Universitat Oberta de Catalunya

uoc.edu

Aproximació a la predicció de la delinqüència amb aprenentatge automàtic

Santiago Herrero Blanco
Màster universitari en Ciència de Dades
Sub-àrea Models predictius

Consultor: Sergi Trilles Oliver
Professor responsable de l'assignatura: Albert Solé Ribalta

Juny 2020

FITXA DEL TREBALL FINAL

Títol del treball:	*Aproximació a la predicció de la delinqüència amb aprenentatge automàtic*
Nom de l'autor:	*Santiago Herrero Blanco*
Nom del consultor/a:	*Sergi Trilles Oliver*
Nom del PRA:	*Albert Solé Ribalta*
Data de lliurament (mm/aaaa):	*06/2020*
Titulació o programa:	*Màster en Ciència de Dades*
Àrea del Treball Final:	*Models predictius*
Idioma del treball:	*Català*
Paraules clau	*Predicció, delinqüència, xarxes neuronals*

Resum del Treball (màxim 250 paraules): *Amb la finalitat, context d'aplicació, metodologia, resultats i conclusions del treball*

La millora en la recollida de dades i la seva anàlisi ha arribat en els darrers anys a l'àmbit de la delinqüència. Com a conseqüència d'aquests avenços i com a evolució dels mapes delinqüencials o de la detecció de punts calents (*hot spots*), policies, matemàtics, informàtics i altres professionals i tècnics han creat programes informàtics i models matemàtics que afirmen predir la delinqüència. Aquest treball pretén comprovar com són d'eficaços aquests models que intenten predir la delinqüència, si són aplicables només en un àmbit territorial o si es poden utilitzar en diversos contextos, per una o més tipologies delictives, o si són millors aquells que utilitzen més variables o menys. Finalment, també s'intentarà dissenyar-ne un utilitzant l'aprenentatge automàtic amb l'objectiu que millori, en un o més aspectes els programes o models ja existents.

Abstract (in English, 250 words or less):

In the recent years, there have been several improvements related to the collection and analysis of crime data. Law enforcement agencies have been using crime mapping and hot spot detection for a long time. As an evolution of those techniques and as a result of the breakthroughs in crime data, police officers, mathematics, computer scientist and other professionals have developed computer programs and mathematic models that assert predict crime. This work aims to check how effective those models are , if they can be applied to a limited territorial space or if it can be used in more than one place; for one or more type of crimes; or if more or less variables lead to a better model. Finally, I will also try to design a model, using machine learning, with the goal of improve the existing programs or models.

Índex

Llista de gràfiques

Llista de figures

Llista de taules

1. Introducció

1.1 Context i justificació del Treball

Descripció de la proposta i justificació de l'interès i la rellevància de la proposta

L'any 1956, Philip K. Dick [1] va publicar el seu relat *Minority Report*, història que, més recentment, es va popularitzar amb la pel·lícula amb el mateix nom i estrenada l'any 2002. La història parla d'una unitat policial encarregada de detenir persones que estan a punt de cometre fets delictius i que són detectats a partir de les visions tres mutants.

Més enllà de la ciència ficció, les ciències s'han anat incorporant a la tasca policial i en l'optimització i, en la predicció de la delinqüència. El registre de les dades sobre delinqüència, especialment les dades policials, la seva anàlisi i la plasmació en mapes han ajudat detectar on es concentra la delinqüència a través dels punts calents (*hot spots*). Aquesta és la base d'una bona part de les estratègies de predicció de la criminalitat, sabent on s'han produït els fets, es pot saber on és més probable que es produeixin més endavant. Però els mapes mostren o detecten dues variables principals, espai i temps, i l'avanç de la tecnologia permet analitzar moltes altres variables, la qual cosa ha permès aprofundir en altres anàlisis, com l'existència de patrons temporals en determinats fets, les característiques físiques dels espais on s'han comés els delictes, les relacions entre diferents tipus delictius, l'anàlisi de les característiques d'autors o de víctimes.

La pregunta principal, per tant, és si es pot predir la delinqüència. I a partir d'aquí apareixen altres qüestions. Poden els models matemàtics ajudar a detectar patrons per optimitzar la tasca policial? Quin valor tenen aquests models? Pot un mateix model servir per més d'un tipus delictiu? Pot necessitar-se més d'un model per un mateix tipus delictiu? Hi ha fets penals amb una major predictibilitat? Pot un mateix model utilitzar-se en més d'una ciutat? Com afecta a la predictibilitat la utilització d'un major o menor nombre de variables?

Per donar resposta a alguna d'aquestes preguntes en primer lloc s'analitzarà l'estat de l'art en aquests moments, si hi ha consensos o no a l'hora de respondre a aquestes preguntes. Finalment, s'intentarà respondre també a aquestes preguntes, desenvolupant i analitzant models predictius utilitzant les dades de fets penals coneguts que diferents ciutats ofereixen en els seus portals de dades obertes i amb altres dades complementàries per comprovar si milloren o no les prediccions realitzades.

Explicació de la motivació personal

El motiu per triar aquest àmbit de treball és perquè fa temps que faig seguiment sobre la predicció de la delinqüència, ja que, a més d'enginyer tècnic en informàtica, sóc llicenciat en Criminologia i treballo al Departament d'Interior de

la Generalitat de Catalunya. Entre d'altres tasques, analitzo les dades sobre la delinqüència (principalment a Catalunya). Per la meva formació i la meva tasca professional un dels objectius que tenia per realitzar el màster era obtenir les bases per poder aprofundir en els sistemes de predicció de la delinqüència i, fins i tot, poder-ne dissenyar o provar algun. Per tant, crec que fer-ho amb el TFM seria una gran oportunitat en la qual, a més de basar-me amb les dades obertes de ciutats d'altres països, intentaria obtenir els permisos per poder utilitzar dades de Catalunya.

1.2 Objectius del Treball

L'objectiu principal del projecte ha estat elaborar un algoritme que, a partir de dades delinqüencials sobre fets coneguts, pogués estimar amb quina probabilitat es tornaran a produir nous delictes en un lloc i moment determinat.

Un altre objectiu principal era determinar quines són les finestres d'espai i de temps que ofereixen estimacions més rellevants.

A més, també s'ha plantejat, com a objectius secundaris intentar confirmar o refutar les següents hipòtesis:

- Un model predictiu serveix per una única tipologia delictiva
- Diferents models poden ser òptims per diferents subtipus d'algunes tipologies delictives
- Per a algunes tipologies delictives no hi ha models predictius
- Els models predictius han de ser específics per a cada ciutat
- Els models predictius són més eficaços quan utilitzen més variables

1.3 Enfocament i mètode seguit

L'anàlisi de l'estat de l'art servirà per acabar de concretar si es el punt de partida és algun model ja existent, per tal d'analitzar-lo, validar-lo o millorar-lo, o si bé és millor dissenyar-ne un de nou utilitzant l'aprenentatge automàtic.

Per fer-ho, s'ha partit de dades disponibles a portals de dades obertes de les ciutats de Nova York, Chicago i Los Angeles (també s'ha considerat la tria d'altres dades o altres repositoris, sobre fets penals de grans ciutats com ara Londres). Inicialment s'havia pensat que es podria plantejar obtenir dades sobre els fets penals coneguts a Catalunya. Ara bé, tenint en compte els criteris de reproduïbilitat i que aquestes dades no es podrien publicar, s'ha treballat només amb dades disponibles a portals de dades obertes.

Inicialment hi havia la voluntat de cercar dades complementàries (per exemple sobre tipus d'edificacions, comerços, meteorologia, etc.) per valorar si els models són millors amb més o menys dades, però finalment no s'ha arribat a aquesta fase.

1.4 Planificació del Treball

La primera fase del treball ha consistit en analitzar l'estat de l'art. En el moment de començar el treball, ja disposava d'una àmplia selecció d'articles i documents

sobre la predicció de la delinqüència. A partir d'aquest catàleg inicial i amb una cerca complementària, s'ha seleccionat aquells articles més rellevants que per detectar alguns models que ja han estat validats i en quins contextos. Aquesta informació serà útil per determinar si es decideix crear un model a partir d'algun de ja existent o si s'opta per crear-ne un de nou, sense renunciar a incorporar elements de models existents.

Com s'ha comentat en l'apartat anterior, bona part de les dades a utilitzar ja estan disponibles, (dades sobre fets penals de, com a mínim, tres grans ciutats). En conseqüència es redueix la tasca relacionada amb l'obtenció de dades. No obstant això, sí que serà necessari homogeneïtzar les dades per poder treballar-les de manera conjunta (per exemple unificar la categorització dels fets penals o la utilització d'un mateix sistema de coordenades geogràfiques).

Un cop conegut l'estat de l'art i preparades les dades es començarà la fase de disseny i elaboració del model, que es realitzarà amb en llenguatge de programació Python i les llibreries necessàries. En funció del temps disponible, el model s'entrenarà i provarà en diversos contextos per intentar confirmar o refutar les hipòtesis plantejades.

La major part del treball s'ha desenvolupat amb maquinari propi, tot i que, en la part final s'ha utilitzat Google Colab[1].

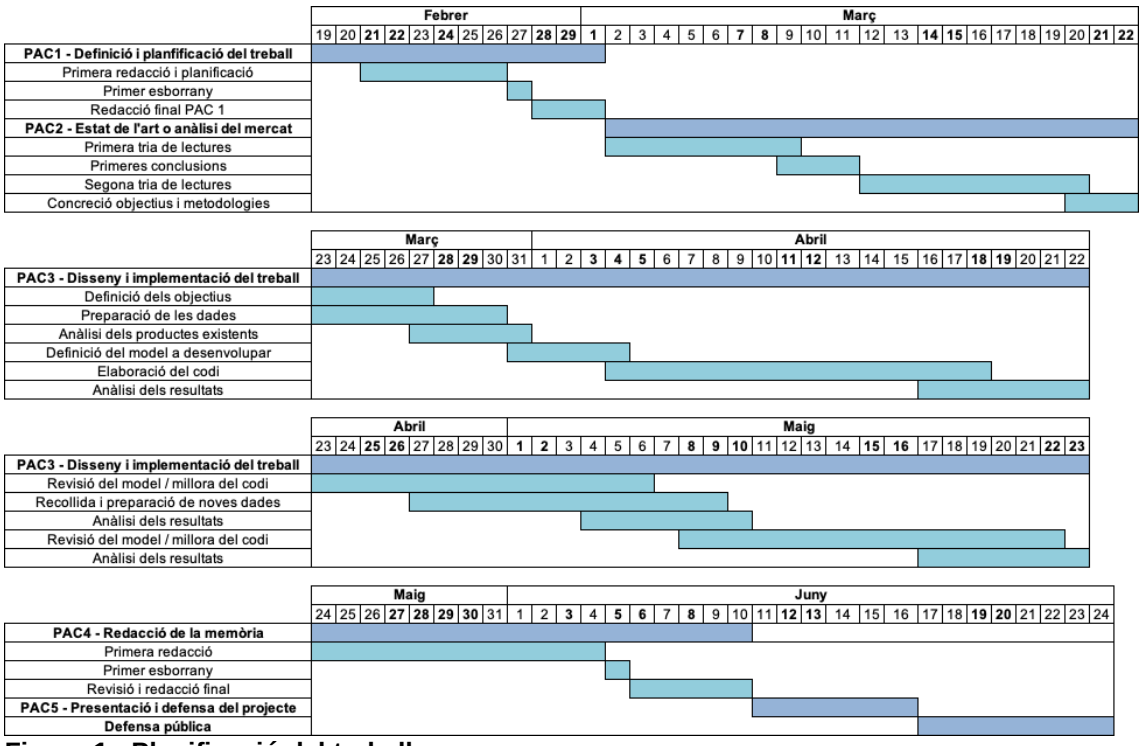


Figura 1 - Planificació del treball

[1] https://colab.research.google.com/

1.5 Breu sumari de productes obtinguts

No s'ha obtingut, com a tal, un producte definitiu. Sí que s'han implementat diferents arquitectures que podrien ser l'origen d'un model de predicció de la delinqüència, malgrat que els resultats inicials no semblen prou precisos i en conseqüència caldria millorar bastant aquesta arquitectura implementada.

1.6 Breu descripció dels altres capítols de la memòria

En el següent capítol es fa un breu repàs de l'estat de l'art en matèria de predicció de la delinqüència. El tercer capítol se centra en el disseny de la metodologia a utilitzar. La implementació realitzada s'explica en el quart capítol. El cinquè capítol s'expliquen les conclusions del treball. Finalment s'incorporen tres apartats més amb un glossari, la bibliografia i els annexos.

2. Estat de l'art

L'any 2008 William Bratton, l'aleshores cap de la policia de Los Ángeles va anunciar que el seu cos policial estava analitzant les dades sobre la delinqüència per tal de realitzar una tasca policial *predictiva*. Aquell és l'origen de l'anomenat *Predictive policing*. Tal i com expliquen Fergusson [2] i Hunt [3], el que realment feia, en aquell moment, la policia de Los Ángeles, era construir mapes de densitat amb els fets penals coneguts, de manera que es poguessin identificar els *hot spots* aquells punts de la ciutat amb una major concentració de fets penals.

La tècnica dels *hot spots* ha estat àmpliament utilitzada i coneguda per la policia. Com relata Hunt [3] s'hi poden trobar antecedents històrics remots (i potser anecdòtics) en el segle XIX. La concentració de fets en determinats espais, sovint es relaciona també amb la victimització repetida [4], quan una mateixa persona (o grup de persones) o espai (gran part dels estudis es concentren en robatoris a habitatges) són més propensos a patir fets delictius. La seva base científica dels *hot spots* s'origina en el darrer terç del segle XX amb les teories criminològiques de les activitats rutinàries de Cohen i Felson [5] i dels patrons delictius de Brantingham i Brantingham [6], que, a més de dels factors personals de víctimes i autors, identifiquen el lloc i el moment com elements claus per tal que els delictes s'acabin cometent. Però no és fins els anys 80 del segle XX, quan es popularitza, en part per la sistematització en la recollida de les dades, que també té com a referència a Bratton, que va impulsar el sistema *Compstat* als anys 90 del segle XX, en la seva etapa com a cap de la policia de Nova York. Aquesta seria la base de la, adoptant el concepte de Mayer-Schönberger i Cukier [7], *datificació* de la delinqüència.

Malgrat que el que feia l'any 2008 la policia de Los Ángeles no es pogués considerar una activitat predictiva, sí que es considera en detonant d'una (o múltiples) línies d'investigació i innovació en matèria de predicció de la delinqüència.

Cal matisar, en aquests moments que el predictive policing, com senyala Brantingham [8] és un procés que consta de tres parts:

1) Ingesta de dades d'un o més tipus
2) Les dades són tractades amb mètodes algorítmics per pronosticar l'ocurrència de delictes en un territori d'interès, i
3) La policia utilitza aquests pronòstics per informar les decisions estratègiques i tàctiques sobre el terreny.

Deixant de banda el fet que Brantingham no descriu algunes fases d'un projecte d'anàlisi de dades (com la neteja o tractament de les dades), s'incorpora aquesta descripció ja que bona part dels estudis avaluen els models en funció de la utilització que realitza la policia, mentre que l'objectiu d'aquest treball no pot arribar a aquesta tercera fase i per això es parla de predicció de la delinqüència.

Les múltiples investigacions, estratègies i models desenvolupats per a la predicció de la delinqüència, es poden agrupar i analitzar en funció de diferents paràmetres.

Per una part, en funció de l'enfocament es podrien considerar dos grans grups de models o estratègies. Per una part, la majoria dels projectes es basen segueixen estratègies derivades, en major o menor mesura, dels *hot spots*, ja que prenen com a punt de referència el lloc on passen els fets. En contraposició, també hi ha un altre conjunt de models que se centren en les persones, principalment en els autors i la seva capacitat de reincidència i aplicats tant per part de la policia com per part dels organismes judicials a l'hora de dictar sentències, però també en víctimes i el seu risc de victimització o revictimització. Aquesta classificació, citada, entre d'altres per Fitzpatrick, Gorr i Neill [9] i Hunt [3], no recull altres estratègies, com la desenvolupada per Quijano-Sánchez i altres [10] que prediu la probabilitat que una denúncia presentada davant de la policia sigui falsa.

A més d'aquests grans grups, altres elements que es tenen en compte són les dades utilitzades en el procés. Gairebé la totalitat dels models utilitzen dades dels registres policials. En aquells supòsits en què es tracta d'aplicacions utilitzades per la policia, les dades són internes, però també hi ha investigacions i models dissenyats a partir de dades obertes com podrien ser aquells desenvolupats en competicions com la convocada pel *National Institute of Justice*, Hunt [3], o algunes competicions o dades obertes de *Kaggle*. A més de les dades policials, en alguns casos s'utilitzen també altres dades complementàries, com ara les dades sociodemogràfiques [11], dades d'entorn, o fins i tot de xarxes socials. Encara que anecdòtiques, també hi ha iniciatives que han volgut predir la delinqüència obviant les dades policials i utilitzant, com a base, dades de xarxes socials, en concret de Twitter [12] i [13].

Encara que es parla de predicció de la delinqüència en genèric, en la majoria dels casos, els models se centren en tipologies de delictes concretes, per tant seria més idoni parlar de predicció o previsió de determinats delictes. I en la tria d'aquestes tipologies delictives influeixen dos factors principals.

Per una part, en l'àmbit criminològic i el policial és àmpliament conegut que no tots els fets penals es denuncien ni tots arriben a coneixement de les institucions de l'administració, i es fa referència a aquest volum de fets com la *xifra negra* de la delinqüència [14]. Però també és conegut que, per motius diversos, no tots els delictes tenen el mateix percentatge de fets desconeguts, fins i tot s'ha començat a parlar de *xifres negres* [14].

Per una altra part, les dades recollides per la policia tenen mancances i imprecisions. Així, com relaten Brantingham [8] i Gerell [15], hi ha errades en la localització geogràfica o temporal dels fets, en alguns casos per errades o incorreccions en la recollida o introducció de les dades, en d'altres per desconeixement sobre el lloc o el moment en què s'han produït els fets. A més, no tots els fets passen en un lloc o un moment concret, sinó que poden realitzar-se en un trajecte (fets que passen en mitjans de transport) o que tenen una durada més o menys.

Per aquests motius, els fets sobre els quals s'ha produït una major quantitat d'investigacions predictives són els robatoris en habitatge i els fets sobre vehicles (sostraccions dels vehicles, sostraccions als vehicles o danys), ja que es tracta de tipologies delictives sobre les quals hi ha un alt volum de denúncies (poca xifra negra) i la precisió sobre el lloc dels fets acostuma a ser alta (no exacta o no en tots casos com es veurà més endavant). En l'àmbit americà, també hi ha bastants estudis que han analitzat fets violents, especialment els relacionats amb bandes juvenils. A més, una altra dimensió que cal tenir en compte respecte a les tipologies delictives, és que encara que no és majoritari, no són infreqüents els estudis que analitzen diverses tipologies penals en conjunt, especialment grans agrupacions de fets. També s'han localitzat iniciatives per predir o preveure altres casos que potser no tenen una vinculació directa amb la delinqüència, com són els accidents de trànsit [16].

Les dues dimensions sobre les quals es realitzen les prediccions són l'espai i el temps i en aquest sentit, cal tenir en compte tant la recollida de dades com la predicció que es realitza. Com s'ha comentat abans, en la recollida de dades no sempre és possible delimitar amb precisió un moment i un lloc en el qual passen els fets. De manera similar, els models atorguen probabilitats més petites (i amb marges d'error més grans) com més petit és l'espai o més breu és el moment. Així, els models desenvolupats rarament realitzen prediccions sobre punts, sinó que ho fan sobre àrees més grans o més petites, en alguns casos resultants de dividir el territori en espais homogenis, i en d'altres utilitzant delimitacions territorials administratives (seccions censals, barris, districtes, poblacions). De la mateixa manera, en la delimitació del temps, la majoria dels models realitzen les prediccions per torns policials o per dies i, en alguns casos, aquestes prediccions es fan per terminis més llargs com ara, setmanes.

Per últim, cal tenir en compte les metodologies de predicció analitzades. En alguns casos, en tractar-se de productes comercials, la informació completa sobre el model, ja sigui pel que fa a l'algoritme o per les dades utilitzades no és completa. No obstant això, en la gran majoria dels casos es realitzen anàlisis estadístics més o menys complexos.

El cas més conegut és l'utilitzat pel sistema *PredPol*, que tenint en compte les teories criminològiques esmentades va fer un paral·lelisme entre les rèpliques d'un terratrèmol i els delictes repetits en un espai proper i va utilitzar el model de seqüència de rèplica de tipus epidèmic (ETAS, pel nom en anglès, *Epidemic-Type Aftershock Sequence*). Aquest model, tal i com explica Brantingham [8], descriu el risc del delicte per a un temps determinat *t* en funció de quatre paràmetres:

$$\lambda(t) = \mu + \theta \sum_{t_i<i} \omega e^{-\omega(t-t_i)}$$

On:

- λ és la taxa de delictes per a aquell moment en un espai concret
- μ representa la taxa estacional històrica de fets en aquell lloc
- θ descriu el nombre de casos de victimització repetida esperada després d'un fet que actua com a detonant, i

- ω és l'escala de temps sobre la qual la repetició dels fets actua.

A més, t_i és el temps de cada fet delictiu i cada esdeveniment *i* previ al temps *t*.

L'altre gran programa és el RTM, modelització del risc del territori (*Risk Terrain Modeling*), introduït per Caplan, Kennedy i Miller el 2011 [17]. En aquest cas s'utilitzen, per una part, les dades dels fets coneguts i, per una altra, les dades de localització de determinats establiments o activitats que poden estar relacionats amb els delictes (espais de venda d'alcohol, aparcaments, farmàcies, etc.) El càlcul es realitza amb diversos passos, aplicant regressions de Poisson, filtrant variables i novament regressions de Poisson i binomials negatives.

Altres estudis, utilitzen eines d'aprenentatge automàtic com els *random forests* [11, 18], amb els quals milloren substancialment els resultats de RTM o de l'aplicació de l'anàlisi de densitat Kernel.

Un programa menys conegut, *HunchLab* (actualment *ShotSpotter® Missions*TM), aplica també l'aprenentatge automàtic (potenciació del gradient, arbres de decisió i mètodes de validació creuada) combinat amb algunes de les tècniques explicades anteriorment RTM, delictes propers, i els processos d'auto excitació dels punts [19].

En els darrers anys, també s'han desenvolupat models que utilitzen diversos tipus de xarxes neuronals per intentar predir la delinqüència.

La presència o l'actuació de la policia en una determinada àrea pot fer incrementar els fets coneguts en aquella àrea. Amb aquesta premissa Ensing i altres, [20] , consideren que aquest factor introdueix una realimentació (*feedback*) a les dades que pot esbiaixar els resultats de programes com *PredPol*, i proposen un model d'aprenentatge automàtic basat en urnes (en concret utilitzen el model generalitzat Pólya[2]) per reduir aquest biaix i que es tinguin en compte possibles delictes que no han arribat a ser coneguts per la policia perquè estava en altres actuacions.

Rummens i altres [21] proposen un model amb una regressió logística binària i una xarxa neuronal amb perceptrons i una capa oculta, que apliquen a la ciutat d'Amsterdam dividint la ciutat en una malla i estimant la probabilitat que en finestres de temps de 2 setmanes es produeixin determinats delictes (entrenen un model per a cada tipus de delicte).

Duan i altres [22] i Wang i altres [23] proposen models de xarxes neuronals convolucionals profundes amb arquitectures complexes. En el primer cas utilitzen una arquitectura amb 7 capes (Conv, Inception, MaxPool, Fractal Block, Conv, Max Pool, Dense) i, aplicada a delictes de la ciutat de Nova York, i obtenen resultats superiors als obtinguts amb models de mitjanes de pesos, *Suport Vector Machines*, *Random Forests* o xarxes neuronals poc profundes completament connectades. En el segon cas apliquen una arquitectura més complexa. Per una part, divideixen la ciutat de Los Angeles en una malla, i per a cada cel·la què

[2] Aquest model calcula el color que tindran les quatre boles d'una urna, després d'un seguit de passos, si a cada pas s'ha extret una bola i substituït per una altra, amb el punt de partida que només hi ha boles de dos colors i que inicialment hi ha dues boles de cada color dins de l'urna.

separen per a cada unitat territorial, extreuen els valors de la tendència, el període i la proximitat del fet. Per l'altra, per a cadascun d'aquests conjunts de valors comparen dos models, en un apliquen una xarxa amb una capa convolucional, 6 capes d'unitats residuals, i una altra capa convolucional, i en fusionen els resultats; en l'altre, no inclouen les capes convolucionals, i n'obtenen millors resultat amb el model que incorpora les capes convolucionals.

Zhuang i altres [24] comparen l'efectivitat d'una xarxa neuronal recurrent, amb dos models de xarxes neuronals espacio-temporals, LSTM i GRU. Utilitzen dades de Portland, Oregon, i obtenen resultats superiors amb les LSTM i GRU que amb la xarxa neuronal recurrent, i, entre aquestes dues, els resultats de la corba ROC tenen una AUC lleugerament superior.

Ramirez-Alcocer i altres [25], per la seva part, utilitzen una xarxa neuronal LSTM. Analitzen dades de la ciutat de Chicago, amb una única capa LSTM, però en aquest cas, en comptes d'intentar predir els fets, intenten pronosticar si, dels delictes coneguts, a partir del tipus de crim i del lloc on ha passat, s'haurà produït alguna detenció, assolint un percentatge d'encert superior al 80% en la majoria dels casos analitzats.

Els programes, sistemes, models i algoritmes de policia predictiva o de predicció de la delinqüència han estat sotmesos a moltes crítiques, posant en evidència algunes de les seves debilitats o mancances. Una part important d'aquestes crítiques es basa en els biaixos existents en les dades policials i que afecten, especialment els models centrats en les persones. Un dels motiu és els models centrats en agressors, es basen en identificacions, escorcolls i detencions prèvies i aquestes accions són molt més nombroses sobre persones amb determinats perfils racials (afroamericans i hispans als EUA), amb la qual cosa, aquest biaix inicial acaba produint resultats esbiaixats [26] [27].

Les crítiques també existeixen cap als models que es basen en l'espai. En alguns casos, es tracta de consideracions que els mateixos autors realitzen sobre les dades que realitzen dels seus models, de cara a avaluar-ne els possibles biaixos. Així Brantingham [8], a més de considerar les possibles deficiències en la ubicació dels fets, també posa de manifest que, en ocasions, les policies tipifiquen incorrectament els fets delictius, agreujant o alleugerint els fets denunciats en funció dels seus interessos. Quant a la ubicació Gerell [15] evidencia errors en la ubicació dels delictes d'incendis en vehicle a Malmö, Suècia, comparant les dades policials amb les provinents dels bombers adreçats a aquells serveis.

3. Disseny

Per tal d'abordar el disseny del projecte cal recordar breument els objectius inicialment previstos.

L'objectiu principal és l'elaboració d'un algoritme que estimi la probabilitat que es tornin a produir nous delictes en un lloc i moment determinat, i quines serien les finestres d'espai i de temps que oferirien estimacions més rellevants. A més d'aquests objectius, també es plantejaven algunes hipòtesis relacionades amb algunes variables de les dades a analitzar, com ara l'efectivitat amb diferents tipologies o subtipologies delictives, o la particularitat dels espais (com ara les ciutats).

Seguint el cicle de vida de les dades, es plantegen les següents passes per aconseguir aquests objectius: captura, emmagatzematge, preprocessat, anàlisi, visualització i publicació.

3.1 Captura de les dades

Les dades necessàries per a aquest projecte són els fets penals que han passat durant un termini de temps en un lloc determinat. És necessari un cert nivell en el detall de des dades, ja que les dades agrupades no permetrien una anàlisi en profunditat. En aquests moments, a Catalunya les dades publicades al portal de dades obertes de la Policia de la Generalitat – Mossos d'Esquadra[3] (que recull dades recollides per aquest cos policial i també per les policies locals de Catalunya) es presenten agregades tant pel que fa al temps (dades mensuals) com pel que fa a l'espai (a nivell d'Àrea Bàsica Policial). Per tant, no serien útils per al projecte. Malgrat que per motius laborals puc tenir accés a aquestes dades, és necessària una autorització per a poder-les utilitzar per motius externs a la organització, com ara per a investigacions acadèmiques. Atès que sovint aquestes autoritzacions es demoren en el temps, m'he estimat més treballar amb dades ja disponibles d'altres ciutats. A la resta d'Espanya, les dades publicades en obert tenen característiques similars, per tant tampoc no són òptimes per poder-hi treballar. Un entorn relativament proper en què hi ha accés a dades amb un major nivell de detall és Anglaterra i Gal·les. No obstant això, encara que les dades estan molt més desagrupades territorialment (els fets es poden consultar per punts on han passat els fets), l'agrupació temporal es mensual, per tant tampoc s'ha considerat adient per als objectius del projecte.

Tenint en compte aquestes qüestions de les dades públiques dels entorns més propers, s'ha optat per utilitzar informació publicada en els portals de dades obertes de les tres ciutats dels EUA amb més població: Nova York, Chicago i Los Ángeles. En aquestes tres ciutats les dades disponibles s'actualitzen de manera periòdica, afegint les dades més recents i, eventualment, actualitzant informació

[3] https://mossos.gencat.cat/ca/els_mossos_desquadra/indicadors_i_qualitat/dades_obertes/cataleg_dades_obertes/dades-delinqueencials/

antiga. No obstant això, inicialment s'ha volgut treballar amb dades d'anys complets, per tant, dades fins finals de 2019.

No obstant això, si el projecte es volgués implantar en alguna localitat concreta, caldria un accés a les bases de dades policials, o com a mínim a un recull de les dades amb una informació mínima: moment del fet, tipus de fet i lloc del fet.

3.2 Emmagatzematge

Tenint en compte que el projecte no pretén treballar amb dades generades ni actualitzades de manera periòdica, no cal una estructura especial pel que fa a l'emmagatzematge. Les dades són publicades i accessibles als portals, però per optimitzat el temps d'execució s'ha optat per descarregar els conjunts de dades en fitxers .csv per treballar en local, i posteriorment publicant les dades a una carpeta privada de Google Drive per tal de poder-les treballar amb Google Colab.

No obstant això, s'ha tingut en compte que la recerca sigui reproduïble. Per aquets motiu, més enllà del fet que les dades siguin accessibles en portals de dades obertes, les dades que s'han filtrat i treballat, corresponents als anys 2018 i 2019 s'han desat en el repositori públic Zenodo (veure apartat 3.6 sobre la publicació).

3.3 Preprocessat

Una de les passes que més temps impliquen en qualsevol projecte de ciència de dades és el tractament de les dades previ a la seva anàlisi. En aquest cas aquestes s'ha concentrat en tres accions diferents: la selecció dels camps que podien aportar informació útil, la conversió d'alguns camps en formats que permetin la seva anàlisi, la creació de noves variables i la neteja d'algunes dades.

En el primer cas, un estudi inicial dels conjunts de dades ha evidenciat que en cadascuna de les ciutats els conjunts de dades publicades contenien variables diferents i que, en alguns casos no totes les variables eren necessàries o podrien arribar a ser útils per als objectius del projecte. Així, per exemple, a més de la informació necessària que s'ha comentat anteriorment (referida al moment, el lloc i el tipus de fet) també hi havia informació relacionada amb el moment de la presentació de la denúncia o de les víctimes, els detinguts o els sospitosos d'haver comès els fets. Per tant, s'ha previst la seva supressió. També és possible que, pel que fa a les dades temporals s'hagi de filtrar per treballar amb uns terminis de temps concrets.

També un estudi preliminar de les dades ha constatat que els formats de les dades quan es carregaven no sempre corresponia amb el format desitjat. Especialment rellevant era el cas dels valors corresponents a dates i hores, motiu pel qual també era necessària la conversió d'aquests valors per passar-los a un valor de temps.

En relació amb aquestes variables temporals, en funció de l'evolució del projecte també pot ser necessària la creació de nous camps, ja sigui per generar agrupacions pròpies, especialment pel que fa a les dades temporals (com ara la creació de franges horàries, el càlcul del dia de la setmana, el mes de l'any o la creació de variables amb el mes dels fets).

Per últim, també és possible que calgui netejar algunes dades, ja que, per experiència pròpia en el tractament d'aquest tipus d'informació, és possible que els conjunts de dades, incloguin informació valors extrems (*outliers*) ja sigui per errades en la introducció de les dades o per desconeixement de la informació. Dos exemples d'aquests casos poden ser fets amb una data del fet molt antiga, o fets amb una ubicació (latitud i longitud o coordenades geogràfiques) desconeguda o fora de les ciutats objecte de l'estudi.

3.4 Anàlisi

Pel que fa a l'anàlisi de les dades l'objectiu és dissenyar un model predictiu utilitzant xarxes neuronals. Atès que es començarà amb agrupacions de temps i espai grans, s'ha considerat oportú començar amb una xarxes neuronals amb cel·les LSTM. Tal i com expliquen Bosch, Casas i Lozano [28] es tracta d'una tipologia de cel·les adient per problemes en què les seqüències poden tenir dependències entre punts de la seqüència allunyats, per tant, podria ser útil per detectar seqüències patrons que es repeteixen en determinats dies, setmanes o mesos. A més, com s'ha explicat anteriorment, aquesta tipologia ja ha donat bons resultat en investigacions prèvies [24 i 25]. De manera resumida, aquestes cel·les tenen un canal de memòria i un control de flux d'informació a través de les portes d'oblit (que decideix quina part de la informació s'ha de conservar, una porta d'entrada que controla la informació que s'incorpora a la memòria de la xarxa, i una porta de sortida, que decideix la sortida de la xarxa en el pas actual, després de calcular el nou estat que tindrà la xarxa. En funció dels resultats i del nivell de detall, si s'arriba a poder es poden utilitzar altres mos

Puntualment, també es planteja que quan s'aprofundeixi amb l'espai i els tipus de fets, es poden considerar un gran número de sèries. En aquest cas es pot utilitzar un autoencoder per eliminar el soroll de les sèries i després agrupar-les amb un algoritme k-means per, finalment, intentar crear diferents models per a cadascun dels grups de delictes.

3.5 Visualització

La visualització de les prediccions pot tenir un doble format. Per una part, la predicció en les sèries temporals, una visualització amb línies de temps dels diferents conjunts, d'entrenament, de test i les prediccions realitzades. Per una altra part, si s'arriben a realitzar prediccions de punts o àrees amb un major risc, es poden representar en mapes, ja sigui ubicant punts o seleccionant les àrees amb un major risc.

3.6 Publicació

En aquest cas el resultat d'aquest treball final de màster es materialitza amb aquest document, el destinataris principals del qual són el personal docent que l'ha d'avaluar. Aquest mateix document, dintre del marc acadèmic, si es considera oportú, també podria ser publicat en el repositori institucional de la UOC. A més, el codi resultat del treball, s'ha publicat en el repositori GitHub[4] a efectes de poder reproduir la recerca si així es considera oportú. També amb aquesta voluntat, s'ha publicat al repositori Zenodo, tres conjunts de dades corresponents als anys 2018 i 2019[5].

[4] https://github.com/hitnas/TFM_Aprox_Prec_Crime
[5] Santiago Herrero. (2020). NYPD Complaint Data-2018-2019 [Data set]. Zenodo. http://doi.org/10.5281/zenodo.3902827
Santiago Herrero Blanco. (2020). Chicago Crime Data-2018-2019 [Data set]. Zenodo. http://doi.org/10.5281/zenodo.3902623
Santiago Herrero. (2020). Los Angeles Crimes-2018-2019 [Data set]. Zenodo. http://doi.org/10.5281/zenodo.3902627

4. Implementació

Com s'ha explicat anteriorment, en aquest treball s'han utilitzat conjunts de dades de delictes de tres ciutats amb més població dels Estats Units d'Amèrica, Nova York, Chicago i Los Ángeles. Tant l'anàlisi de les dades com un primer estudi de la distribució geogràfica dels fets, així com les primeres proves amb els algoritmes de predicció, s'han fet de manera individual per a cadascuna de les ciutats. Un cop comprovat el funcionament dels algoritmes de predicció, per a cadascuna de les ciutats, en una segona etapa, s'han analitzat els resultats comparant els de les tres ciutats. Per últim, s'ha provat els resultat amb un conjunt de dades un conjunt de dades amb agrupacions En una segona etapa, s'han analitzat els resultats de les tres ciutats. Per últim, s'ha creat un conjunt de dades seleccionant les 100 combinacions de tipus penal i districte, amb més fets a cada ciutat i s'ha comprovat el funcionament de la predicció amb aquest conjunt de dades.

4.1 Estudi de les dades i pre-processat

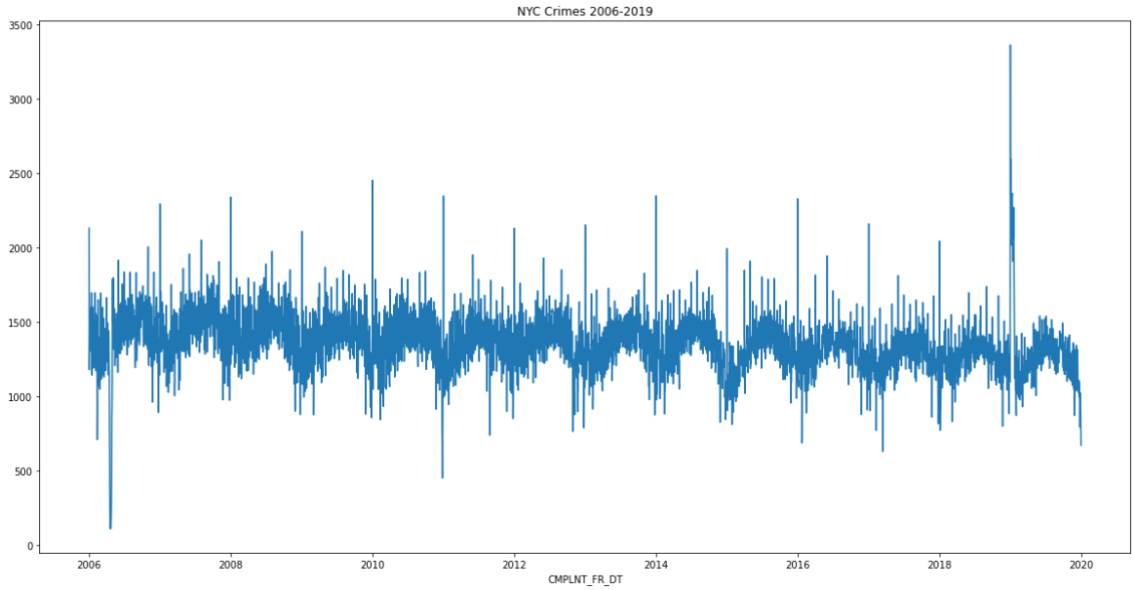
En primer lloc, s'han treballat el conjunt de dades de la ciutat de Nova York, que contenien registres de 2006 a 2019. Hi havia un total de 36 camps (s'adjunta la llista i la descripció a l'Annex). D'aquests 36 camps, se n'han eliminat 10, dos corresponents a la data final dels fets[6] (la data de final i l'hora de final) ja que és l'única de les tres ciutats que té aquesta informació, dos en relació amb la recollida de la denúncia (la data i la comissaria), tres camps amb la informació sobre el sospitós dels fets i tres camps amb informació sobre la víctima.

En referència als camps que tenien informació sobre dies i hores, en importar les dades no quedaven registrats amb un format de data, per tant ha estat necessari transformar-les per tal de poder-les interpretar com a tals. Prèviament, també ha estat necessari canviar alguns registres que constaven amb una hora de *24:00:00* per *00:00:00*. També s'ha detectat que hi havia fets amb una data anteriors a 2006, la data en què inicialment hi havia registres. S'ha considerat que aquests registres podien ser considerats valors extrems (*outliers*) i s'ha decidit eliminar-los.

Un altre àmbit en què s'ha fet una neteja de les dades, eliminant alguns valors ha estat el referent a la ubicació geogràfica dels fets. Ja en la descripció de les dades, la latitud i la longitud presentaven alguns valors extrems, i també s'ha decidit obviar els registres que contenien fets que podien estar ubicats fora de la ciutat de Nova York.

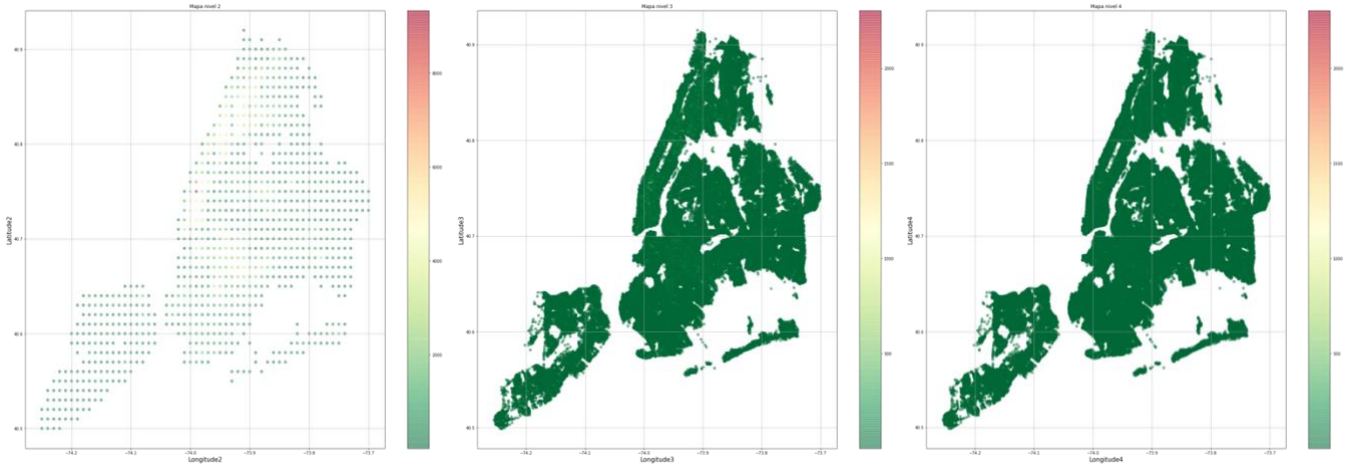
Després d'aquest procés de neteja de les dades, s'ha passat de 7.309.655 a 6.690.245 registres. La seva evolució diària queda representada a la gràfica 1.

[6] Tot i que el més habitual és treballar amb una data dels fets, sovint no es un moment exacte sinó que és un interval, que pot quedar recollit amb una data d'inici del fet i una data de final dels fets. El fet de recollir un interval pot venir donat perquè el delicte ha tingut una durada més o menys llarga en el temps (un atracament pot durar diversos minuts o, fins i tot, hores, o un segrest que pot durar dies), o bé perquè es desconeix el moment exacte en què han passat els fets (un robatori amb força en un habitatge, una empresa o un comerç quan no hi ha ningú en aquell immoble i no hi ha cap alarma que faci que es pugui concretar el moment exacte).



Gràfica 1. Evolució diària dels delictes coneguts a la ciutat de Nova York 2006 2009

L'última transformació realitzada fa referència a les dades de latitud i longitud. Les dades originals tenen sis decimals. Aquesta informació és molt precisa, i pot ser útil en l'estudi en detall de determinats o en l'estudi d'àrees petites, però per una representació del conjunt de la ciutat havia de mostrar 198.349 punts i, en conseqüència no resultava eficient. A més, per a un possible anàlisi territorial podria ser útil establir unitats geogràfiques de diferents dimensions, més enllà de les divisions administratives. Aquest efecte es podia aconseguir reduint els decimals dels camps, i d'aquesta manera es creen agrupacions de punts i es divideix la ciutat en quadrats. S'ha provat amb tres configuracions diferents creant nous camps de latitud i longitud amb 2, 3 i 4 decimals (Gràfica 2) i s'obtenien, respectivament 924, 53.397 i 114.686 punts. Amb aquestes dades s'ha considerat oportú realitzar la mateixa tasca d'arrodoniment de les dades de latitud i longitud de les altres ciutats, però només amb 2 decimals.



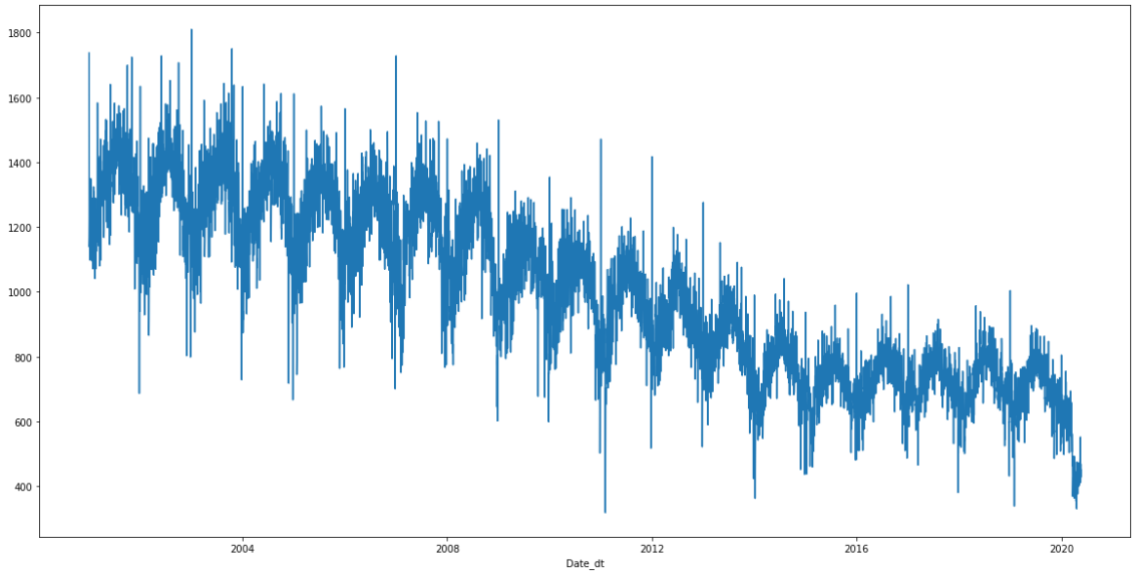
a b c

Gràfica 2. Distribució geogràfica dels delictes coneguts a la ciutat de Nova York per Latitud i longitud (a- 2 decimals, b- 3 decimals, c- 4 decimals).

El segon conjunt de dades que s'ha analitzat ha estat el de la ciutat de Chicago. En aquest cas els registres eren des de 2001 i arriben a 2020[7] (a la ciutat de Nova York el seu portal de dades obertes ofereix les dades recents en un conjunt de dades diferent), i contenen un total de 22 camps (s'adjunta la llista i la descripció a l'Annex). Dels 22 camps, se n'han eliminat tres, un que correspon al número de cas (ja hi ha un camp d'identificació del registre), un altre amb la data d'actualització del registre, i l'últim amb localització ja que conté una tupla amb la latitud i la longitud, informació que ja es troba en uns altres camps.

Inicialment només es fa una transformació en el camp que conté la informació del dia i la hora, que, a més de transformar-lo a format data, se n'extreuen dos camps més, separant dia i hora.

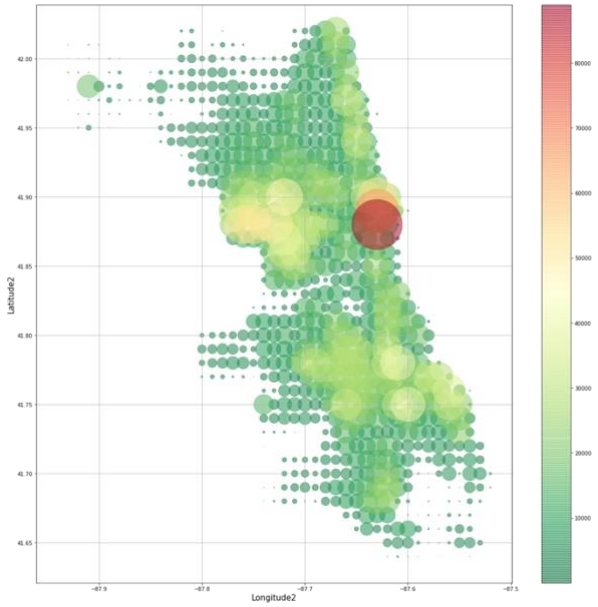
Com a la ciutat de Nova York, s'han netejat les dades de fets que per latitud i longitud queden fora de la ciutat i després de la neteja es passa de 7.122.391 registres a 7.053.723. La seva evolució diària queda representada a la gràfica 3.



Gràfica 3. Evolució diària dels delictes coneguts a la ciutat de Chicago 2001 2020

Com s'ha comentat, s'ha creat els camps de latitud i longitud arrodonida a dos decimals. En aquest cas, s'ha provat alguna altra visualització, de manera que els punts amb més densitat de fets tinguin una mida més gran (Gràfica 4).

[7] Les dades s'actualitzen setmanalment i, per exemple, en l'actualització de 21 de juny de 2020 el registre més recent era del 13 de juny de 2020.

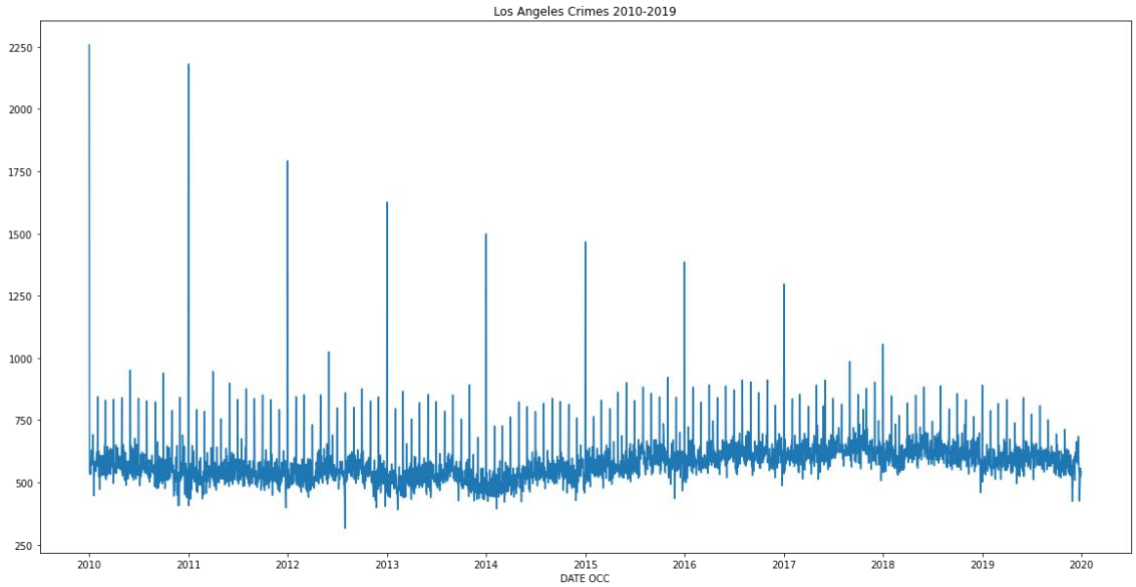


Gràfica 4. Distribució geogràfica dels delictes coneguts a la ciutat de Chicago per latitud i longitud arrodonida a dos decimals.

El tercer conjunt de dades amb què s'ha treballat és el de la ciutat Los Angeles, amb registres comencen l'any 2010 i finalitzen el 2019, i amb informació de 28 camps diferents (s'adjunta la llista i la descripció a l'Annex). Dels 28 camps originals, se n'han eliminat vuit, que corresponen a informació de la víctima (tres), a la data de recollida de la denúncia, al codi i la descripció de l'estat del cas, l'adreça del lloc dels fets i l'encreuament de carrers.

Com amb les dades de les altres ciutats, s'ha realitzat una transformació de les dades referents al moment del fet per passar-los a un format de data i hora, i també s'han eliminat les dades que contenen una informació de latitud i longitud allunyada de la ciutat, tot i que en quest cas el nombre de registres eliminats és molt menor i passa de 2.114.374 a 2.113.469.

Finalment en aquesta ciutat es realitza una transformació addicional, i és que els registres poden tenir fins a quatre tipologies penals diferents (i n'hi ha quatre camps destinats a aquesta funció). S'opta per conservar tota la informació i el que es fa és crear nous registres de manera que consti la informació de tots els fets que s'han conegut. Es creen dos nous camps, un on es recull la informació del tipus penal i un altre per recollir la informació del camp en què estava, i finalment s'eliminen els quatre camps originals. Finalment queden 2.256.204 registres. La seva evolució diària queda representada a la gràfica 4.



Gràfica 5.– Evolució diària dels delictes coneguts a la ciutat de Los Angeles entre 2010 i 2019

A la ciutat de Los Angeles s'observa que els dies 1 de cada mes i els dies 1 de cada any hi ha un gran volum de fets. Molt probablement siguin valors de fets de data desconeguda durant el mes i que s'assignen, per defecte a aquell mes o a aquell any. Aquesta distorsió s'ha anat reduint (especialment pel que fa als valors anuals, al llarg del temps. No obstant això, es manté en els valors mensuals.

4.2 Anàlisi dels fets per ciutats

La primera anàlisi que es va realitzar, era a la ciutat de Nova York. Atès el gran volum de dades, gairebé 6,7 milions de registres, inicialment es va començar a treballar amb una mostra d'aquestes dades, i es van seleccionar aleatòriament el 10% dels registres. Ara bé, seguint les indicacions del director del treball, en el sentit que era millor treballar amb les dades d'anys sencers, es va modificar aquest criteri i es va treballar amb els dos darrers anys complets, els anys 2018 i 2019. A més, en seleccionar els dos anys complets, s'ha pogut tenir les dades del mateix període temporal per a les tres ciutats.

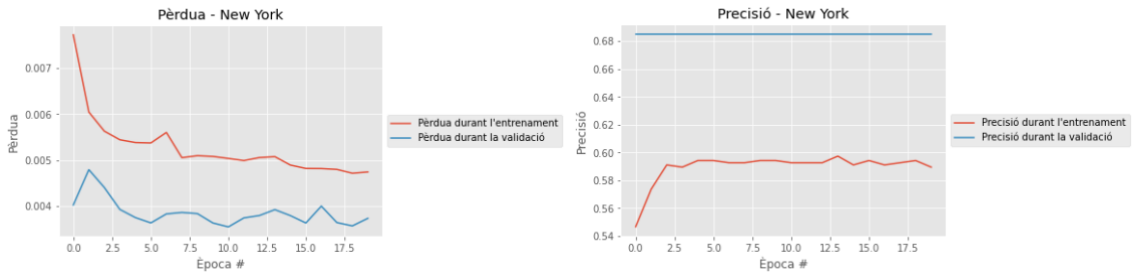
Com s'ha comentat, la primera anàlisi s'ha fet de manera individual per a cada ciutat, intentant realitzar una predicció diària per àrees geogràfiques. Cal tenir en compte que les divisions i subdivisions territorials disponibles tenien algunes diferències entre les tres ciutats. Per una primera anàlisi era poc operatiu treballar amb les unitats territorials més petites. Inicialment es va fer alguna prova amb la ciutat de Nova York amb els 6 districtes, però a les ciutats de Chicago i Los Angeles les divisions territorials més grans (districtes i àrees), n'eren 23 i 21, respectivament. Finalment, s'ha seleccionat els districtes policials, que en són 8 àrees, ja que la següent divisió, les comissaries, n'eren 78, una xifra ja excessiva per aquesta anàlisi inicial.

Cal reconèixer, també, que no s'inclouen resultats de les primeres proves realitzades, ja que malgrat que es van obtenir molt bons resultats, aquests eren un miratge, conseqüència d'una incorrecta separació dels conjunts d'entrenament i validació, ja que s'havien incorporat els valors de validació dins del conjunt d'entrenament.

Un cop esmenat l'error es van entrenar la xarxa amb les següents opcions, obtenint resultats no gaire bons pel que fa la predicció.

Com s'ha comentat anteriorment, les xarxes neuronals amb xarxes LSTM (*Long short term memory*) són adients per realitzar prediccions en sèries temporals.

En primer lloc s'ha utilitzat una arquitectura no gaire complicada, amb dues capes LSTM, cadascuna d'elles amb 200 unitats. L'entrenament s'ha realitzat amb 20 èpoques i amb una finestra de temps de 7 dies.

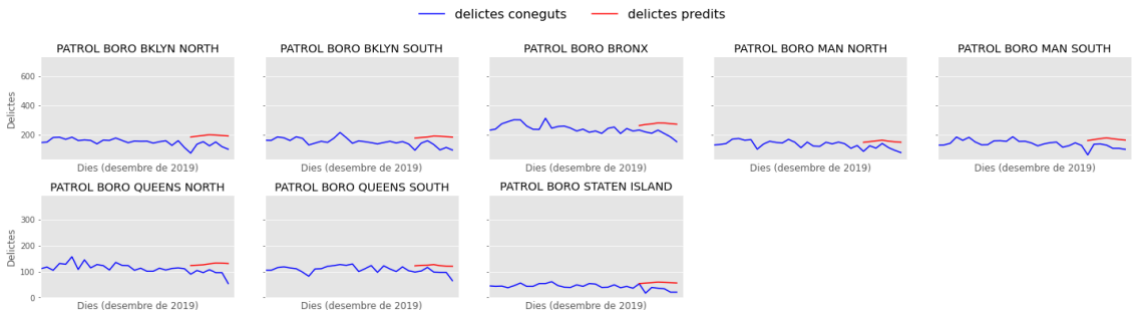


a b

Gràfica 6.Resultats de l'entrenament amb la primera arquitectura a la ciutat de Nova York, pèrdua (a) i precisió (b)

Els primers resultats han estat una mica sorprenents. Els primers valors, corresponents a la pèrdua (gràfica 6.a) tant en l'entrenament com en la validació poden ser esperables, amb una lleugera reducció de la pèrdua més accentuada. En la precisió (gràfica 6.b) durant l'entrenament ha millorat en les primeres dues èpoques fins arribar a un valor proper al 60%, però s'ha estabilitzat amb lleugeres variacions, sense arribar, en cap moment a superar aquest 60%. Ara bé, la precisió durant la validació ha estat constant, en les 20 èpoques, amb un valor de 0,6852. Encara que aquest valor no és òptim, podria ser un bon començament, però el fet que sigui un valor constant introdueix dubtes sobre la seva validesa.

A més, quan s'analitza la predicció amb els valors reals (gràfica 7), s'observa, per una part, que la predicció és lleugerament elevada respecte els valors reals, i, per l'altra, que l'evolució de la predicció és basant estable, respecte els valors reals. Aquesta desviació a l'alça, pot estar condicionada per uns valors atípics a començament de l'any 2019.

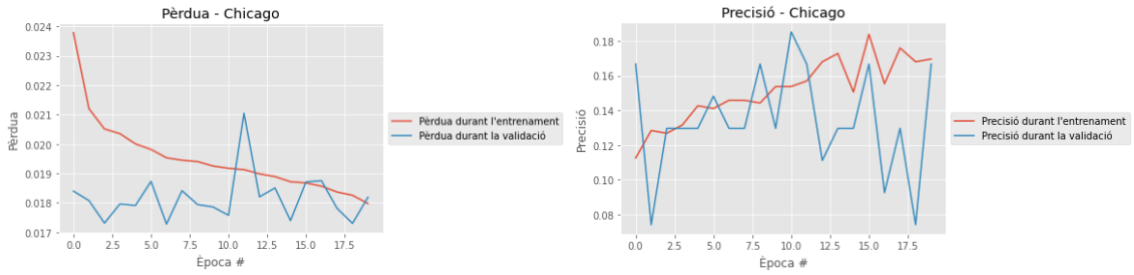


Gràfica 7. Prediccions amb l'entrenament la primera arquitectura a la ciutat de Nova York, pèrdua durant l'entrenament

Per valorar aquestes prediccions, s'ha utilitzat la mètrica de l'arrel quadrada de l'error quadràtic mitjà (root mean square error, RMSE), que ha estat de 22,54. És a dir, que les prediccions s'han equivocat, en 22,54 fets, quan la mitjana de fets per districte oscil·la entre els 280 dels Bronx i els 55 d'Staten Island.

La mateixa arquitectura s'ha utilitzat per entrenar sengles xarxes neuronals també a les ciutats de Chicago i Los Angeles. Malgrat que en aquests dos casos, els resultats de precisió han estat molt més baixos, l'evolució de les dades ha estat més natural.

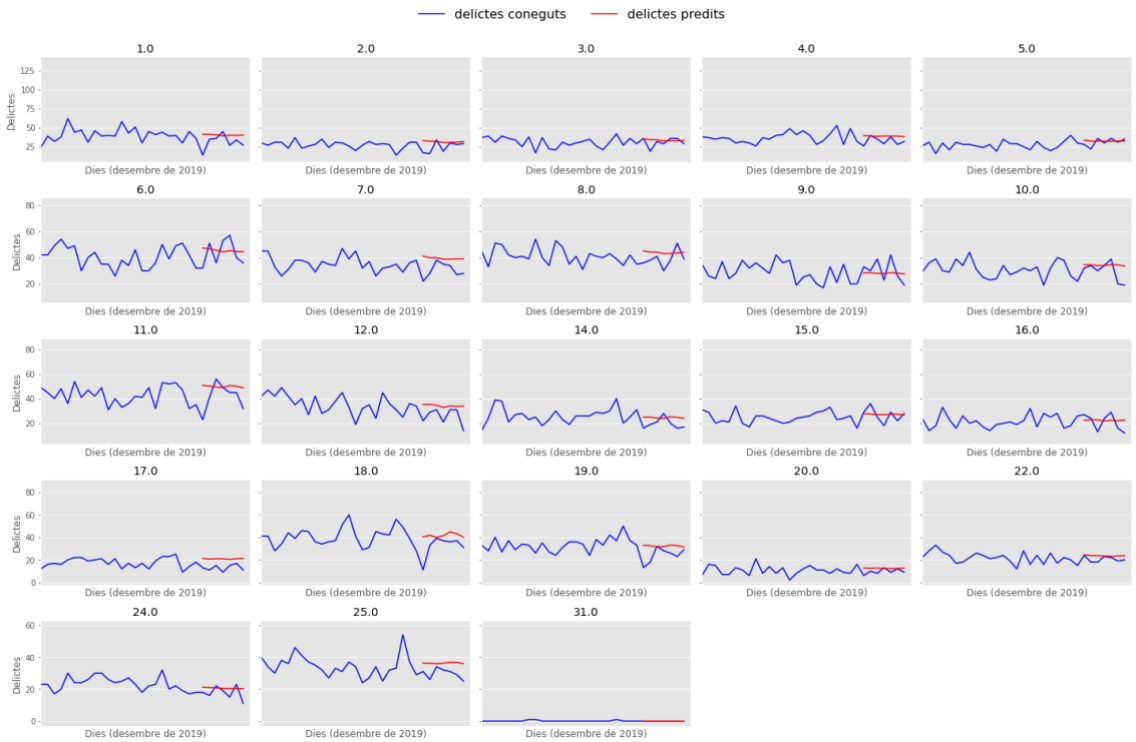
Així, en per a la ciutat de Chicago, la pèrdua durant l'entrenament ha baixat de manera constant, la qual cosa sembla indicar que la xarxa neuronal seguia un procés d'aprenentatge correcte (gràfica 8.a). Pel que fa a la precisió (gràfica 8.b), durant l'entrenament augmentava de manera bastant contínua, malgrat que durant la validació aquesta evolució era molt més irregular, i els valors no superaven el 20%.



a b

Gràfica 8. Resultats de l'entrenament amb la primera arquitectura a la ciutat de Chicago, pèrdua (a) i precisió (b)

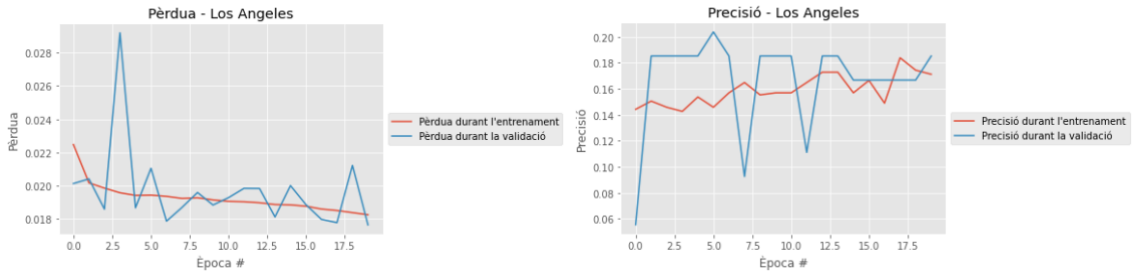
Si observem les prediccions realitzades amb aquest model, podem comprovar que malgrat els resultats de la validació són pitjors que per a la ciutat de Nova York, les prediccions semblen més properes als valors reals. Ara bé, malgrat que a la majoria dels districtes els valors predits estan són molt més propers als valors reals, se segueix donant el mateix problema que a la ciutat de Nova York i és que la predicció per als 7 dies és bastant estable per a cadascun dels districtes i sembla que no s'adapta a les fluctuacions diàries.



Gràfica 9. Prediccions amb l'entrenament de la primera arquitectura a la ciutat de Chicago

En aquest cas la RMSE ha estat de 12,53, i els valors mitjans dels districtes oscil·len entre els 50,84 del districte 11, i els 12,55 del districte 20 (sense comptar les dades del districte 31 (que té una mitjana de 0,02, i que recull valors només alguns dies).

Les mètriques d'entrenament del model a la ciutat de Los Angeles, presenten uns valors bastant semblants a les de la ciutat de Chicago. Així, la pèrdua durant la validació (gràfica 10.a) es va reduint de manera progressiva en totes les èpoques i la precisió (gràfica 10.b) també amb bastantes oscil·lacions en la validació i, malgrat que a l'època 6 supera el 20%, més endavant baixa, fins i tot per sota del 10%) i amb algunes oscil·lacions, a partir de l'època 15, s'estabilitza al voltant del 17% i acaba repuntant lleugerament en l'última època.

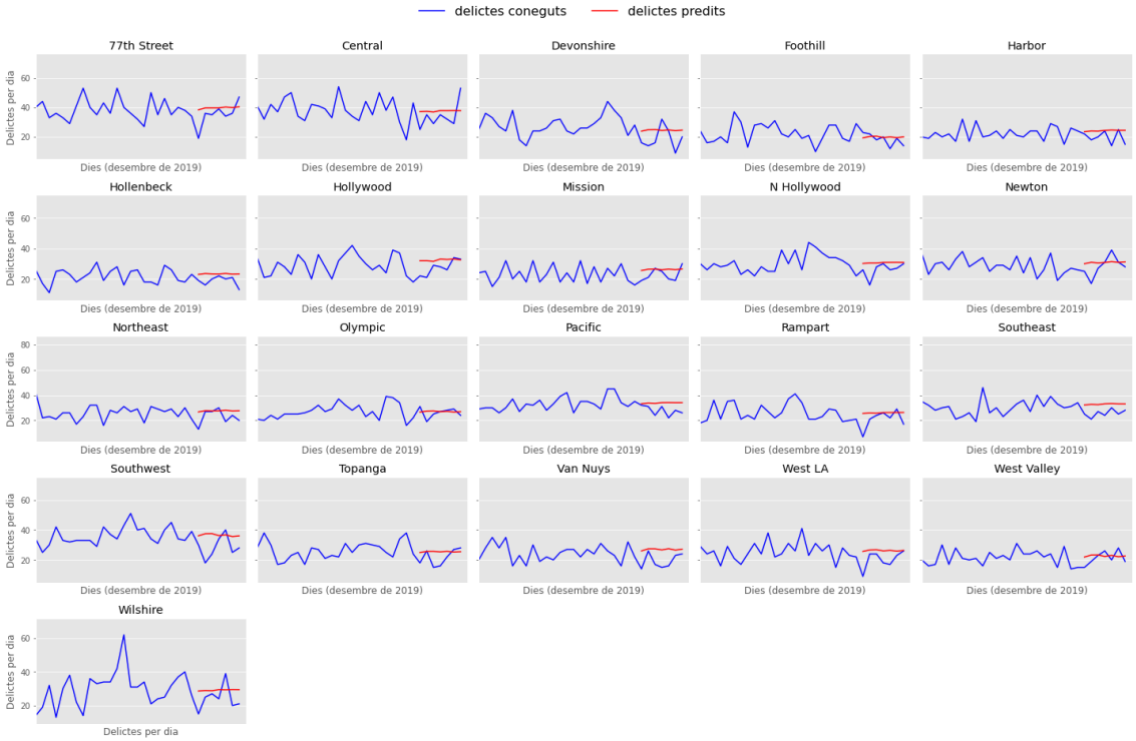


a b

Gràfica 10. Resultats de l'entrenament amb la primera arquitectura a la ciutat de Los Angeles, pèrdua (a) i precisió (b)

En analitzar la predicció del model pels darrers set dies de l'any, com a les altres dues ciutats, s'observa que la predicció es bastant estable, i no presenta

fluctuacions accentuades en els dies (com passa amb les xifres reals durant tot l'any). Ara bé, els valors predits són més aproximats que a la predicció de Nova York, i, fins i tot, sembla també que siguin lleugerament millors que als de la predicció de Chicago.



Gràfica 11. Prediccions amb l'entrenament de la primera arquitectura a la ciutat de Los Angeles

Aquesta última afirmació sembla confirmar-se, ja que la RMSE per a aquesta predicció de Los Angeles és de 5.04, amb uns valors de mitjana als districtes de la ciutat que fluctuen entre els 39,74 del districte 77th Street i els 21,34 de Foothill.

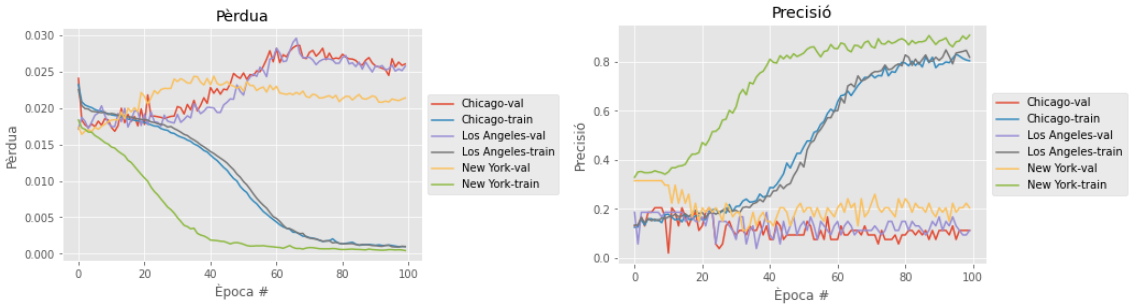
Els resultats, en ser tan dispars, han estat poc concloents. Per tant, s'ha considerat oportú seguir provant aquesta estructura amb altres configuracions, augmentant el número d'èpoques i provant amb una finestra de temps més gran, per poder valorar de manera adequada si resultava una estructura que pogués arribar a ser útil.

La segona arquitectura que s'ha entrenat, ha estat amb la mateixa estructura i configuració que el primer model, però augmentant el nombre d'èpoques, de 20 a 100. Recordem que constava de dues capes LSTM amb 200 cel·les cadascuna, i que la finestra de temps era de 7 dies. S'ha entrenat un model diferent per a cadascuna de les ciutats.

En aquest cas, la pèrdua durant l'entrenament es comporta de la manera esperada amb una evolució molt semblant a les ciutats de Los Angeles i Chicago, i amb un descens més accentuat a la ciutat de Nova York (gràfica 12.a). Ara bé, la pèrdua en la validació, es comporta d'una manera completament diferent, i,

gairebé inversa a la validació, ja que a la ciutat de Nova York augmenta la pèrdua en les 30 primeres èpoques, i després inicia un lleuger descens, encara que sense arriar als valors de les primeres èpoques. A les ciutats de Chicago i Los Angeles, l'evolució és bastant semblant, i comença amb un augment lleuger, que s'accentua entre les èpoques 30 i 60 (aproximadament) i s'estabilitza a la baixa a partir d'aquell moment.

Pel que fa a la precisió (gràfica 12.b), igual que amb la pèrdua, mentre els valors d'entrenament indiquen un aprenentatge, més ràpid a la ciutat de Nova York que a les altres dues ciutats, les dades de validació apunten en sentit contrari, de manera que no sembla que el model aprengui, amb valors de precisió molt baixos (al voltant del 20% per a la ciutat de Nova York i entre el 15 i el 20% per a les altres dues ciutats. Hi ha moltes oscil·lacions entre èpoques i després d'un primer descens en les primeres 30 èpoques els valors s'estabilitzen.

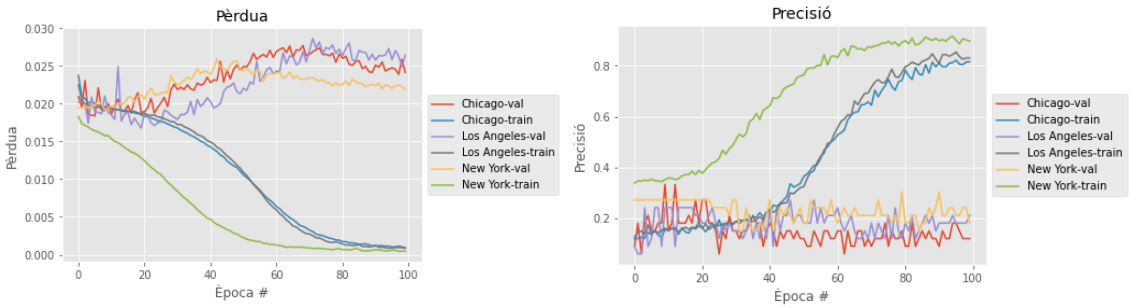


a b

Gràfica 12. Resultats de l'entrenament amb la segona arquitectura a la ciutat de Los Angeles, pèrdua (a) i precisió (b)

El fet que l'evolució de les dades de Nova York tinguin una evolució tan diferent a les de Chicago i Los Angeles, podria estar condicionat pel menor número de característiques del model, que en aquest cas ve representada per les àrees territorials, 8 en el cas de Nova York, 23 i 21, respectivament, en el els casos de Chicago i Los Angeles.

La mateixa arquitectura, de 2 capes LSTM amb 200 cel·les, entrenades durant 2 èpoques, s'ha utilitzat per provar una finestra de temps diferent, de 28 dies. S'ha entrenat un model amb aquesta arquitectura per a cadascuna de les ciutats. Els resultats obtinguts a la gràfica 13 mostren uns resultats i un comportament molt semblants al de 7 dies. La diferència més significativa podria ser una lleugera millora en el resultat de la precisió en la validació per a la ciutat de Chicago, amb valors més propers al 20%, que segueixen sent baixos.



a b

Gràfica 13. Resultats de l'entrenament amb la tercera arquitectura a la ciutat de Los Angeles, pèrdua (a) i precisió (b)

4.3 Anàlisi per principals combinacions de tipus de fet i àrees geogràfiques

Com a anàlisi complementària, s'ha comprovat quins eren els resultats d'aplicar un model a un conjunt de dades que integrés registres de les tres ciutats. En primer lloc, per a cadascuna de les ciutats s'ha creat una taula en què es mostraven els fets diaris per a cada tipologia de fets penals en una àrea geogràfica concreta.

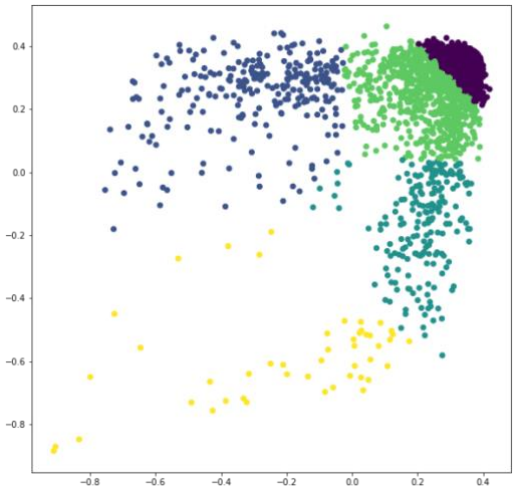
Les combinacions escollides han estat les següents:

- A Nova York, s'ha seleccionat la comissaria del lloc on han passat els fets (ADDR_PCT_CD) i el codi del fet (KY_CD.
- A Chicago, s'ha triat la comunitat (Community Area) i els principals tipus de fet (Primary Type).
- A Los Angeles, s'ha escollit l'àrea policial (AREA NAME), i el codi del delicte després de des agrupar els possibles fets secundaris (Crm Cd n).

Entre les tres ciutats s'han obtingut un total de les 8.031 combinacions d'àrea i tipus de fet, que corresponen cadascuna a una sèrie temporal.

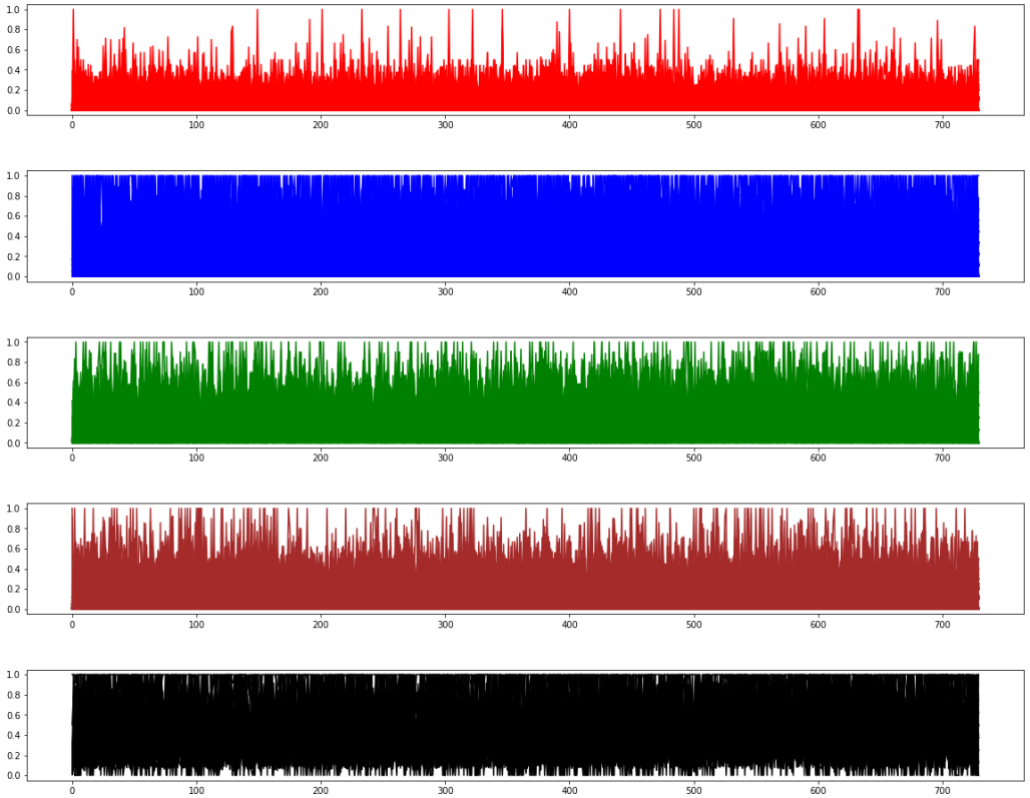
Sembla lògic que no totes aquestes combinacions hauran tingut la mateixa evolució durant els darrers dos anys. Amb aquesta premissa s'ha volgut classificar aquestes combinacions de fet i àrea en cinc categories diferents. Per a crear aquestes categories, s'ha utilitzat un autoeencoder amb dues neurones centrals. Els valors d'aquestes dues neurones per a cada sèrie s'han utilitzat per agrupar les categories amb un algoritme k-means[8] (gràfica 14).

[8] Aquest mètode es va utilitzar en una pràctica de l'assignatura Models Avançats de Mineria de Dades, d'aquest màster.



Gràfica 14. Agrupació de les 8.031 sèries amb l'algoritme *k-means*

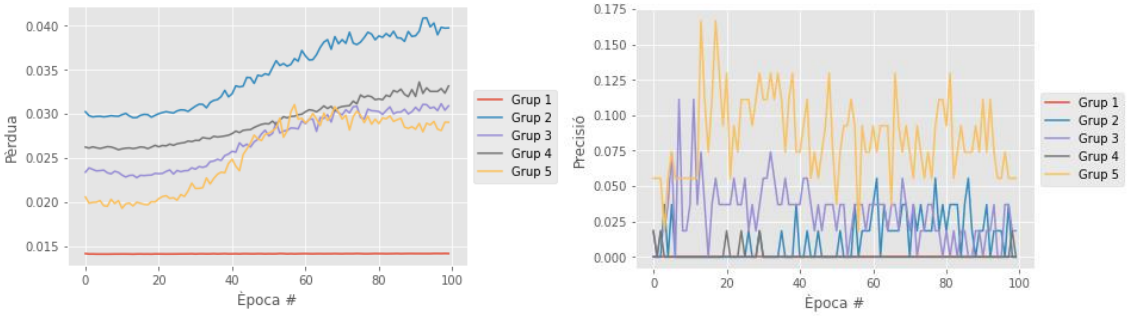
Tot i que per la quantitat de sèries de cada grup i la quantitat de dades resulta difícil comprovar l'homogeneïtat dels grups, a la representació gràfica de les sèries de cada grup (gràfica 15) s'observen algunes diferències entre els diferents grups creats.



Gràfica 15. Representació dels valors normalitzats de les sèries de les 5 agrupacions obtingudes amb l'algoritme *k-means*

Val a dir, també, que els grups tampoc han estat homogenis pel que fa a la quantitat de sèries en cadascun. Els grups tenien 6.805, 251, 203, 725 i 47 sèries.

A continuació, s'ha utilitzat la segona de les arquitectures utilitzades en l'apartat 4.2 (2 capes LSTM amb 200 unitats, durant 100 èpoques i una finestra de temps de 7 dies) per entrenar un model per a cadascun d'aquests grups.

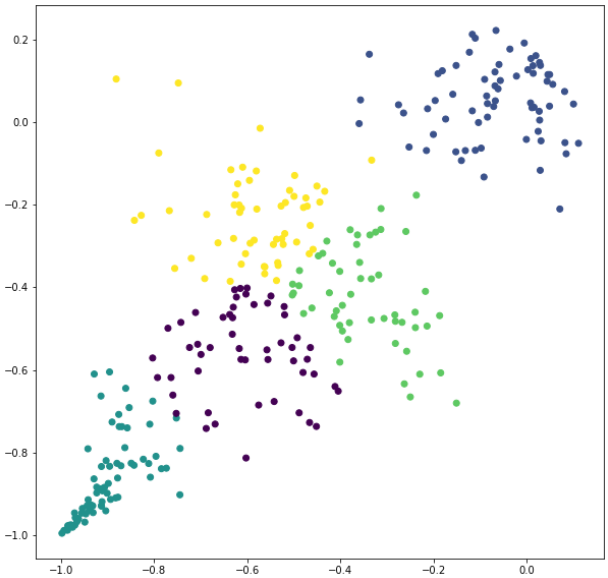


a b

Gràfica 16.Resultats de pèrdua (a) i validació (b) amb la segona arquitectura als diferents grups obtinguts amb l'algoritme *k-means* per a les 8.031 sèries

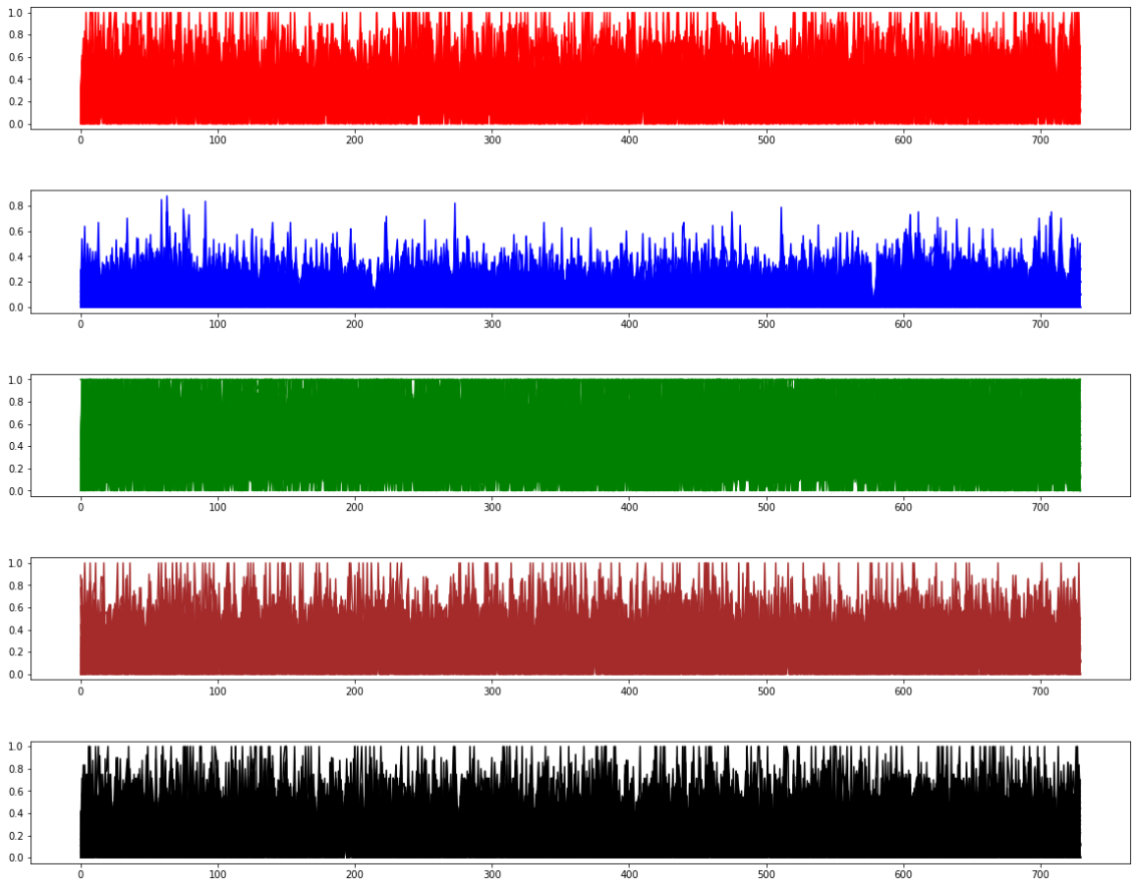
Encara que, novament els valors de pèrdua i precisió en la validació per als diferents models no son gaire bons, sí que s'observen diferències substancials en el comportament dels diferents grups. Especialment rellevant és que el grup més nombrós té una pèrdua constant i cap precisió en la predicció, probablement per la quantitat de sèries que incorpora. Ara bé caldria analitzar si les diferències en la precisió en la resta de grups són fruit de la pròpia distribució dels grups o bé si hi ha una certa homogeneïtat interna en la precisió en les sèries dintre dels diferents grups.

Tenint en compte el gran volum de sèries amb què es treballava, més de 8.000, s'ha fet una prova semblant, però seleccionant només les 100 sèries de cadascuna de les tres ciutats, amb la hipòtesi que sèries amb un major volum de dades poden realitzar unes prediccions més consistents.

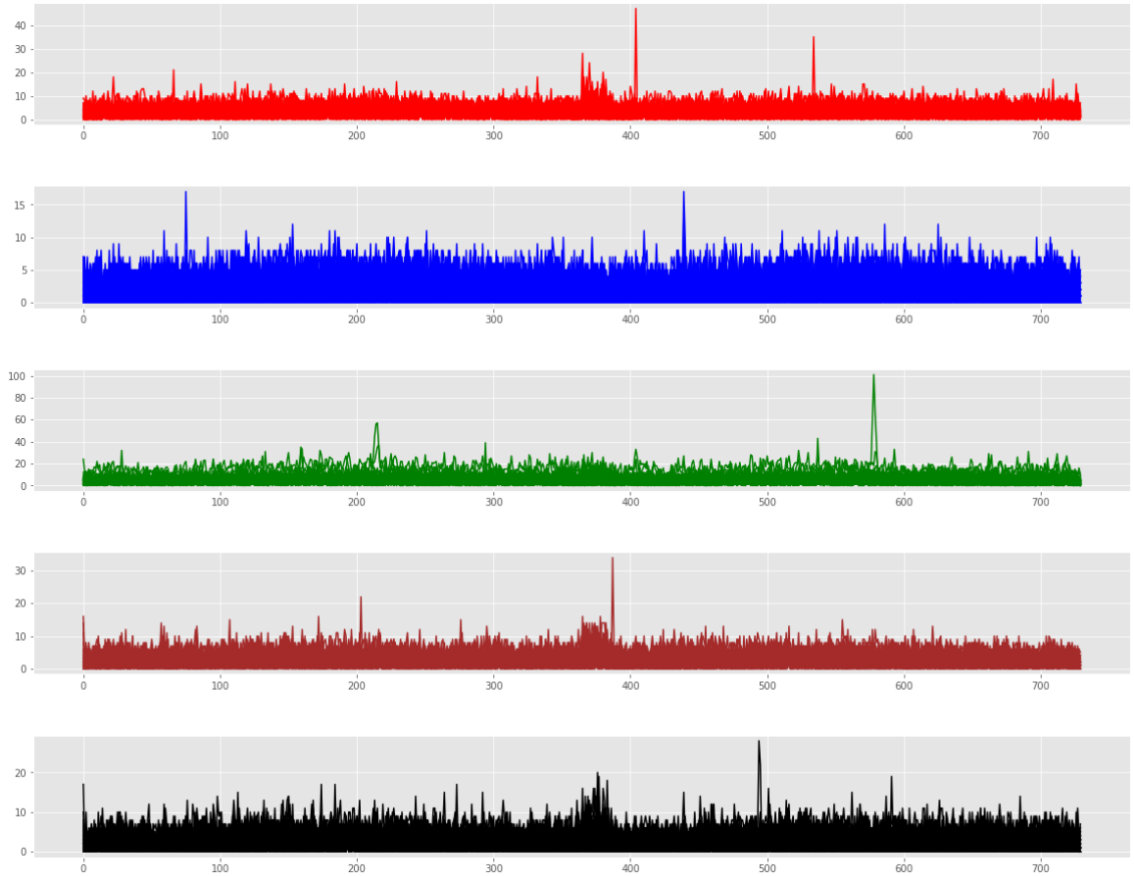


Gràfica 17. Agrupació de les 300 sèries seleccionades amb l'algoritme *k-means*

El resultat de l'agrupament de les sèries (aplicant un autoeencoder amb dues neurones centrals i utilitzant els els valors d'aquestes dues neurones per a cada sèrie per agrupar les categories amb un algoritme k-means), han resultat amb uns grups bastant més equilibrats que per al total de les sèries. Així els grups creats tenien 38, 70, 76, 53 i 63 sèries.



Gràfica 18.Representació dels valors normalitzats les 300 sèries seleccionades en les 5 agrupacions obtingudes amb l'algoritme *k-means*

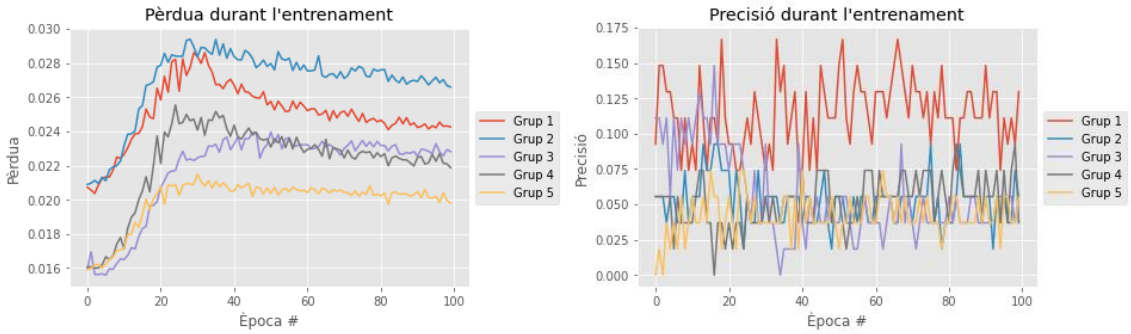


Gràfica 19. Representació dels valors de les 300 sèries seleccionades en les 5 agrupacions obtingudes amb l'algoritme *k-means*

A les gràfiques 18 i 19 es poden observar els les sèries de temps, en valors normalitzat i en valors reals en els diferents grups. Es poden comprovar les diferències en les sèries, no només per la diferent quantitat de registres diaris (gràfica 19), sinó també per una major regularitat o irregularitat en l'evolució de les sèries (gràfiques 18 i 19).

Finalment, s'ha utilitzat, un cop més, la segona arquitectura proposada en l'apartat 4.2 (2 capes LSTM amb 200 unitats, durant 100 èpoques i una finestra de temps de 7 dies) per entrenar un model per a cadascun d'aquests grups.

Encara que, com es pot observar a la gràfica 20 els resultats segueixen sense ser òptims, s'observa unes diferències respecte els resultats del total de les sèries. Per una part, els valors de pèrdua durant la validació (gràfica 20.a) mostren que la pèrdua augmenta de manera considerable, per a totes les sèries en les primeres 20-25 etapes i, a partir d'aquest moment inicia una millora, tot i que molt suau fins al final de l'entrenament. Per una altra part, els valors de la precisió segueixen sense ser gaire alts, però hi ha dues diferències destacades. La primera és que tots els grups tenen una precisió encara que baixa (recordem que en un dels grups del punt anterior la precisió era de 0). La segona és que un dels grups té valors lleugerament superiors a la resta, i presenta valors gairebé sempre superiors al 12,5, mentre en l'anterior cap dels grups superava, de manera constant aquest valor.



a b

Gràfica 20. Resultats de pèrdua (a) i validació (b) amb la segona arquitectura als diferents grups obtinguts amb l'algoritme k-means per al conjunt de les 300 sèries seleccionades.

4.4 Resum de resultats

Configuració	Ciutat / Grup	Pèrdua validació	Precisió validació
20 èpoques, input = 7	Nova York	0,0037	0,6852
	Chicago	0,0179	0,2222
	Los Angeles	0,0117	0,1852
100 èpoques, input = 7	Nova York	0,0201	0,1852
	Chicago	0,0257	0,0185
	Los Angeles	0,0248	0,1296
100 èpoques, input = 28	Nova York	0,0231	0,2727
	Chicago	0,0255	0,0909
	Los Angeles	0,0275	0,0606
100 èpoques, input = 7	Grup 1 (6.805/8.031)	0,0141	0,00E+00
	Grup 2 (251/8.031)	0,0398	0,00E+00
	Grup 3 (203/8.031)	0,0309	0,0185
	Grup 4 (725/8.031)	0,0332	0,00E+00
	Grup 5 (47/8.031)	0,029	0,0556
100 èpoques, input = 7	Grup 1 (38/300)	0,0243	0,1296
	Grup 2 (70/300)	0,0266	0,0556
	Grup 3 (76/300)	0,0228	0,0556
	Grup 4 (53/300)	0,0219	0,037
	Grup 5 (63/300)	0,0198	0,0556

Taula 1. Resum de resultats de les diferents arquitectures provades.

Els resultats obtinguts en les diferents configuracions, no són gaire òptims pel que fa a la precisió. El valor màxim obtingut ha estat una 0,6852 per a la ciutat de Nova York, amb una configuració de 20 èpoques i un input de 7. Ara bé, tal i com s'observa a la gràfica 6, el fet que aquesta precisió sigui constant durant les 20 èpoques resulta estrany, més encara tenint en compte la gran diferència

amb els resultats de les altres dues ciutats, amb precisions de 0,2222 i de 0,1852.

Ara bé, comparant els resultats de les tres configuracions aplicades a les ciutats, aquesta primera ha estat la que ha donat millors resultats per a les tres ciutats. El millor comportament amb un menor nombre d'èpoques s'observa també amb una major pèrdua, que, com s'observa a les gràfiques 12.a i 13.a augmentava amb les dades de validació a patir de les 20-25 èpoques.

Entre els resultat de les dues configuracions amb 100 èpoques per les ciutats, sí que s'observa un millor resultat de la que té una finestra de temps més àmplia, si més no per les ciutats de Nova York i Chicago. A la ciutat de Los Angeles no passa això, però probablement, la incidència dels valors extrems dels dies 1 de cada més, pot afectar, en major mesura aquesta finestra de temps més àmplia.

Per últim, entre els resultats de les dues proves de grups generats amb *autoencoder* i *k-means*, la precisió del primer conjunt és gairebé nul·la, molt probablement perquè la gran quantitat d'elements en cada grup provoca una gran heterogeneïtat dintre dels grups. En el segon cas, amb una selecció de les 100 combinacions de tipus penal i àrea geogràfica amb més fets, s'observa un grup on s'ha produït un major aprenentatge. Ara bé, l'increment de la pèrdua les gràfiques 16.a i 20.a, poden indicar que la configuració tenia un excessiu nombre d'èpoques.

5. Conclusions

En primer lloc cal destacar que el treball ha servit per confirmar una de les màximes en l'àmbit de la ciència de dades i és que la preparació de les dades és una de les tasques que comporta una major quantitat de temps. A més d'aquest temps dedicat a la preparació de les dades, la utilització de xarxes neuronals profundes (encara que amb poques capes) també ha comportat una gran inversió de temps en l'execució (malgrat que moltes proves s'han fet amb poques èpoques o amb un valuós recurs com és Google Colab).

No obstant això, i com acostuma a passar, els majors aprenentatges són fruit dels errors comesos, alguns dels quals no han pogut ser esmenats. Penso que els resultats poc rellevant dels algoritmes utilitzats poden són conseqüència de tres factors principals.

El primer té l'origen en la matèria primera del treball. Com es pot observar a les gràfiques 1, 3 i 5, les dades de totes tres ciutats contenen valors atípics, especialment concentrats en els dies 1 de cada mes i, puntualment, en el dia 1 de cada any. Més enllà d'un possible increment de fets vinculats a aquests dies (festes de cap d'any, possible cobraments de diners, etc.), i que es podria tenir en compte introduint una variable per esdeveniments que passin determinats dies (festivitats, grans esdeveniments populars, etc.) sovint els dies 1 s'utilitzen com a valor utilitzat com a data dels fets en què es desconeix quan han passat els fets, aquesta circumstància distorsiona els valors reals i caldria tractar d'alguna manera aquests increments "irreals" de fets en determinats dies pe tal d'obtenir un resultat predictiu més acurat.

El segon està relacionat amb el fet de no trobar una arquitectura i una configuració de paràmetres adequada. Les xarxes LSTM s'han demostrat útils per a aquest tipus de problemes, no obstant això, la manca de temps i, probablement, una tria inadequada dels paràmetres, no han permès obtenir uns millors resultats.

El tercer error pot haver estat també un error en la tria de les mètriques. Els resultats de la precisió i de la RMSE han resultat contradictoris. Així, la ciutat de Nova York, amb uns resultats de precisió en la validació força alts, penalitzaven més en la RMSE, metre les altres dues ciutats, amb valors més baixos de precisió en la validació tenien resultats millors en la RMSE. Les gràfiques 7, 9 i 11, sembla que corroboren una major validesa de la RMSE respecte la precisió en la validació.

Amb aquestes i altres qüestions, malgrat que s'ha desenvolupat un algoritme amb voluntat predictiva, els resultats no han estat gaire efectius i, en conseqüència considero que el grau d'assoliment de l'objectiu principal ha estat baix. Com a conseqüència de no haver arribat a uns millors resultats amb aquest objectiu principal, tampoc no s'ha pogut experimentar amb finestres d'espai i

temps que ofereixin estimacions rellevants i considero que els resultats de la prova amb 7 o 28 dies no són concloents.
L'anàlisi per principals combinacions de tipus de fet i àrees geogràfiques pretenia aportar respostes a algunes de les hipòtesis plantejades com a objectius secundaris, especialment si els diferents models podien ser òptims per diferents subtipus d'algunes tipologies delictives o si per a algunes tipologies delictives no hi ha models predictius. Els resultats són poc concloents i, per tant, amb les dades disponibles no es poden confirmar ni refutar aquestes hipòtesis.

Una qüestió no esmentada fins ara, i que pot haver estat un factor important per no assolir dels objectius ha estat la impossibilitat de seguir la planificació inicialment prevista. El motiu és que la situació atípica viscuda durant aquest semestre, com a conseqüència de la crisi sanitària i la declaració de l'estat d'alarma ha fet que pugui destinar menys temps a la realització d'aquest treball. Afortunadament, no he tingut problemes de salut (tampoc ningú del meu entorn els ha tingut). Ara bé, en l'àmbit professional, per una part he hagut de teletreballar, la qual cosa ha suposat un canvi d'hàbits, i també algun encàrrec laboral addicional han suposat que tingués menys temps disponible per als estudis. Per una altra part, l'adequació a la nova situació i la preocupació per la situació també dificultaven la concentració en el poc temps disponible. En algun moment, fins i tot, havia pensat en no finalitzar el treball, però gràcies al temps addicional que se'ns ha facilitat, i als ànims del meu director, he pogut acabar presentant aquest treball. Tot i així, les urgències per poder finalitzar-lo han suposat treballar d'una manera menys metòdica del que seria necessari per un treball d'aquestes característiques.

En conseqüència, queden obertes moltes línies de futur, de les quals només s'esmenten, breument, algunes:

- Analitzar com reduir el biaix que poden suposar els valors atípics del dia 1. Dues possibles solucions serien fer una estimació dels valors del dia i eliminar l'excés o distribuir els fets d'aquell dia entre l'espai de temps del qual s'hi ha incorporat dades.

- Incorporar variables basades en el calendari de manera que es tinguin en compte esdeveniments que passen en moments puntuals.

- Aprofundir en l'anàlisi territorial dels fets. Únicament s'ha aprofitat les agrupacions territorials, però caldria analitzar si altres segmentacions territorials són més útils.

- Provar configuracions de xarxes neuronals més profundes per obtenir uns millors resultats predictius. Amb més disponibilitat de temps es podran provar altres configuracions sense haver de patir per la dilació en els temps d'execució.

- Analitzar dades de més de 2 anys. Per comprovar si una major quantitat de dades pot ajudar a millorar la predicció o no. Novament el fet de disposar de més temps, permetrà fer proves encara que requereixin una major quantitat de temps.

- Dades de Barcelona o Catalunya. En el cas d'obtenir algun model que obtingui valors predictius alts, podria intentar provar-lo amb dades de l'entorn proper amb les quals treballo professionalment.

6. Glossari

Autoencoder. Tipus de xarxa neuronal completament connectada, en què les capes intermèdies acostumen a tenir tenen menys neurones que les capes d'entrada i de sortida. Tenen una estructura simètrica, de manera que la sortida intenta replicar els valors d'entrada. Entre d'altres funcions, s'utilitzen per reduir la dimensionalitat dels conjunts de dades [28].

Hot spot (punt calent). Es tracta d'una zona o una àrea territorial relativament petita (places, parcs, cruïlles, segments de carrers, illes de cases, etc.) on es produeix una elevada concentració de fets delictius.

K-means. Mètode d'agrupació de conjunts de dades, basat en la classificació supervisada i en què es divideix el conjunt de dades en *k* grups, i cada element es classifica en el grup al quan hi ha una menor distància [29].

LSTM (*long short term memory*). Arquitectura de cel·la per a xarxes neuronals, proposada per Hochreiter y Schmidhuber el 1997, en què les entrades estan formades pel valor de la seqüència en el pas corresponent i la concatenació de l'estat i la sortida de la xarxa en el pas anterior [28].

RMSE (*root mean square error*). Arrel quadrada de l'error quadràtic mig. Mètrica utilitzada per a l'avaluació del rendiment de models de regressió.

7. Bibliografia

1. Dick, Philip K. The Minority Report. Pantheon, 2002
2. Fergusson Andrew G. Policing Predictive Policing, 94 Wash. U. L. Rev. 1109 (2017). https://openscholarship.wustl.edu/law_lawreview/vol94/iss5/5
3. Hunt, Joel. From Crime Mapping to Crime Forecasting: the Evolution of Place-Based Policing. July 10, 2019, nij.ojp.gov: https://nij.ojp.gov/topics/articles/crime-mapping-crime-forecasting-evolution-place-based-policing
4. Farrell, G. and W. Sousa (2001) Repeat Victimization and Hot Spots: The Overlap and Its Implications for Crime Control Crime Prevention Studies, volume 12, pp. 221-240
5. Cohen L.E., Felson, M. (1979). “Social change and crime rate trends: A routine activity approach”. American Sociological Review, 44, 588-608. Republicat a Andersen, M. A.; Brantingham, P. J., i Kinney, J. B. (Eds.) (2010). Classics in Environmental Criminology. Burnaby: Simon Fraser University Publications.
6. Brantingham, P. L. i Brantingham , P. J. (1993) “Environment, Routine and Situation. Toward a Pattern Theory of Crime”. Advances in Criminological Theroy, 5, 259-294. Republicat a Andersen, M. A.; Brantingham, P. J., i Kinney, J. B. (Eds.) (2010). Classics in Environmental Criminology. Burnaby: Simon Fraser University Publications.
7. Mayer-Schönberger, Viktor i Cukier, Kenneth. Big data. La revolucíon de los datos masivos. Turner Publicaciones, Madrid, 2013
8. Brantingham, P. Jeffrey. The Logic of Data Bias and Its Impact on Place-Based Predictive Policing. Ohio State Journal Of Criminal Law. Vol. 15 (2018) http://paleo.sscnet.ucla.edu/Brantingham-2018-OSJCL.pdf
9. Fitzpatrick, Dylan; Gorr, Wilpen L; Neill, Daniel B. Keeping Score: predictive Analytics in Policing. Annual Review of Criminology. 2009. 02:7. https://www.annualreviews.org/doi/10.1146/annurev-criminol-011518-024534
10. Quijano-Sánchez, Lara; Liberatore, Federico; Camacho-Collados, José; Camacho-Collados, Miguel. Applying automatic text-based detection of deceptive language to police reports: Extracting behavioral patterns from a multi-step classification model to understand how we lie to the police. Knowledge-Based Systems, Volume 149 (2018) Pages 155-168, ISSN 0950-
7051
https://doi.org/10.1016/j.knosys.2018.03.010

11. Wheeler, Andrew; SteenBeck, Wouter (2020). Mapping the risk terrain for crime using machine Learning. https://osf.io/preprints/socarxiv/xc538/
12. Wang X., Gerber M.S., Brown D.E. (2012) Automatic Crime Prediction Using Events Extracted from Twitter Posts. In: Yang S.J., Greenberg A.M., Endsley M. (eds) Social Computing, Behavioral - Cultural Modeling and Prediction. SBP 2012. Lecture Notes in Computer Science, vol 7227. Springer, Berlin, Heidelberg https://link.springer.com/chapter/10.1007/978-3-642-29047-3_28
13. Xinyu Chen; Youngwoon Cho; Suk Young Jang. Crime prediction using Twitter sentiment and weather. 2015 Systems and Information Engineering Desing Symposium. https://ieeexplore.ieee.org/document/7117012
14. Mur Petit , Rosa (2016) Aproximación a la victimización no denunciada (cifras negras) a partir de la última Encuesta de seguridad pública de cataluña (ESPC). Actes del XII Congreso Español de Sociología.
15. Gerell, Manne. Quantifying the Geographical (Un) realiability of Police Data. Nordisk politiforskining. 02/2018 (Volum 5). https://doi.org/10.18261/issn.1894-8693-2018-02-05
16. Giménez-Santana, A., Medina-Sarmiento, J. E., & Miró-Llinares, F. (2018). Risk terrain modeling for road safety: identifying crash-related environmental factors in the province of Cádiz, Spain. *European Journal on Criminal Policy and Research*, *24*, 451–467. https://doi.org/10.1007/s10610-018-9398-x
17. Caplan, J. M., Kennedy, L. W., & Miller, J. (2011). Risk terrain modeling: Brokering criminological theory and GIS methods for crime forecasting. Justice Quarterly, 28, 360-381. https://doi.org/10.1080/07418825.2010.486037
18. Mohler G.O., Porter, M.D. (2018) Rotational grid, PAI-maximizing crime forecasts. Statistical Analysis and Data Mining 11: 227-236
19. Ratcliffe, J.H, Taylor, R.B., Askey, A.P., *et. al.* (2020) The Philadelphia predictive policing experiment. Journal of Experimental Criminology https://doi.org/10.1007/s11292-019-09400-2
20. Ensing, D., Fiedler, S. A., Neville, S. Scheidegger, C., i Venkatasubramanian, S. (2018). Runaway Feedback Loops in Predictive Policing. Proceeding of Machine Learning Research 81:1-12. 2018.
21. Rummens, A, Hardyns, W, & Pauwels, L (2017) The use of predictive analysis in spatiotemporal crime forecasting: Building and testing a model in an urban context. Applied Geography 86: 255-261
22. Duan, L., Hu, T. Chen, E. Zhu, J., Gao, C. (2017). Deep Convolutional Neural Networks for Spatiotemporal Crime Prediction. International Conference Information and Knowledge Engineering. https://csce.ucmss.com/cr/books/2017/LFS/CSREA2017/IKE3704.pdf

23. Wang, B., Zhang, D., Zhang, D., Brantingham, J., Bertozzi, A. L. (2017) Deep Learning for Real time Crime Forecasting. https://arxiv.org/abs/1707.03340
24. Zuang Y., Almeida, M., Morabito, M., Ding, W. (2017) “Crime Hot Spot Forecasting: A Recurrent Model with Spatial and Temporal Information.” 2017 IEEE International Conference on Big Knowledge. https://www.cs.umb.edu/~ding/papers/08023406.pdf
25. Ramirez-Alcocer, U., Tello-Leal, E., Mata-Torre, J. (2019). “Predicting Incidents of Crime through LSTM Neural Networks in Smart City Domain.” SMART 2019 : The Eighth International Conference on Smart Cities, Systems, Devices and Technologies https://www.thinkmind.org/download.php?articleid=smart_2019_2_30_40010
26. O’Neil, Cathy. Armas de destucción Matemática. Cómo el Big Data aumenta la desigualdad y amenaza la democracia. Capitán Swing Libros. Madrid. 2017
27. Ferguson, Andrew Guthrie. The Rise of Big Data Policing. New York University Press. New York. 2017
28. Bosch, A.; Casas, J.i; Lozano, T. (2019) Deep Learning. Principios y fundamentos. Editorial UOC. Barcelona, 2019.
29. Gironés, J., Casas, J., Minguillón, J., Caihuelas, R. (2017). Minería de datos. Modelos y Algoritmos. Editorial UOC. Barcelona, 2017.

8. Annexos

Camps originals a les dades de cadascuna de les tres ciutats:

Nova York[9]:

- 'CMPLNT_NUM': És l'identificador de cada cas, les dades són de tipus int64, per tant, serà útil per identificar cada registre
- 'CMPLNT_FR_DT': La data d'inici què han passat o s'estima que han passat els fets. Cal transformar-ho a format de data i també s'extrauran altres variables relacionades amb la data.
- 'CMPLNT_FR_TM': L'hora d'inici en què han passat o s'estima que han passat els fets. Cal transformar-ho a format dd'hora i es crearà un camp unificat amb la data per tenir un moment concret.
- 'CMPLNT_TO_DT': La data de final en què han passat o s'estima que han passat els fets. En la majoria dels casos hi ha una hora fixa i aquest camp és buit. Quan s'indica podria comportar una manca de fiabilitat de l'hora d'inici. No obstant això, s'obviarà aquest camp.
- 'CMPLNT_TO_TM': L'hora de final en què han passat o s'estima que han passat els fets. En la majoria dels casos hi ha una hora fixa i aquest camp és buit. Quan s'indica podria comportar una manca de fiabilitat de l'hora d'inici. No obstant això, s'obviarà aquest camp.
- 'ADDR_PCT_CD': El codi de la comissaria del lloc on han passat els fets, n'hi ha 78.
- 'RPT_DT': Data en què s'ha recollit la denúncia. No es considera útil i s'elimina.
- 'KY_CD': Codi del delicte comès, 74 valors diferents.
- 'OFNS_DESC': Descripció del delicte que s'ha comès, hauria de correspondre amb el camp KY_CD, però hi ha 72 valors diferents.
- 'PD_CD': Codi per un segon nivell de classificació dels delictes. Arriba a 432 valors diferents.
- 'PD_DESC': Descripció del segon nivell de classificació dels delictes. Arriba a 422 valors diferents.
- 'CRM_ATPT_CPTD_CD': Indica si el valor s'ha consumat o és una temptativa, hi ha alguns valors buits. Es descarta aquest camp.
- 'LAW_CAT_CD': Indica el nivell del delicte en tres graus: *felony*, *misdemeanor* o *violation*, i correspondria al nivell més baix de classificació dels delictes.
- 'BORO_NM': Nom del districte on han passat els fets.
- 'LOC_OF_OCCUR_DESC': Indicacions sobre el lloc dels fets, per si ha passat al voltat, dins, davant de o darrere del lloc.
- 'PREM_TYP_DESC': Indica el tipus de lloc on han passat els fets.
- 'JURIS_DESC': Nom de la jurisdicció competent en el delicte.
- 'JURISDICTION_CODE': Codi de la jurisdicció competent en el delicte.
- 'PARKS_NM': Quan els fets han passat en un parc, uns jardins infantils o lloc semblant, s'indica el nom. Es descarta aquest camp.
- 'HADEVELOPT': En els cassos vinculats a habitatge públic, indica el nom. Es descarta aquest camp.
- 'HOUSING_PSA': Codi de l'Àrea de servei policial
- 'X_COORD_CD': Coordenada X del lloc on ha passat el fet, en la projeccio NAD 83.
- 'Y_COORD_CD': Coordenada X del lloc on ha passat el fet, en la projeccio NAD 83.
- 'SUSP_AGE_GROUP': Grup d'edat del sospitòs. Es descarta aquest camp.

9 https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i

- 'SUSP_RACE': Raça del sospitòs. Es descarta aquest camp.
- 'SUSP_SEX': Sexe del sospitòs. Es descarta aquest camp.
- 'TRANSIT_DISTRICT': Districte de trànsit.
- 'Latitude': Latitud del lloc dels fets.
- 'Longitude': Longitud del lloc dels fets.
- 'Lat_Lon': Tupla amb la latitud i la longitud del lloc dels fets.
- 'PATROL_BORO': Districte de patrullatge.
- 'STATION_NAME': Nom de lestació
- 'VIC_AGE_GROUP': Grup d'edat de la víctima. Es descarta aquest camp.
- 'VIC_RACE': Raça de la víctima. Es descarta aquest camp.
- 'VIC_SEX': Sexe de la víctima. Es descarta aquest camp.

Chicago[10]:

- 'ID': És l'identificador de cada cas, les dades són de tipus int64, per tant, serà útil per identificar cada registre
- 'Case Number': És un identificació dels procediments. La xifra de registres únics és molt propera a la de l'identificador, per tant, probablement hi hagi algun valor repetit o nul. Les dades són alfanumèriques. La informació que aporta no ens resultarà d'utilitat, per tant eliminarem la columna.
- 'Date': La data en què han passat o s'estima que han passat els fets. Cal transformar-ho a format de data i també s'extrauran altres variables relacionades amb la data.
- 'Block': Informació sobre l'illa de cases on ha passat el fet. Es podria utilitzar com l'agrupació de punts més petita.
- 'IUCR': Informació sobre el tipus de fet. Hi ha una gran quantitat de tipologies (402), i cal recòrrer a una informació externa per interpretar els codis.
- 'Primary Type': Informació sobre els principals tipus de fet, són 36 categories, en format text.
- 'Description': Informació sobre els tipus de fet secundaris, són 517 categories, en format text.
- 'Location Description': Informació sobre el lloc on han passat els fets.
- 'Arrest': Indica si s'ha produït una detenció. És una variable booleana.
- 'Domestic': Informa si es tracta d'un fet de violència domèstica. És una variable booleana.
- 'Beat': Informació geogràfica en relació a les zones de patrullage, que en són 304.
- 'District': Informació geogràfica en relació als districtes policials. N'hi ha 25 registres únics, malgrat que a la informació municipal en parla de 22.
- 'Ward': Informació geogràfica que fa referència a districtes municipals, diferents dels districtes policials.(51)
- 'Community Area': Una altra divisió territorial,amb àrees una mica més petites, ja que són 79 valors únics.
- 'FBI Code': Una altra classificació dels tipus de fets, en aquest cas de l'àmbit nacional dels EUA, i que té 26 tipus de fets.
- 'X Coordinate': Coordenada X del lloc on ha passat el fet, en la projeccio NAD 1983. La ubicació pot estàr parcialment moguda, però hauria de correspondre a la mateixa illa de cases.
- 'Y Coordinate': Coordenada Y del lloc on ha passat el fet, en la projeccio NAD 1983. La ubicació pot estàr parcialment moguda, però hauria de correspondre a la mateixa illa de cases.
- 'Updated On': Data en què s'ha actualitzat el registre. No es considera útil i s'elimina.

[10] https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2

- 'Latitude' : Latitud del lloc dels fets. Igual que les coordenades, la ubicació pot estàr parcialment moguda, però hauria de correspondre a la mateixa illa de cases.
- 'Longitude' : Longitud del lloc dels fets. Igual que les coordenades, la ubicació pot estàr parcialment moguda, però hauria de correspondre a la mateixa illa de cases.
- 'Location' : Tupla amb la latitud i la longitud. Atès que la informació està en altres camps, s'omet també aquesta columna.

Los Angeles[11]:

- 'DR_NO': És l'identificador de cada cas, les dades són de tipus int64, per tant, serà útil per identificar cada registre
- 'Date Rptd': Data en què s'ha recollit la denúncia. No es considera útil i s'elimina.
- 'DATE OCC': La data en què han passat o s'estima que han passat els fets. Cal transformar-ho a format de data i també s'extrauran altres variables relacionades amb la data.
- 'TIME OCC': L'hora en què han passat o s'estima que han passat els fets. Cal transformar-ho a format dd'hora i es crearà un camp unificat amb la data per tenir un moment concret.
- 'AREA ': Codi numèric de l'àrea policial on han passat els fets. N'hi ha 21.
- 'AREA NAME': Nom de l'àrea policial on han passat els fets. N'hi ha 21 i corresponen amb els codis d'AREA.
- 'Rpt Dist No': Codi numèric de la subàrea geogràfica. N'hi ha 1303 registres únics.
- 'Part 1-2': Camp numèric amb dos valors únics, 1 i 2, encara que no s'ha pogut determinar a què corresponen.
- 'Crm Cd': Codi del delicte comès, 142 valors diferents.
- 'Crm Cd Desc': Descripció del delicte comès, 142 valors diferents.
- 'Mocodes': Codis del modus operandi. Cada camp pot contenir més d'un codi. Inicialment no es tractarà amb aquesta informació, que caldria separar per poder analitzar-la de manera adequada.
- 'Vict Age': Camp numèric amb l'edat de la víctima. Es descarta el seu ús.
- 'Vict Sex': Camp amb el codi del sexe de la víctima (Dona, Home o desconegut). Es descarta el seu ús.
- 'Vict Descent': Camp amb una codificació de lletra per identificar l'origen de la víctima. Es descarta el seu ús.
- 'Premis Cd': Codi sobre el tipus de lloc on han passat els fets.
- 'Premis Desc': Informació sobre el tipus lloc on han passat els fets.
- 'Weapon Used Cd': Codi sobre el tipus d'arma utilitzat en el delicte.
- 'Weapon Desc': Descripcío del tipus d'arma utilitzada.
- 'Status': Codi sobre l'estat del cas, indica, per exemple, si s'ha detingut una persona o si contniua la investigació (situació per defecte). Es descarta el seu ús
- 'Status Desc': Descripció de l'estat del cas, indica, per exemple, si s'ha detingut una persona o si contniua la investigació (situació per defecte). Es descarta el seu ús
- 'Crm Cd 1': Codi del fet principal que ha tingut lloc.
- 'Crm Cd 2': Codi d'un eventual fet secundari que ha tingut lloc en el mateix fet. Caldria afegir noves línies per aquests fets.
- 'Crm Cd 3': Codi d'un eventual fet terciari que ha tingut lloc en el mateix fet. Caldria afegir noves línies per aquests fets.
- 'Crm Cd 4': Codi d'un eventual quart fet que ha tingut lloc en el mateix fet. Caldria afegir noves línies per aquests fets.
- 'LOCATION': Adreça del lloc on han passat els fets. Es descarta aquest camp.
- 'Cross Street': Creuament dels carrers. Es descarta aquest camp.
- 'LAT' : Latitud del lloc dels fets. Igual que les coordenades, la ubicació pot estàr parcialment moguda, però hauria de correspondre a la mateixa illa de cases.

11 https://data.lacity.org/A-Safe-City/Crime-Data-from-2010-to-2019/63jg-8b9z

- 'LON' : Longitud del lloc dels fets. Igual que les coordenades, la ubicació pot estàr parcialment moguda, però hauria de correspondre a la mateixa illa de cases.