Per la primera cerca trobem que els millors resultats s'obtenen quan el nombre d'arbres a construir és 200, la profunditat màxima de cada arbre és 6, la taxa d'aprenentatge que controla la magnitud d'actualització dels pesos per cada iteració és 0.1, el nombre mínim de mostres necessàries per formar una fulla és 5, la fracció de les dades utilitzades per construir cada arbre és 0.8 i que la fracció de característiques utilitzades per construir cada arbre també és 0.8.

A la segona quan el número d'arbres en el model és 250, la profunditat màxima dels arbres 5, la taxa d'aprenentatge 0.1155, el nombre mínim de mostres per formar una fulla 4, la fracció de les dades d'entrenament utilitzades per construir cada arbre 0.925, la fracció de característiques utilitzades per construir cada arbre 0.85, el paràmetre *gamma* que redueix l'overfiting 0, i les regularitzacions *L1* i *L2* són 0.25 i 5.5 respectivament.

Per la tercera, els paràmetres seguint l'ordre anterior són 200, 5, 0.1333, 3, 0.925, 0.9, 0.2, 0.0 i 4.

Light Gradient Boosting Machine (LGB)

Tot seguit, passem a entrenar el model LGB. Igual que abans, primer entrenem un model més general i després l'intentem millorar buscant els hiperparàmetres més òptims. El segon model l'hem entrenat amb els següents hiperparàmetres: *subsample*: 0.925,

reg_lambda: 10, *reg_alpha*: 0.5, *num_leaves*: 15, *n_estimators*: 300, *min_child_samples*: 20, *max_depth*: 7, *learning_rate*: 0.1525, *colsample_bytree*: 0.7750.

CatBoost (CB)

Ananem a entrenar l'últim algorisme que provarem de *Gradient Boosting*. Repetim el mateix procedement realitzat anteriorment. El segon l'entrenarem fixant les iteracions a 300, fent que la profunditat màxima dels arbres sigui 5, la taxa d'aprenentatge 0.1525, la regularitzaió *L1* 1, el nombre de bins per discretitzar característiques numèriques 128, la diversitat entre els arbres 0.1 i seguint una estretègia de construcció dels arbres *SymmetricTree*.

6.2.3 Stacking

També hem decidit entrenar models amb **Stacking** perquè ens permet combinar múltiples algoritmes alhora que redueix el risc de sobreajustament. Per el primer model hem combinat els primers algorismes que hem provat: el *Random Forest* i el *XGBoost*. Per el segon hem utilitzats els algorismes que ens han donat prediccions més bones fins el moment: *XGBoost* i *LGBoost*

6.2.4 Xarxes Neuronals (XN)

També hem provat d'entrenar dos xarxes neuronals perquè aquestes són més útils a l'hora d'interpretar patrons no lineals entre les característiques. Però, són més difícils d'entrenar i d'interpretar.

Hem provat dues xarxes neuronals. Les dues estan formades per tres capes i utilitzant *adam* com a optimitzador. En la primera xarxes, la primera capa té 64 neurones, la segona de 32 i l'última de 3.

En la segona, la primera capa té 128 neurones, la segona 64 i l'última 3. A més, hem afegit regularització en les capes i també un `Dropout` entre les capes per evitar l'overfiting.

6.2.5 Utilitzant una Funció Objectiu Personalitzada

També hem provat d'utilitzar un Funció Objectiu Personalitzada amb l'objectiu que els models s'entrenessin amb la finalitat de maximitzar la sobreestimació i reduir la infraestimació. Això, ho hem fet penalitzant més les classificacions on es feia una infraestimació i penalitzant menys en els casos de sobreestimació.

6.3 Avaluació dels models

El rendiment dels models de predicció s'avalua mitjançant el subconjunt de dades de validació. Per tal d'avaluar la precisió predictiva i la fiabilitat del model, utilitzem varies mètriques derivades de la matriu de confusió. Mirem l'estructura d'aquesta matriu a partir de la Taula 6.1.

Reals \ Predit	0	1	2
0	TP_0	$FP_{0,1}$	$FP_{0,2}$
1	$FP_{1,0}$	TP_1	$FP_{1,2}$
2	$FP_{2,0}$	$FP_{2,1}$	TP_2

Taula 6.1: Estructura Matriu de Confusió

On TP_i són les mostres de la classe i predites correctament i $FP_{i,j}$ són les mostres predites a la classe j però que realment són i.

A partir d'aquí podem extreure les formules per a les mètriques. L'**accuracy** és la proporció de les dades que s'ha classificat correctament.

$$Accuracy = \frac{\sum_{i=0}^{2} TP_i}{\#Mostres}$$

La **precisió** mesura, per a cada classe, la proporció de les mostres predites per la classe són realment d'aquesta classe.

$$Precision_i = \frac{TP_i}{TP_i + \sum_{j \neq i} FP_{j,i}}$$

La **sensibilitat** mesura, per a cada classe, quina proporció de les mostres que realment pertanyen a la classe han estat predites correctament.

$$Recall_i = \frac{TP_i}{TP_i + \sum_{j \neq i} FP_{i,j}}$$

L'**F1-Score** és la mitjana harmònica, per a cada classe, de la precisió i la sensibilitat.

$$F1Score_i = 2 \cdot \frac{Presicion_i \cdot Recall_i}{Presicion_i + Recall_i}$$

La **Macro Avg** és la mitjana aritmètica de les mètriques de cada classe.

A més, també hem creat una **mètrica personalitzada** per tal que la sobreestiamció del model sigui menys penalitzada que la subestimació. Per el nostre estudi, és millor que el model classifiqui un incendi a grau alt quan en realitat és de la classe 2 (mitja) que no pas que la classifiqui com a baix. Aquesta estimació consisteix en penalitzar més els elements de la *Matriu de Confusió* que estan a la part inferior de la diagonal en vers els de la part superior.

Capítol 7

Resultat dels models entrenats

Els resultats de tots els models entrenats els hem agrupat en un única Taula 7.1 per facilitar la comprensió d'aquests. Aquesta taula inclou les mètriques clau, descrites en l'avaluació dels modes, juntament amb els valors numèrics obtinguts fent una *CrossValidation* per a cada model.

Tot i així, totes les matrius de confusió de cada model estan agrupades a l'Annex A.

7.1 Taula amb els resultats dels diferents models

Model	**Accuracy**	**Precisió**	**Recall**	**F1-Score**	**Mètr. Per.**
RF	0.613	0.652	0.610	0.610	0.565
RF amb GS	0.637	0.692	0.631	0.631	0.527
RF amb 31 cols	0.598	0.633	0.595	0.595	0.597
RF amb 56 cols	0.612	0.652	0.609	0.609	0.566
XGB	0.637	0.686	0.633	0.633	0.537
XGB amb GS	0.644	0.698	0.640	0.640	0.531
XGB amb RS 1	0.644	0.698	0.640	0.640	0.531
XGB amb RS 2	0.644	0.698	0.639	0.640	0.531
LGB	0.642	0.696	0.637	0.637	0.530
LGB amb RS	0.644	0.698	0.639	0.639	0.531
CB	0.640	0.695	0.636	0.636	0.535
CB amb RS	0.642	0.698	0.637	0.637	0.533
Stacking (RF + XGB)	0.629	0.678	0.625	0.625	0.546
Stacking (XGB + LGB)	0.638	0.689	0.633	0.633	0.530
XN Bàsica	0.582	0.620	0.581	0.581	0.643
XN Avançada	0.599	0.655	0.597	0.597	0.581
Func. Obj. Per.	0.537	0.582	0.531	0.531	0.644

Taula 7.1: Resultats dels models entrenats

7.2 Discussió dels Resultats

Anem columna a columna. Si ens fixem a l'accuracy podem veure que els millors models que hem entrenats són els que tenen com a algorisme el *XGBoost*. Però els models amb l'algorisme *LGBoost* no es queden enrere.

Tenim un cas diferent a l'hora de centrar-nos amb la pressició. Tots els models tenen una pressició molt semblant excepte el primer model entrenat amb *Random Forest*, les Xarxes Neuronals i el model entrenat amb la Funció Objectiu Personalitzada.

Per la sensibilitat, la recall, els millors resultats els trobem amb els models *XGBoost* i *LGBoost*.

En el cas de la F1-Score, una altre vegada són els models *XGBoost* que tenen els millors resultats.

Finalment, tenim la mètrica que penalitza menys la sobreestimació que la infraestimació. Com ja pensàvem, el millor model respecte a aquesta mètrica és el que hem entrenat canviant la Funció Objectiu amb aquesta finalitat. Tot i així, la diferència amb la *Xarxa Neuronal Bàsica* entrenada és molt poca. La resta si que tenen valors més baixos.

Tenint en compte tot això tenint en compte les tres primeres mètriques, ens quedaríem amb els models *XGBoost* utilitzant el `GridSearchCV` i el primer utilitzant `RandomizedSearchCV`. Però per altre banda, si prestem més atenció a la nostre mètrica personalitzada, escolliríem la primera Xarxa Neuronal que hem entrenat o el model que hem entrenat amb la nostre Funció Objectiu Personalitzada. En el tema que estem tractant, el fet d'infraestimar un incendi pot ser molt perillós. Per aquest motiu, finalment, decidim que el model millor que tenim és el de la *Xarxa Neuronal Bàsica* perquè és un dels models que tracta millor la infraestimació i el mateix temps tampoc ens dona resultats gaire dolents en les altres mètriques.

Capítol 8

Conclusions

Aquest treball tenia com a objectiu crear un model prediu per tal de poder determinar la gravetat dels incendis en habitatges utilitzant la Base de Dades de la *NFIRS*. Durant el treball s'han identificat com a variables claus per aconseguir models amb rendiment notable, la mediana dels ingressos anuals i les pèrdues totals que ha causat l'incendi. Els resultats indiquen que el model XGBoost ajustat ha sigut el model més efectiu en quan a l'accuracy i l'F1-Score, mentre que les xarxes neuronals una millor capacitat de minimitzar la infraestimació de la gravetat, factor molt important en temes de gestió relacionats amb emergències. L'anàlisi també mostra que que els factor socioeconòmics d'una família, com els ingressos anuals baixos, i la falta de sistemes de prevenció, com detectors de fum, estan associats amb un grau de gravetat d'incendi més elevat. A més, també hem vist com els incendis estructurals són particularment perillosos.

Tot i així, hem tingut algunes limitacions. Tot i que la Base de Dades *NFIRS* és molt rica, ens ha faltat més informació per tal de predir amb més precisió la gravetat de l'incendi. Saber el material general de la casa, si és de fusta o està construïda de formigó ens mostraria la resistència d'ella. Perquè no és el mateix quedar-te amb una casa buida tota cremada per dintre que directament quedar-te només amb un escampat. També estaria bé saber si està prop de cables elèctrics. Com he vist anteriorment, els incendis originats per culpa d'un error tècnic són significatius.

Com a futures propostes de cerca, miraríem d'incloure més informació com la que hem comentat. Una altre cosa a provar seria crear un altre dataset on ajuntéssim les classes a predir de tal manera que n'acabéssim tenint dos. Tot seguit, podríem fer un primer entrenament dels models amb aquest dataset de classe binària i posteriorment acabar-lo d'ajustar a una tercera classe amb el dataset que hem treballat nosaltres. Per acabar, també podríem provar models de Xarxes Neuronals amb arquitectures més complexes.

En conclusió, aquest treball és una mostra del potencia que tenen les dades històriques per prevenir, gestionar i predir riscos d'incendis. La combinació de Bases de Dades robustes juntament l'ús d'eines d'aprenentatge automàtic, poden fer una gran millora de la seguretat del poble de cara els incendis.

Bibliografia

[1] U.S. Fire Administration. *Residential Fire Estimate Summaries (2013-2022)*. 2024. URL: https://www.usfa.fema.gov/statistics/residential-fires/.

[2] United Sttes Census Bureau. *Income by Zip code tabulation area*. 2024. URL: https://data.census.gov/table?q=Income%20by%20Zip%20code%20tabulation%20area&g=860XX00US30165,31905,35004,35005,35006.

[3] Fundació 'la Caixa'. *Incendios forestales en España. Importancia, diagnóstico y propuestas para un futuro más sostenible*. 2017. URL: https://elobservatoriosocial.fundacionlacaixa.org/es/-/incendios-forestales-en-espana-importancia-diagnostico-y-propuestas-para-un-futuro-mas-sostenible.

[4] Govern d'Espanya. *ESTADÍSTICA GENERAL DE INCENDIOS FORESTALES (EGIF)*. URL: https://www.miteco.gob.es/va/biodiversidad/temas/incendios-forestales/estadisticas-datos.html.

[5] FEMA. *Annual NFIRS Public Data*. URL: https://www.fema.gov/about/openfema/data-sets/fema-usfa-nfirs-annual-data.

[6] FEMA. *National Fire Incident Reporting System Version 5.0 Fire Data Analysis Guidelines and Issues*. 2011. URL: https://www.miteco.gob.es/va/biodiversidad/temas/incendios-forestales/estadisticas-datos.html.

[7] FEMA. *National Fire Incident Reporting System. Complete Refernce Guide*. 2015. URL: https://www.usfa.fema.gov/downloads/pdf/nfirs/nfirs_complete_reference_guide_2015.pdf.

[8] MAPFRE Fundación. *Estudio de víctimas de incendios 2023: análisis y prevención*. 2024. URL: https://www.fundacionmapfre.org/blog/estudio-victimas-de-incendios-2023/.

[9] D. Bruck M. Barnett i A. Jago. *MEAN ANNUAL PROBABILITY OF HAVING A RESIDENTIAL FIRE EXPERIENCE THROUGHOUT A LIFETIME: DEVELOPMENT AND APPLICATION OF A METHODOLOGY*. URL: https://publications.iafss.org/publications/aofst/7/85/view/aofst_7-85.pdf.

[10] Juan Manuel Chaves Posada. *Análisis de Probabilidad de Muerte en Incendios Residenciales*. 2014. URL: https://repositorio.uniandes.edu.co/server/api/core/bitstreams/25da10f9-6d09-44aa-bc36-3b688b57dd1b/content.

Capítol A

Matrius de confusió dels models entrenats

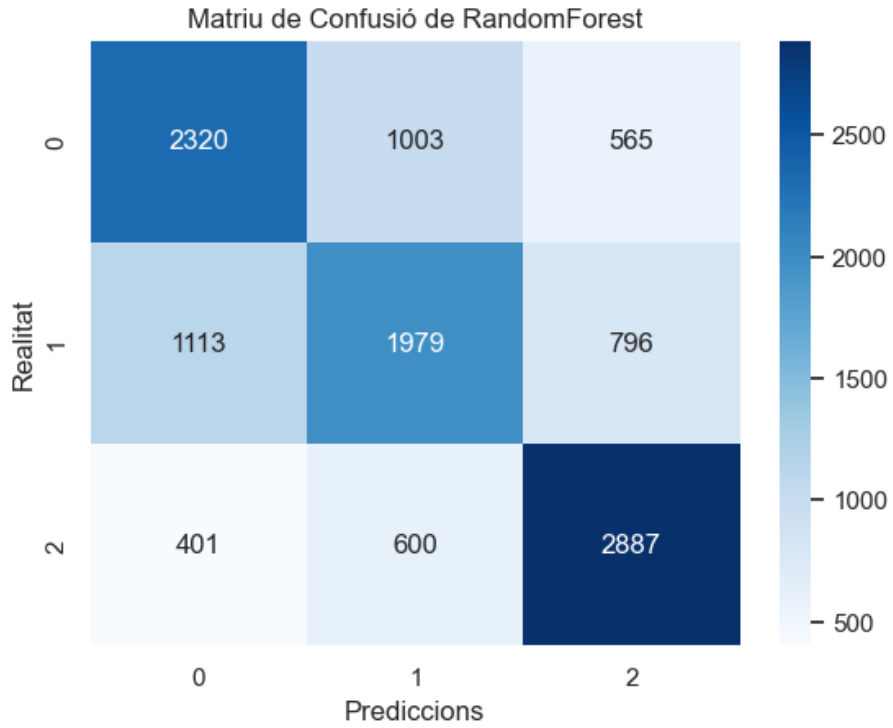

Figura A.1: Matriu de Confusió de RF Simple

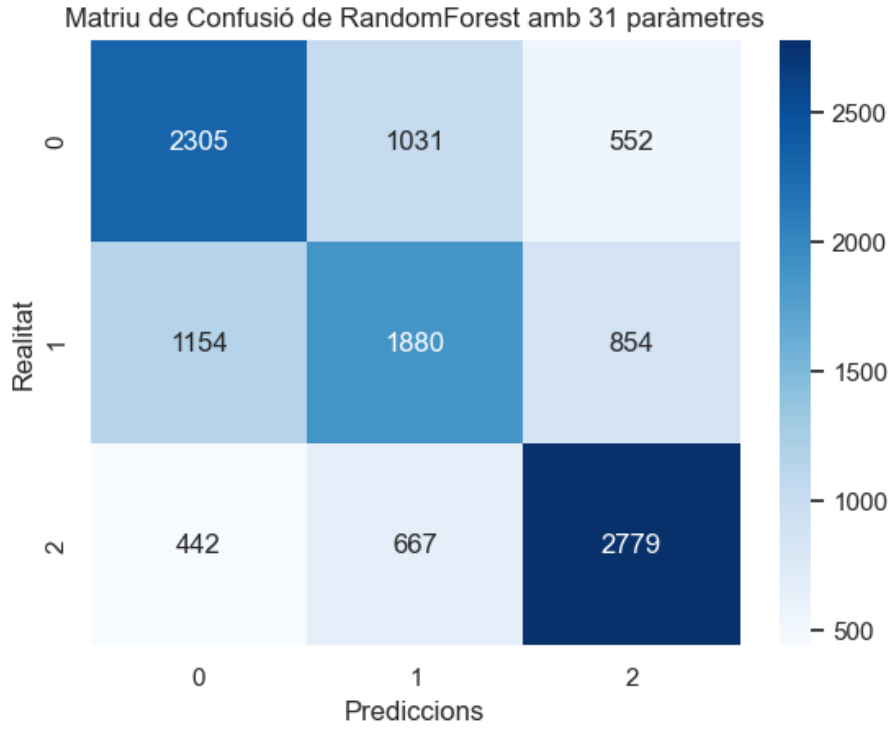

Figura A.2: Matriu de Confusió de RF amb GrindSearchCV

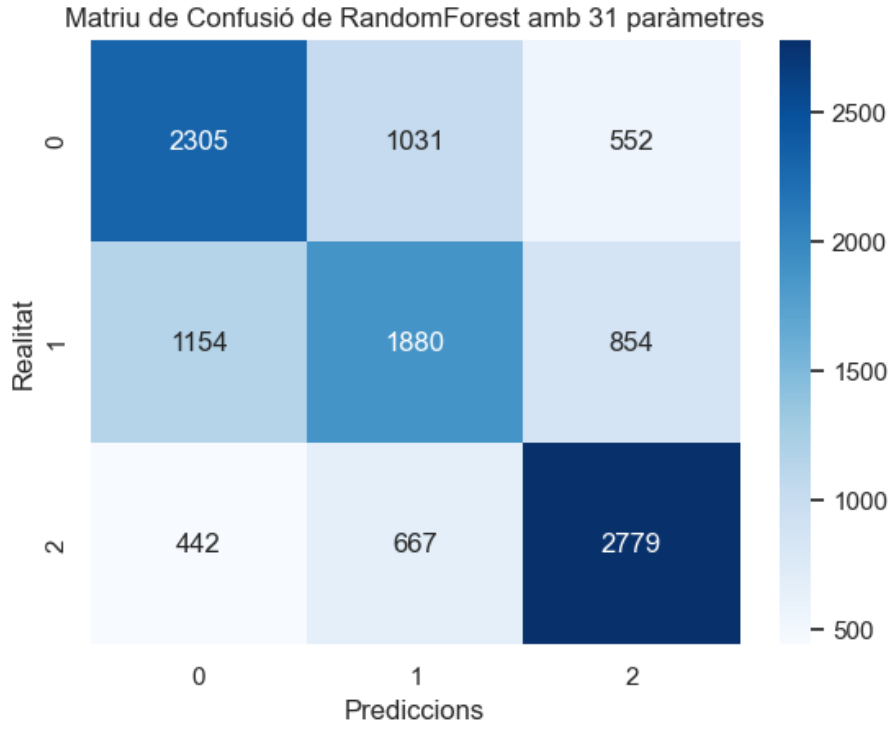

Figura A.3: Matriu de Confusió de RF amb 31 característiques

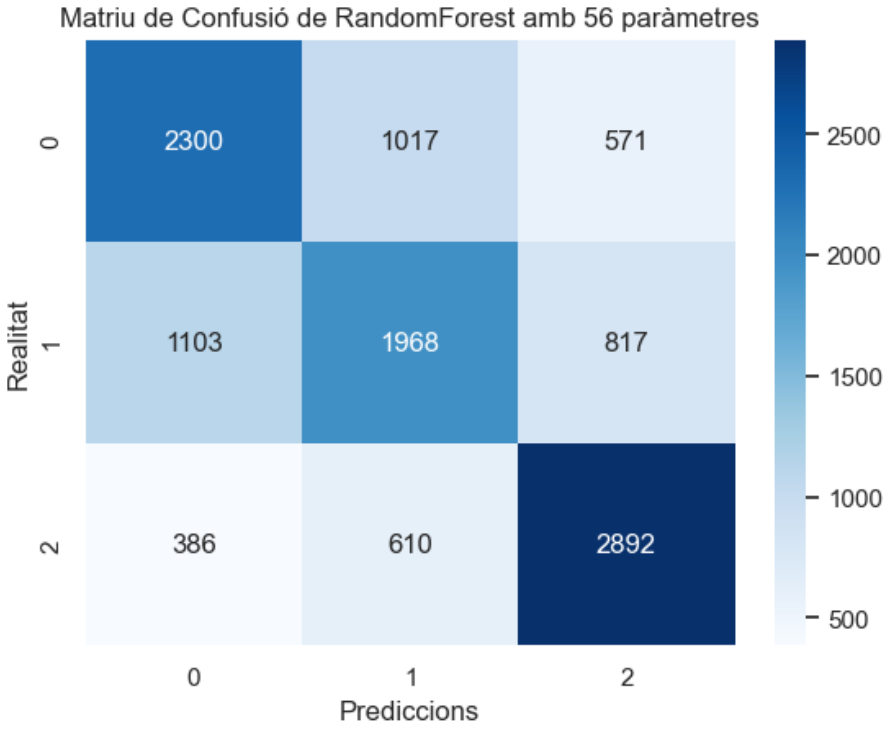

Figura A.4: Matriu de Confusió de RF amb 56 característiques

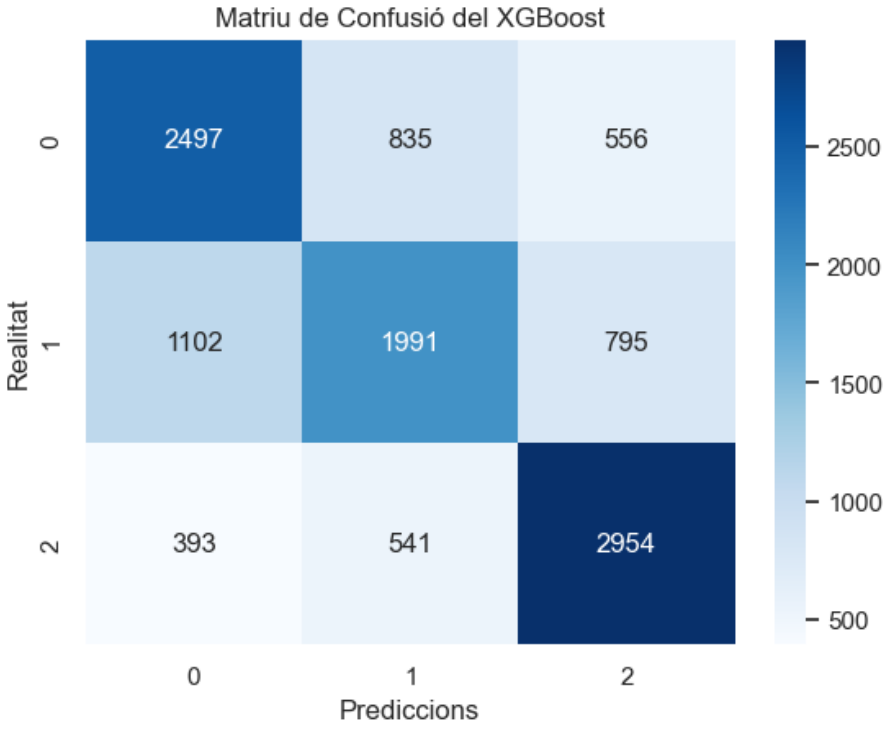

Figura A.5: Matriu de Confusió de XGB

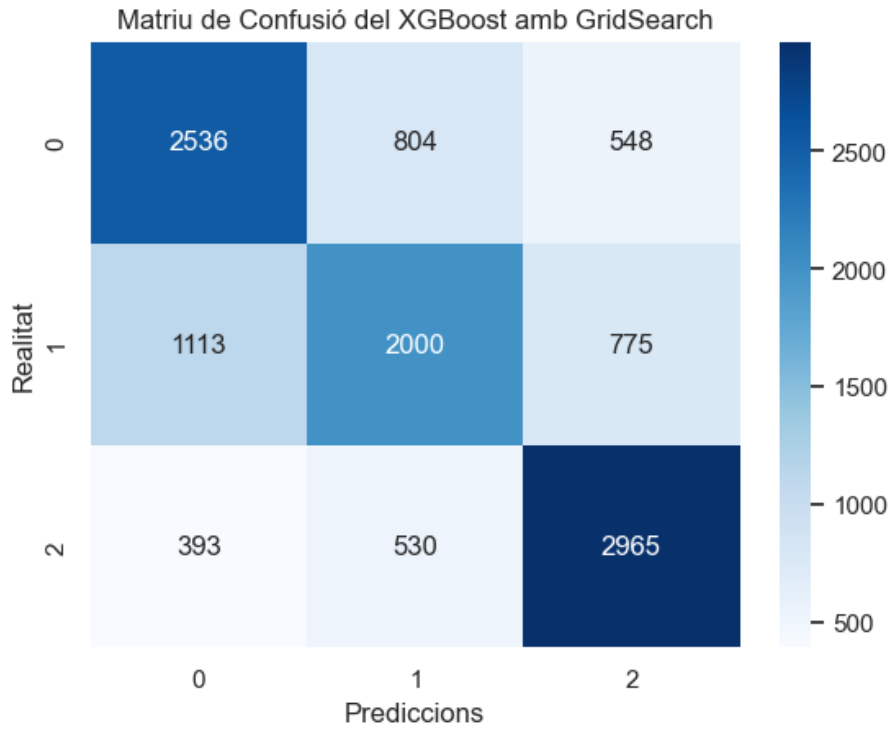

Figura A.6: Matriu de Confusió de XGB amb GrindSearchCV

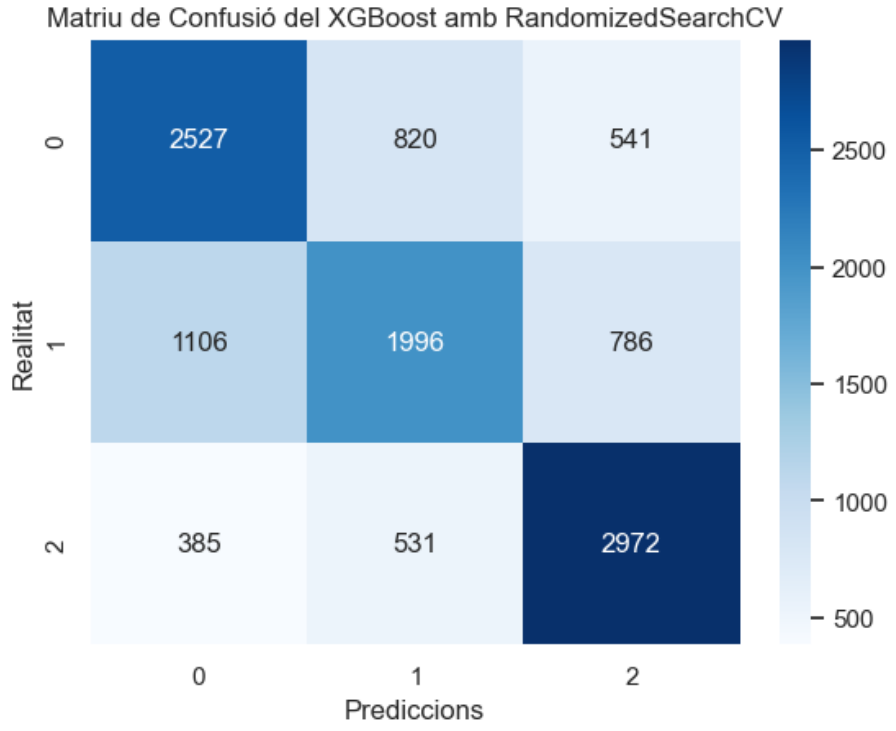

Figura A.7: Matriu de Confusió de XGB amb RandomizedSearchCV

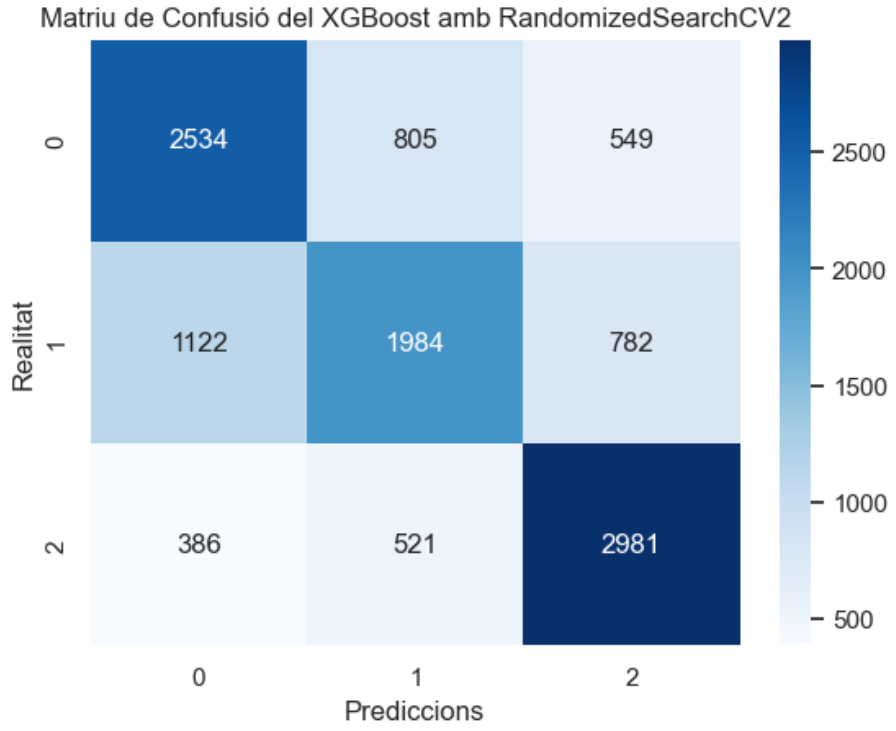

Figura A.8: Matriu de Confusió de XGB amb RandomizedSearchCV 2

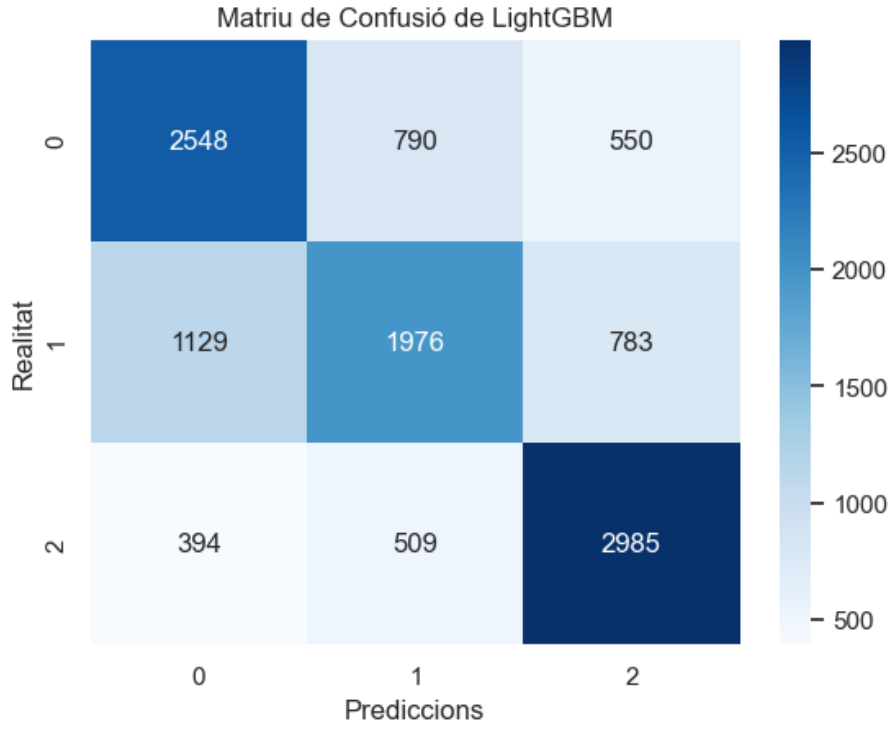

Figura A.9: Matriu de Confusió de LGB

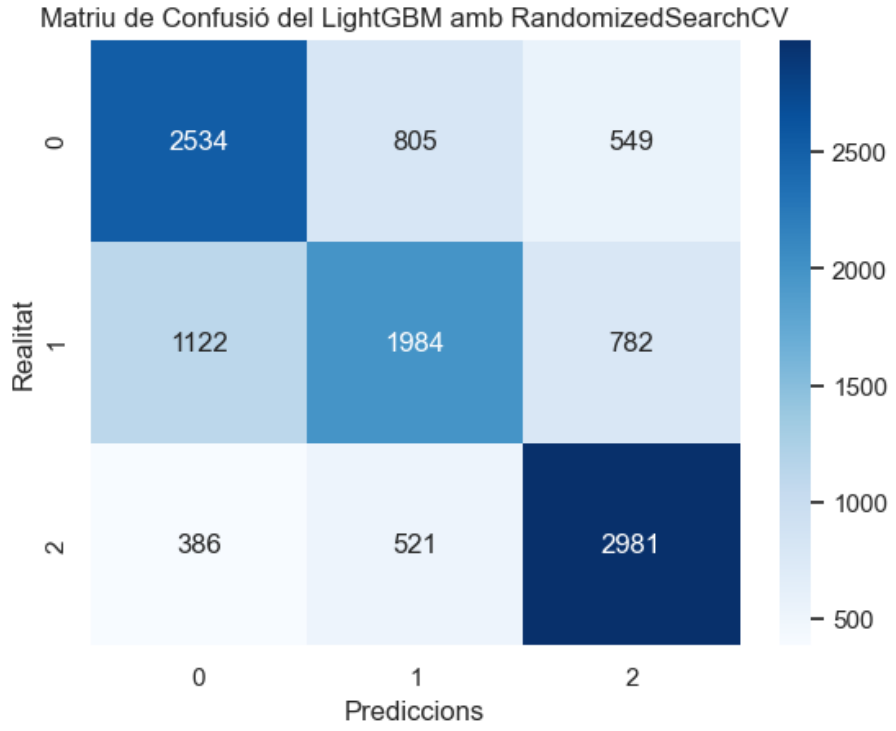

Figura A.10: Matriu de Confusió de LGB amb RandomizedSearchCV

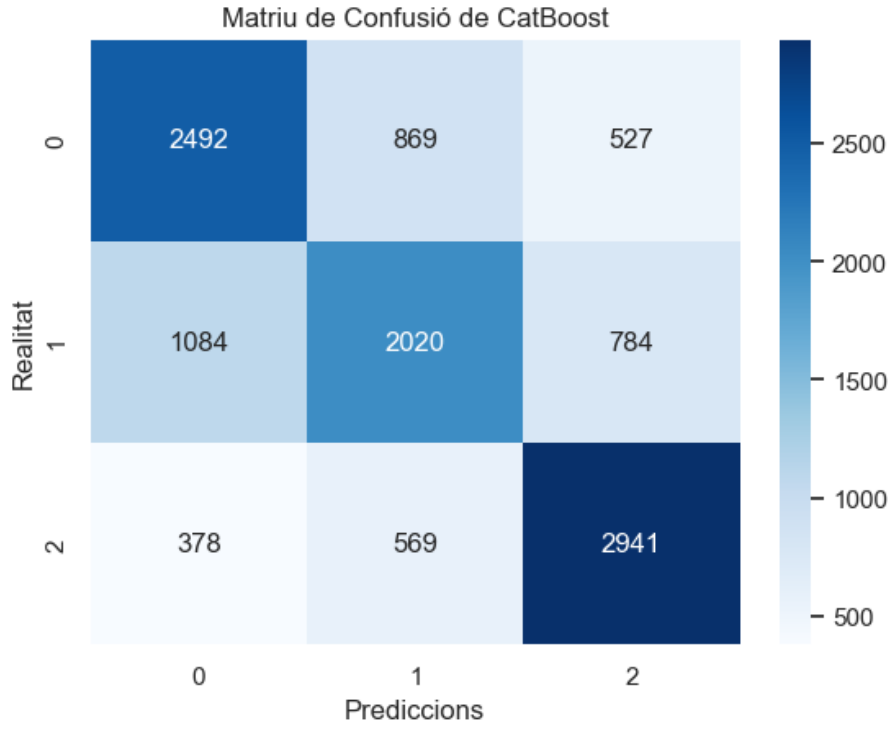

Figura A.11: Matriu de Confusió de CB

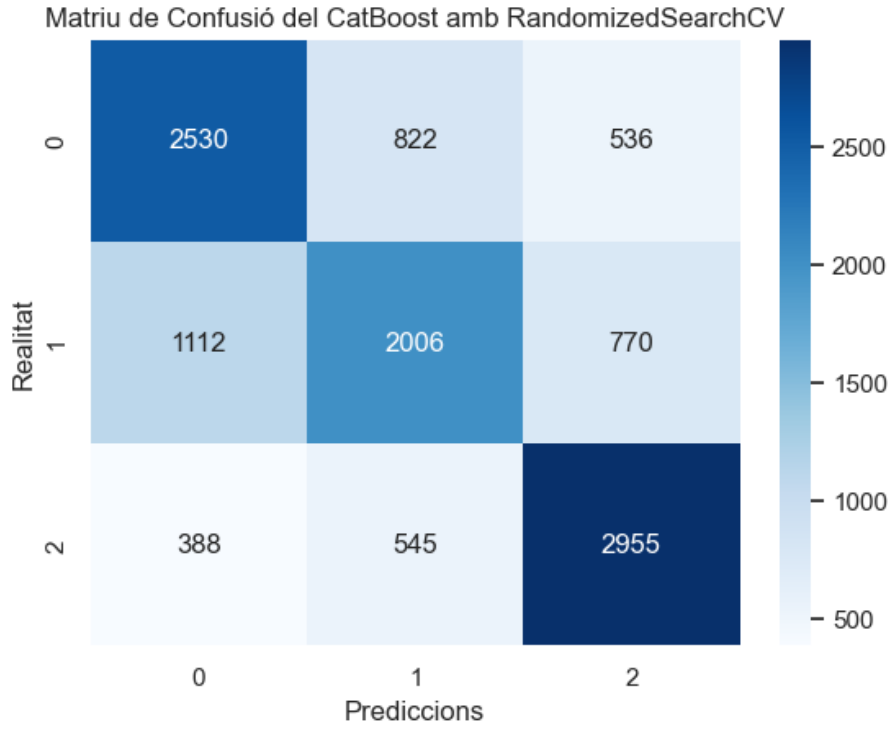

Figura A.12: Matriu de Confusió de CB amb RandomizedSearchCV

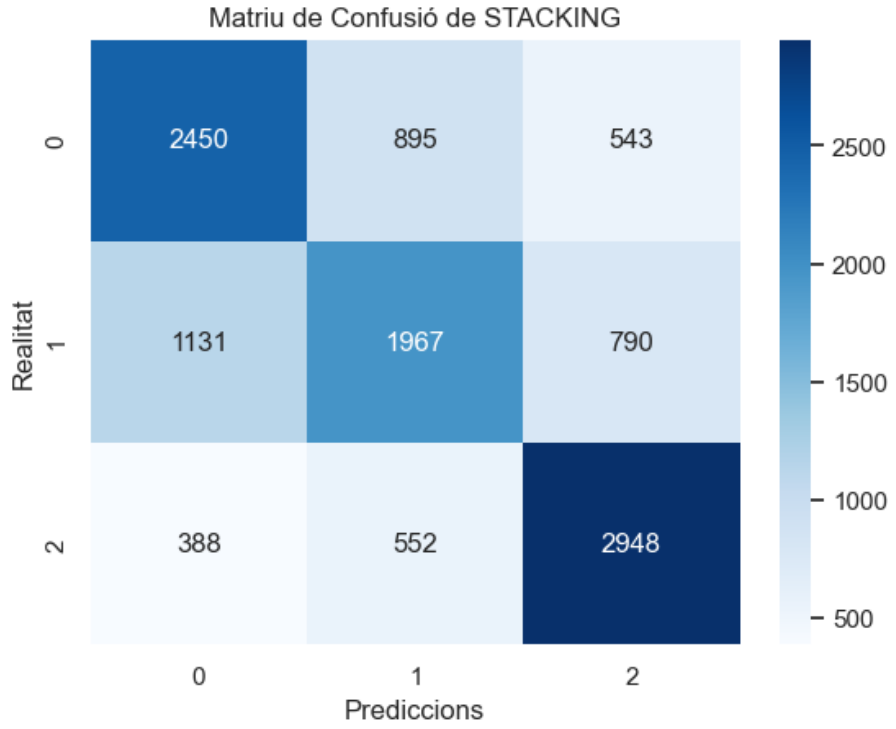

Figura A.13: Matriu de Confusió de Stacking RF i XGB

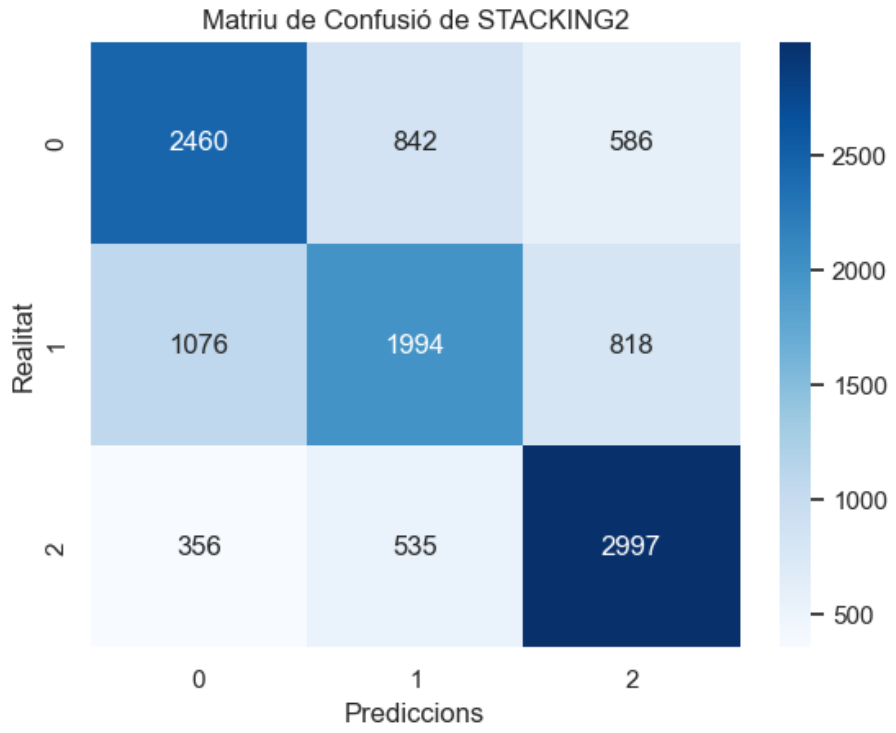

Figura A.14: Matriu de Confusió de Stacking XGB i LGB

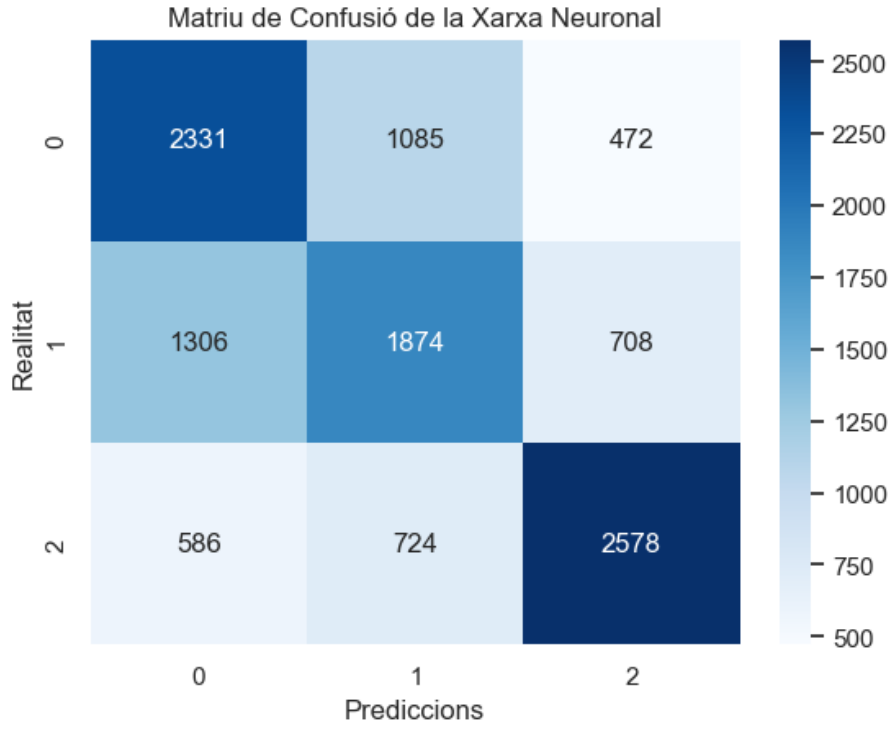

Figura A.15: Matriu de Confusió de XN Bàsic

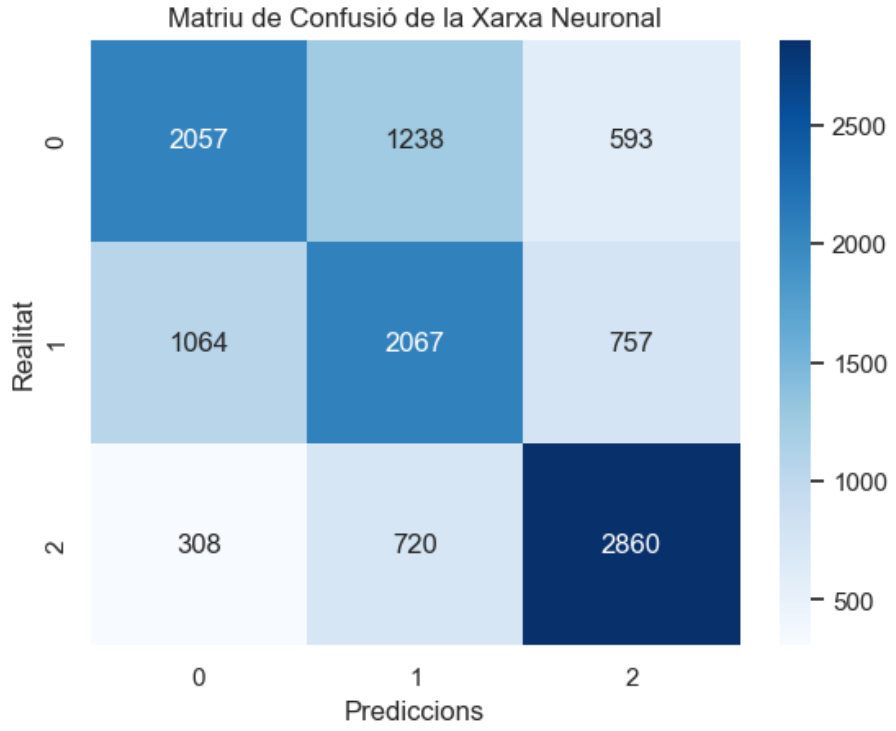

Figura A.16: Matriu de Confusió de XN Avançat

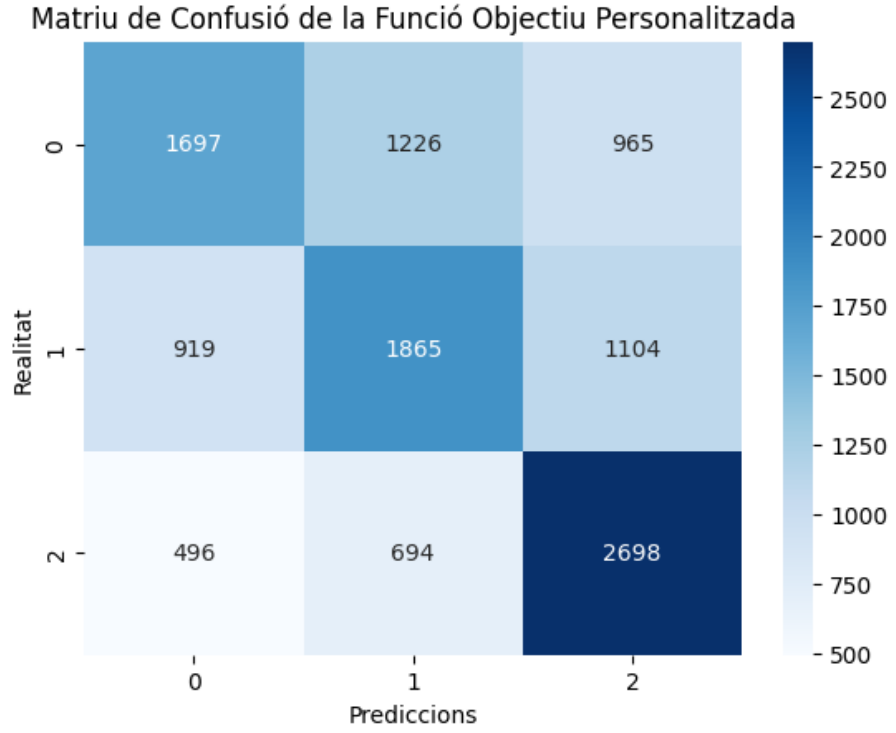

Figura A.17: Matriu de Confusió amb la Funicó Objectiu Personalitzada