Grau en Matemàtica Computacional i Analítica de Dades

Cas Kaggle

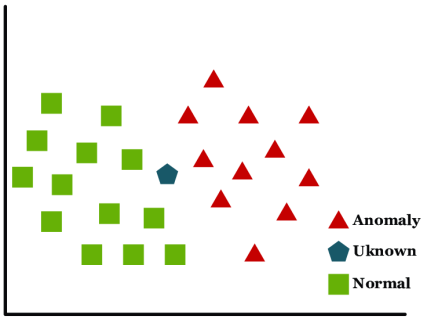
Aprenentatge Computacional

Xavier Seminario Monllaó (1603853)

Índex

1 Presentació del problema i del dataset

En aquest treball es demana fer un anàlisi d'un cas particular d'un dataset extret de Kaggle. S'han realitzat diversos gràfics per tal d'entendre millor les dades i classificacions per veure si realment les dades proporcionades tenen algun tipus de correlació o dependència entre elles.



Un cop vistes les dades de les que es disposa, per tal de tenir un aprenentatge més eficient, s'estudiarà si cal o no normalitzar les dades i es tractaran els valors nulls de les columnes (si és que hi ha). Posteriorment, s'escollirà sobre quina o quines tècniques d'aprenentatge computacional supervisat *(sklearning)* es vol treballar i s'aplicaran al dataset. També es farà una cross-validació dels diferents models *(cross-validation)* per tal de conèixer quin serà el resultat esperat d'aquests amb dades mai vistes abans.

Tot això s'acompanyarà amb les explicacions i justificacions pertinents i diversos elements visuals que facilitaran la comprensió del text.

2 Descripció de la Base de dades

En aquest primer apartat l'objectiu serà analitzar els diferents atributs que composen la base de dades amb la que s'està treballant i fixar quin és l'atribut objectiu a predir.
La base de dades que es tractarà està extreta de la web kaggle.com. Aquesta base de dades conté la informació del 7941282 crims comesos a la ciutat de Chicago en l'interval de temps del 2001 fins al 2017. Concretament, en aquest treball només s'ha treballat en l'interval de 2012 a 2017, la raó d'aquesta reducció ha sigut l'excessiu temps d'execució per a qualsevol comanda.

2.1 Descripció dels atributs

- **Col.1: ID** Identificador únic del crim. (int)
- **Col.2: Case Number** Número únic per crim que és assignat pel departament de policia de Chicago. (categòric)
- **Col.3: Date** Data del incident (categòric)
- **Col.4: Block** Adreça parcial on va succeïr l'incident. (categòric)
- **Col.5 IUCR** Codi de report de l'incident. Altament relacionat amb el tipus de crim comès. (categòrica)
- **Col.6: Primary Type** Descripció del crim general, extreta de IUCR. Aquest serà l'atribut objectiu d'aquest treball.(categòrica)
- **Col.7: Description** Descripció més detallada del crim. (categòrica)
- **Col.8: Location description** Descripció del lloc de l'incident. (categòrica
- **Col.9: Arrest** Indica si es va realitzar un arrestament. (booleana)
- **Col.10: Domestic** Indica si va ser un incident de violència domèstica. (booleana)
- **Col.11: Beat** Indica el "Beat"del crim. Un beat es la regió de policia més petita. (int)
- **Col.12: District** Districte de l'incident. (float)
- **Col.13: Ward** Ward on va succeïr. (float)
- **Col.14: Community Area** Àrea comunitaria on va succeïr, Chicago en té 77. (float)
- **Col.15: FBI Code** Classificació del crim segons l'FBI. (categòrica)
- **Col.16,17 X,Y Coordiante** Coordenades de l'incident. (float)
- **Col.18: Year** Any en el que es va cometre el crim. (int)
- **Col.19: Updated On** Dia i hora en el que es va reportar. (categòrica)
- **Col.20,21: Longitud, Latitud** Longitut i latitut d'on es va cometre el crim. (float)
- **Col.22: Location** Longitut i latitut en conjunt. (categòrica).

2.2 Primeres observacions del dataset i columnes categòriques

Abans de fer cap anàlisi sobre les dades és molt important netejar-les. Convenientment, la majoria de dades categòrques d'aquest dataset o tenen relació directa amb el nostre atribut objectiu o són valors únics que no aporten informació útil. Per tant podem rebutjar directament per a la classificació els atributs: Unnamed: 0, Location(es literalment el mateix que Latitude i Longitude), IUCR, Description, Block, FBI Code i Case Number. I només queden els atribut categòrics: Updated On, Date i Location description.

2.2.1 Updated on / Date

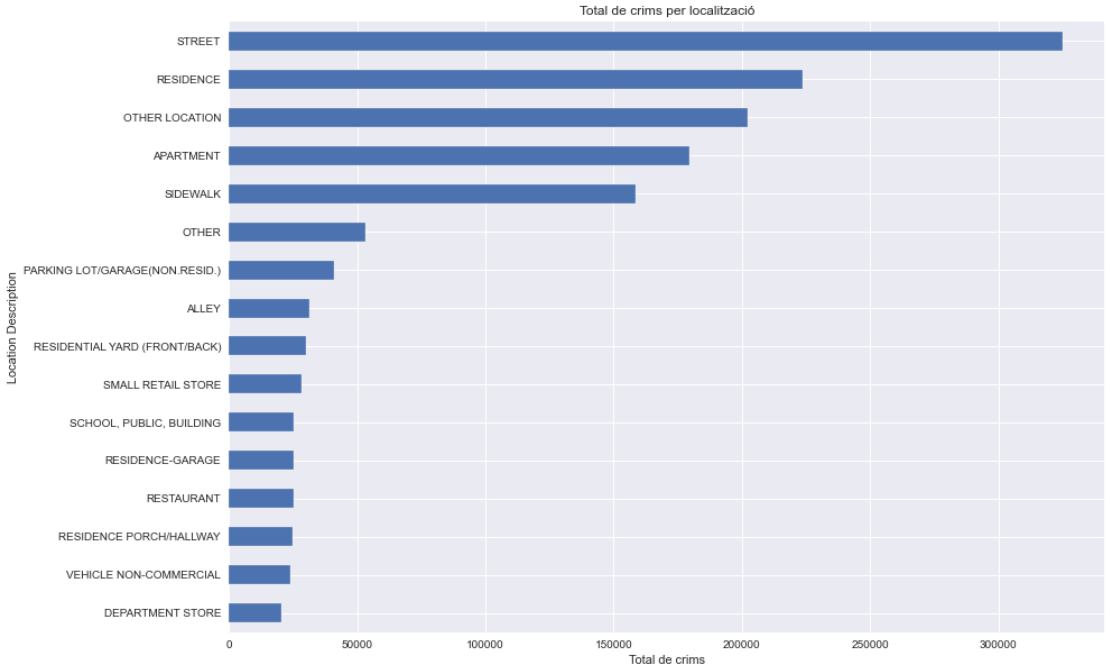
Aquestes dos columnes conten contiguts similars, per al cas del problema a estudiar, és més interessant la columna Date que Updated On, per tant també es pot eliminar aquesta columna. Pel que fa la columna Date, per tal de que deixés de ser un atribut categòric, s'ha dividit en 6 columnes diferents: Year, Month, Day, Hour, Minute i Second, on el contingut de cada atribut és el que indica el seu propi nom, per tant les 6 són ints.

2.2.2 Location Description

En un principi també es va rebutjar aquest atribut ja que tenia moltes categories diferents i fer encoding seria massa problemàtic. Més endavant, es va plantejar la idea de que segurament aquest atribut podria ajudar bastant a la regressió, ja que certs crims eren més comuns depenent del lloc, com per exemple el robatori d'un cotxe era més probable a un garatge que a un cinema. La solució al gran número de possibles valors va ser només agafar les 15 primeres localitzacions i ajuntar les restant a una comú "OTHER LOCATION". La llista resultant va ser la següent:

ALLEY/APARTMENT/DEPARTMENT STORE/OTHER/OTHER LOCATION/PARKING LOT/GARAGE(NON.RESID.)/RESIDENCE/RESIDENCE/(PORCH/HALLWAY)/RESIDENCE-GARAGE/RESIDENTIAL YARD (FRONT/BACK)/RESTAURANT/ SCHOOL, PUBLIC, BUILDING/SIDEWALK/SMALL RETAIL STORE/STREET/VEHICLE NON-COMMERCIAL

Amb aquesta llista, un one-hot encoding era més raonable i és el que es va fer, quedant-nos amb la distribució següent:



2.3 Tractament de Nulls

No s'ha dedicat temps al tractament de nulls, ja que només hi havia 38349 que en comparació a les 1456714 dades de les que disposem no representa un gran canvi, per tant s'han eliminat directament aquestes dades.

3 Mapping de variables

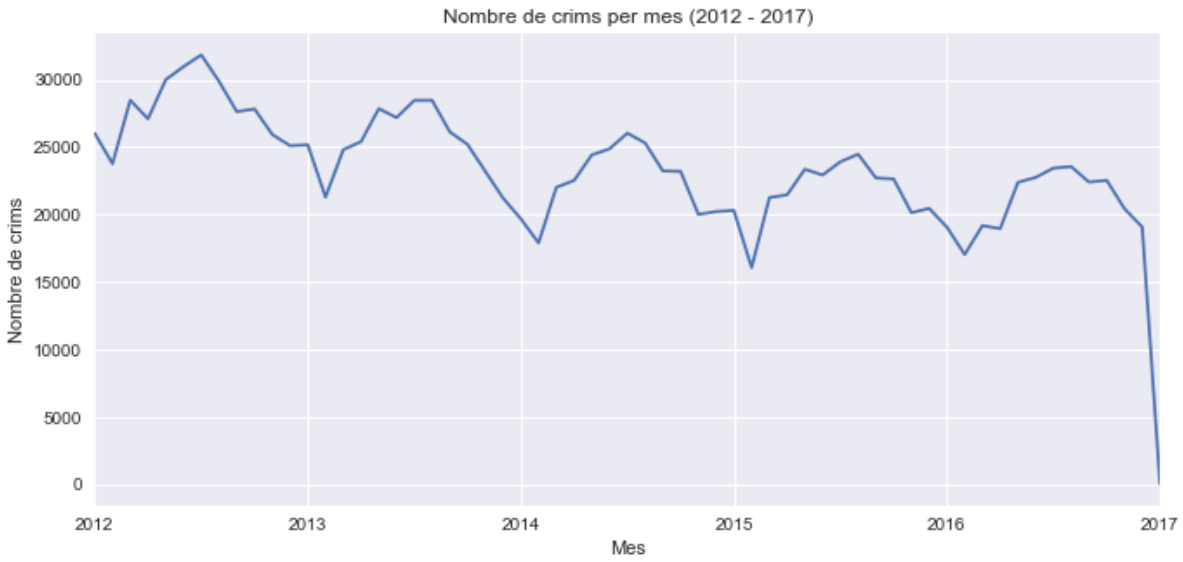
El principal objectiu d'aquest apartat és entendre com ha avançat el crim a Chicago amb els anys y de quina forma está distribuïda la ciutat en funció dels crims.

3.1 Crims en funció del temps

En aquest apartat es veuràn diferents gràfics per veure com ha anat evolucionant el crim a Chicago amb els anys, i també si hi ha alguna relació amb els mesos o dies de la setmana.

3.1.1 Crims per mes

Aquest és un gràfic del nombre total de crims en funció dels mesos al llarg dels anys.



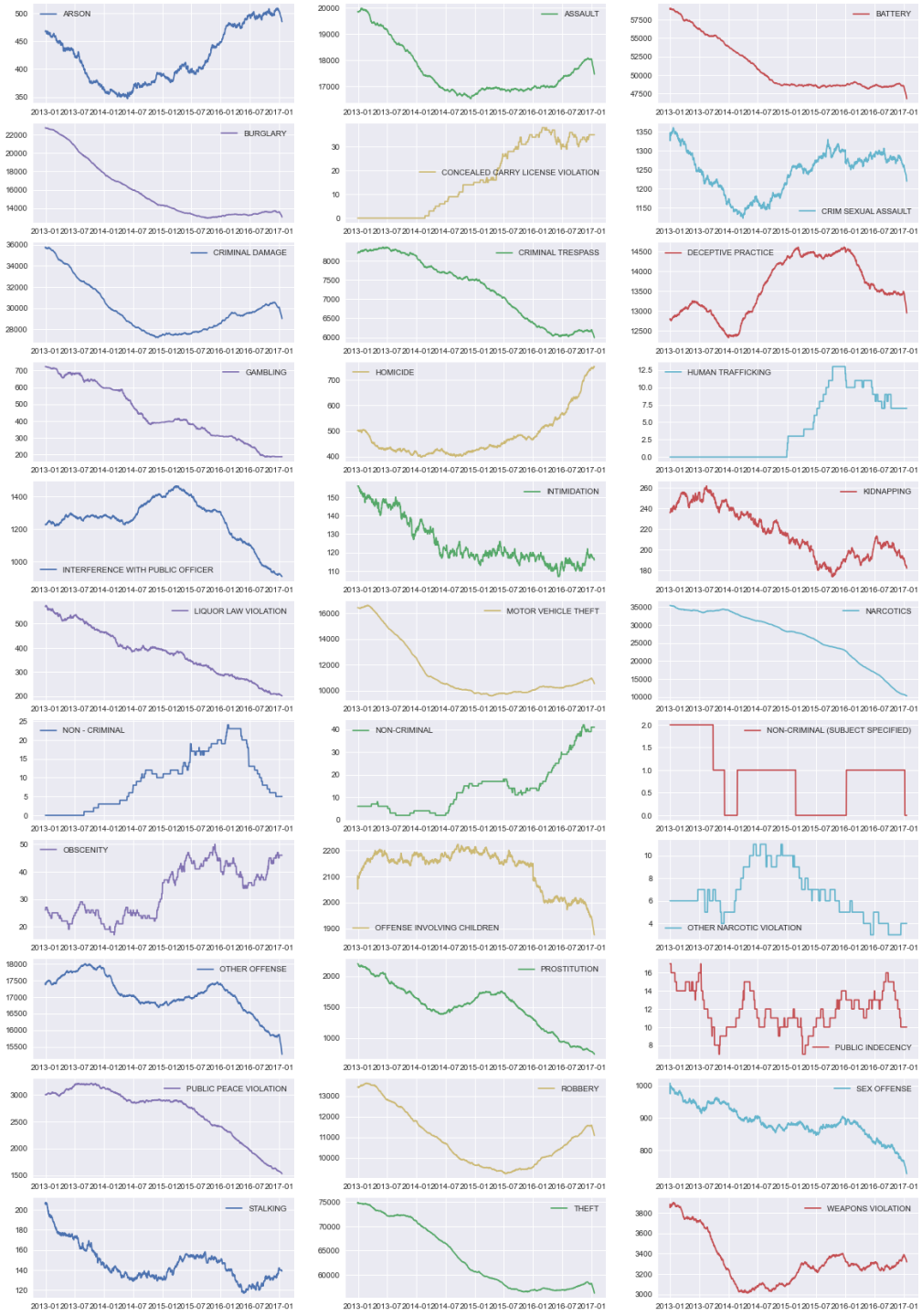
Es pot observar com amb els anys, el nombre de crims ha anat disminuïnt, i que en els mesos intermitjos de l'any, s'acostuma a realitzar més crims.

3.1.2 Crims per dia

Aquest és un gàfic del nombre total de crims dia per dia del 2012 fins al 2017.

Com s'ha vist prèviament, amb el temps el nombre total de crims al dia ha anat disminuïnt.

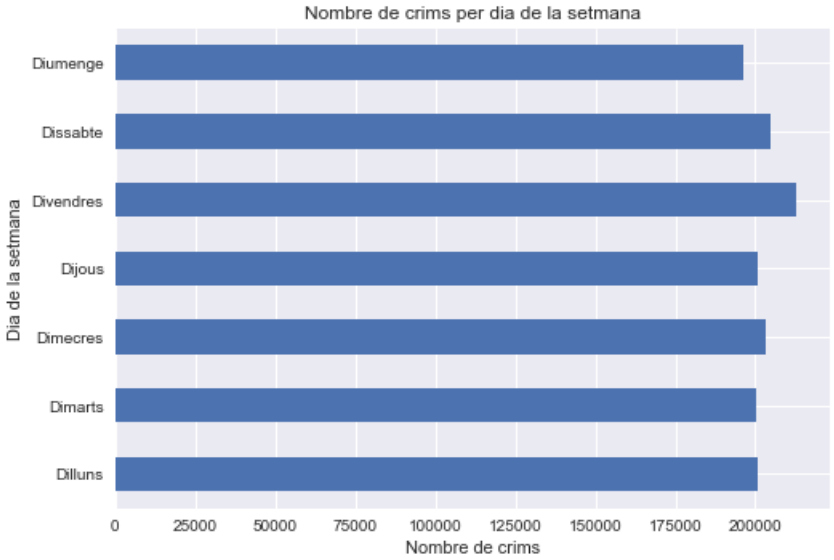
Aquest és el mateix gràfic d'abans però individualitzat per tipus de crim.

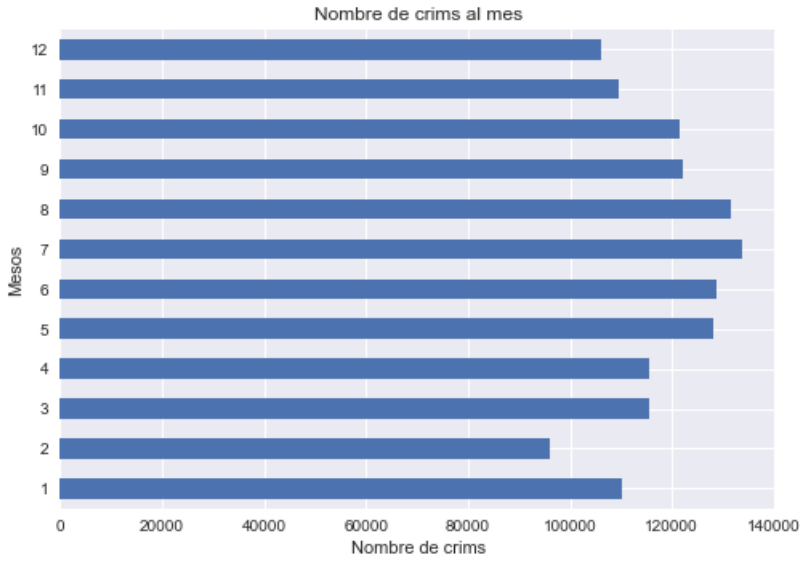


Es pot observar com no tots els crims han disminuït al llarg del temps, alguns com homicidis o robatoris acaben el gràfic en creixement. Aquests gràfics en poden indicar que el temps pot ser un factor a tenir en compte en les classificacions.

3.1.3 Crims a la setmana i al mes

A continuació es mostraràn dos gràfics del nombre de crims per setmana i per mes en total.

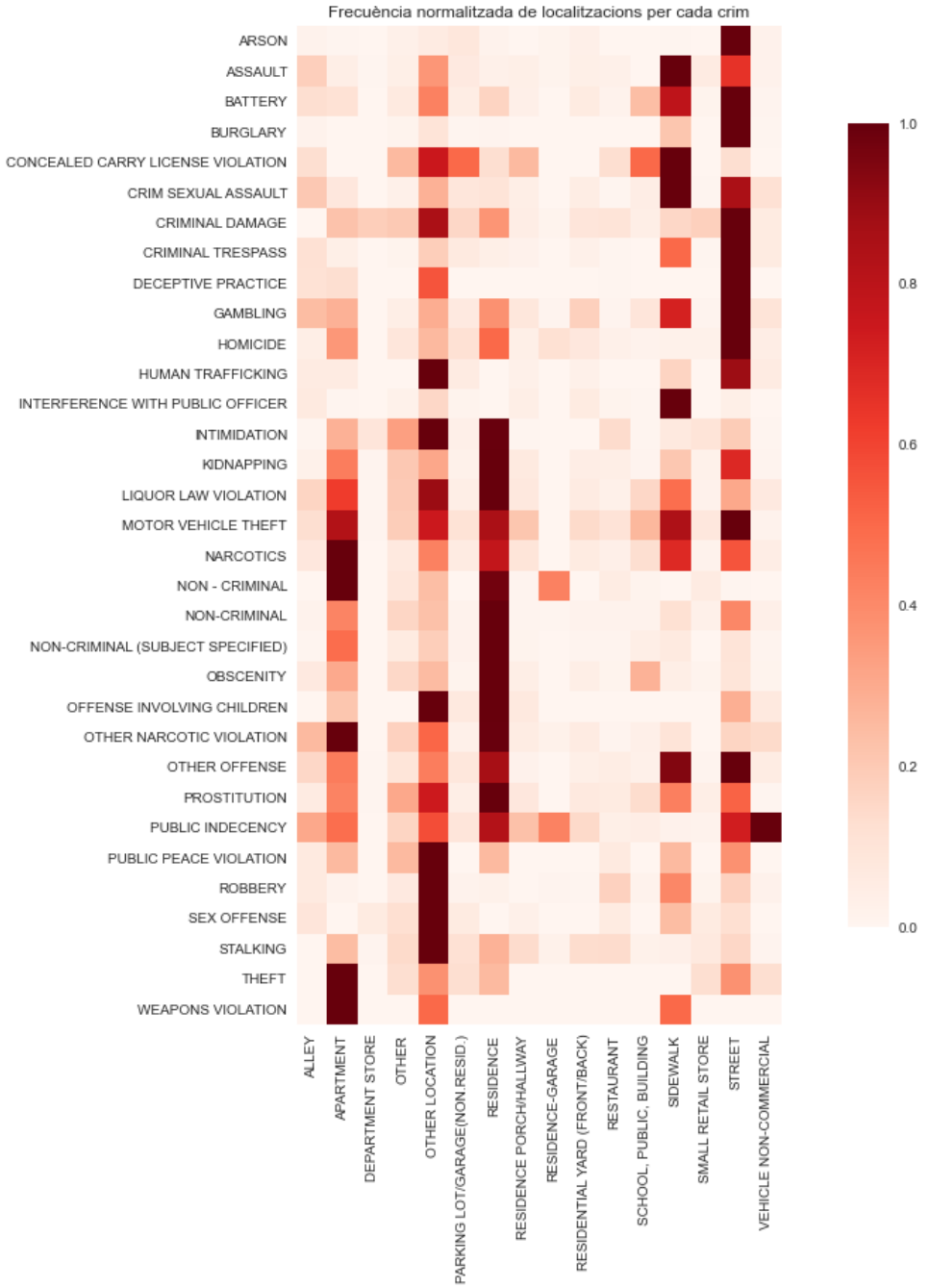




Es pot veure que el divendres és el dia en el que es realitzen més crims en promig, encara que la diferència no és molt notòria. En quant als mesos podem observar el que ja haviem vist abans, que en els mesos intermitjos de l'any hi ha més tendència a que es realitzin més crims.

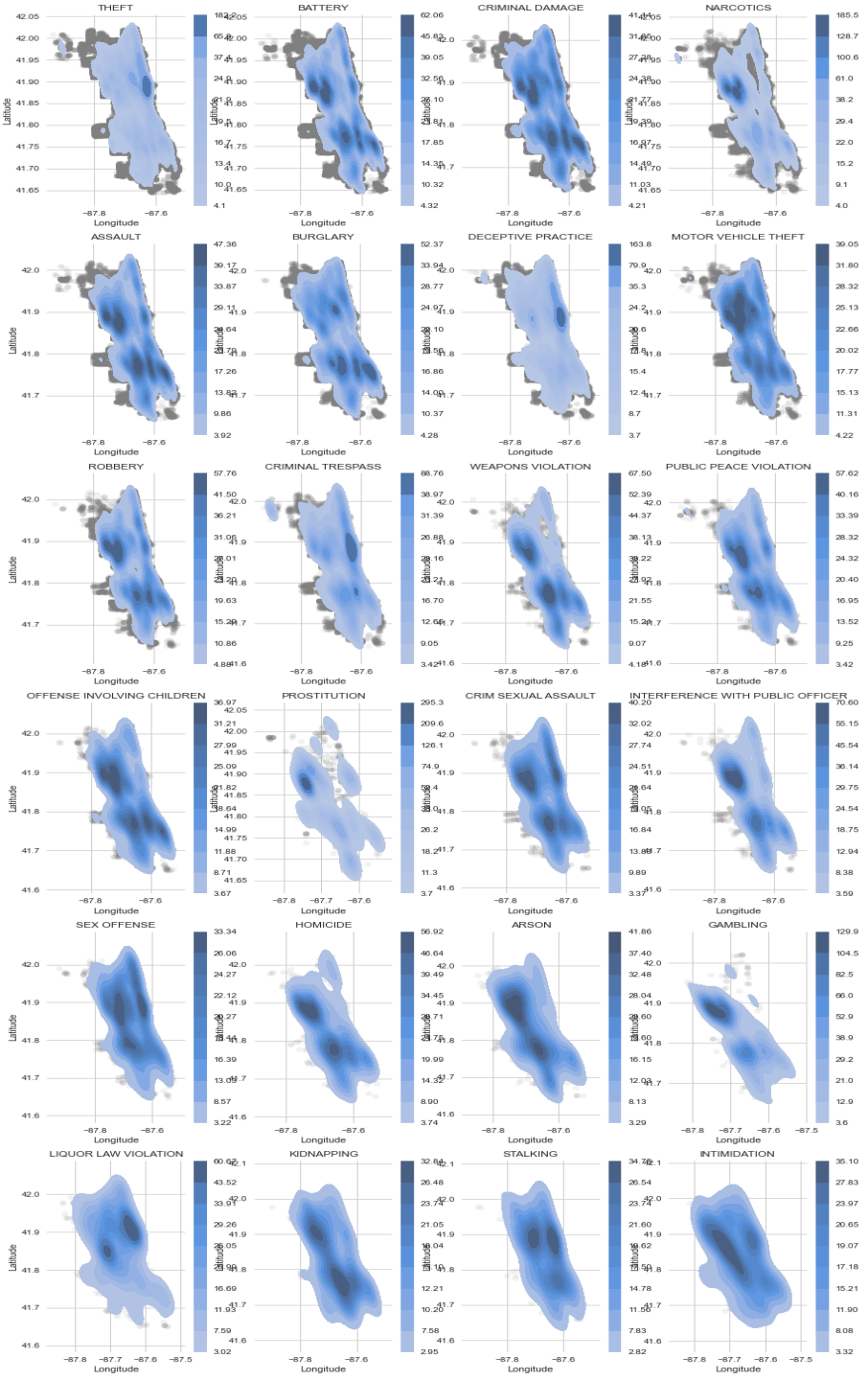
3.2 Crims per localització

En aquest apartat es mostrarán 2 taules o mapes molt importants en aquest treball que han servit per millorar la precisió de les prediccions. El primer gràfic és una taula de frecuències del crims segons la descripció de la seva localització.



En aquesta taula podem observar com hi ha clarament 5 localitzacions que són més rellevants que les altres, concretament APARTMENT, OTHER LOCATION, RESIDENCE, SIDEWALK i STREET.

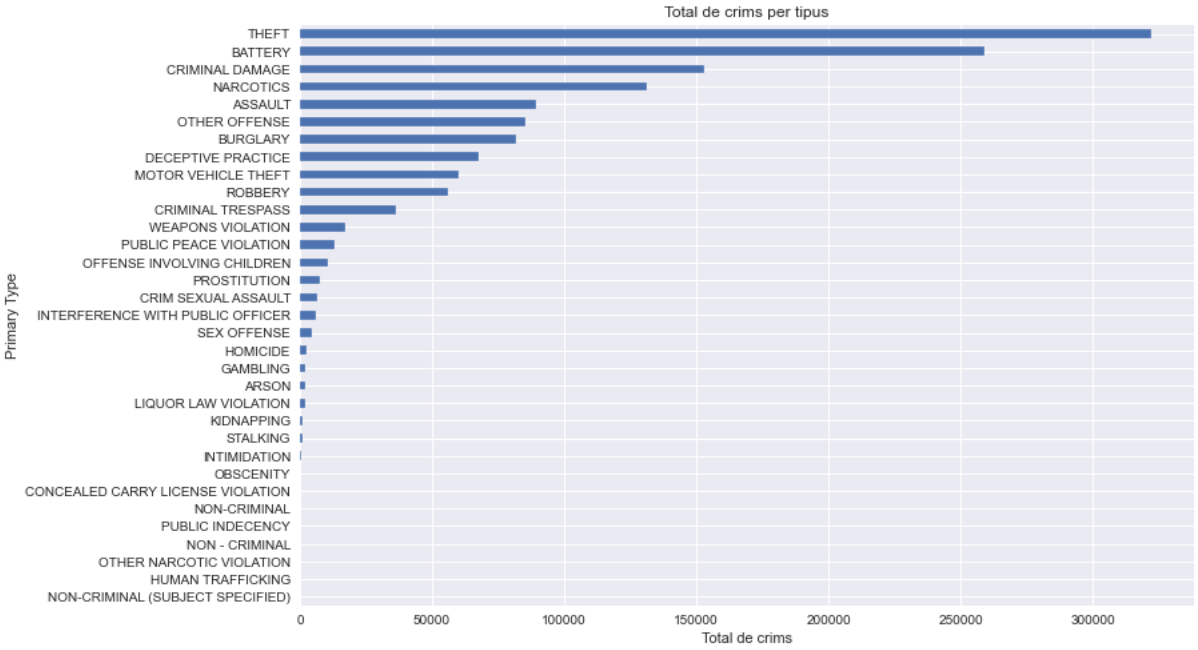
Els següents gràfics són mapes de calor dels diferents crims a la ciutat de Chicago.



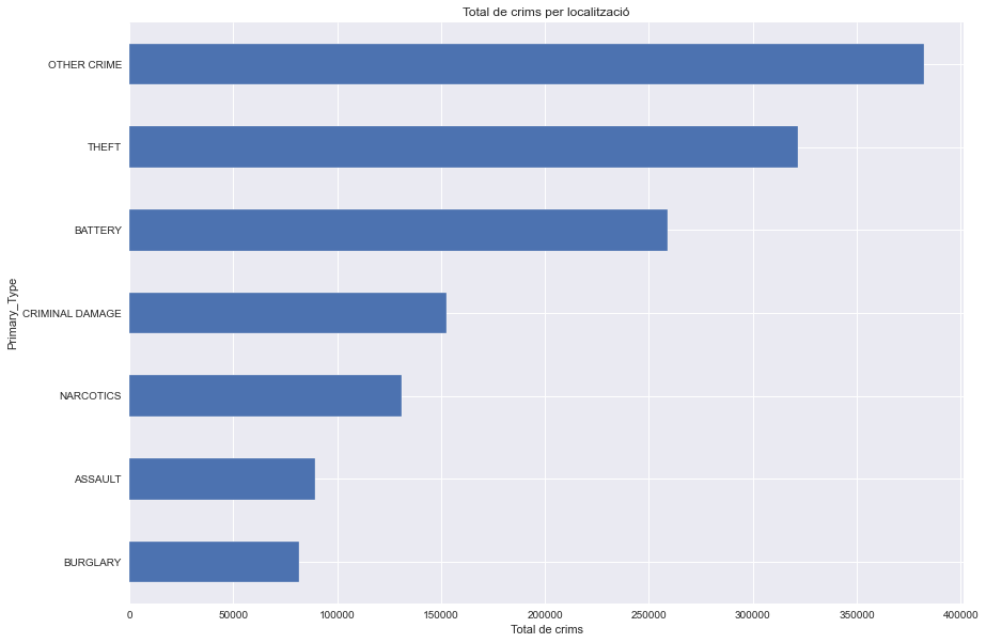
Es pot veure aqui també, com hi ha certa tendència a que alguns crims passin en segons quines zones, per exemple en aquests mapes es poden arribar a diferènciar 4 distribucions de mapes diferents.

4 Tractament d'atribut objectiu

En un principi, l'atribut objectiu del dataset tenia moltes categories diferents, així que en aquest treball s'ha optat en reduïr el nombre de categories agrupant els crims seguint diferents criteris. Aquest és l'atribut objectiu sense cap tipus de modificació.

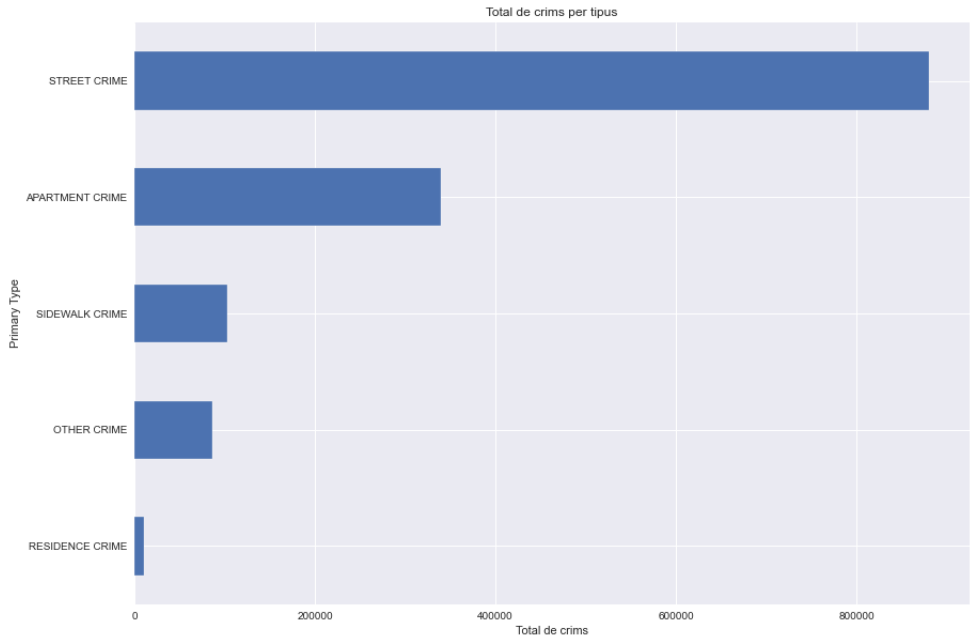


La primera modificació que es va plantejar va ser reduïr el dataset a les 6 primeres classes de la taula anterior. Les demés categories s'acumularien a la classe OTHER CRIME. Aquesta distribució és la que deixa l'atribut objectiu millor repartit, però és la que dona pitjors resultats.



La segona modificació va ser agrupar els tipus de crims segons el primer gràfic de l'apartat

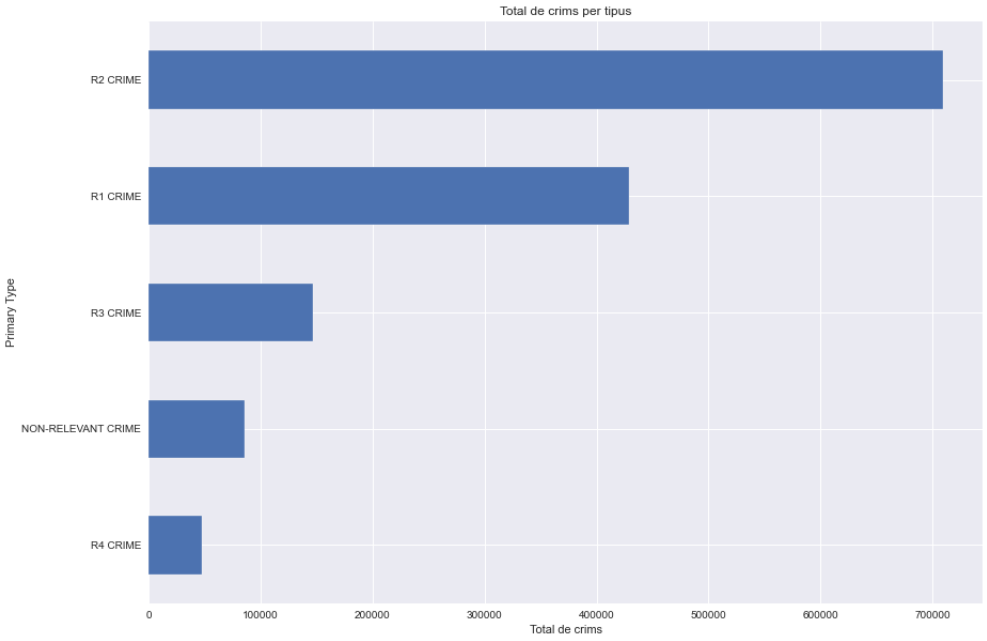
3.2, com s'ha dit en el apartat, es poden veure 5 classes principals, per tant aquestes han sigut les classes resultants.



Auesta distribució está molt més descompensada que la primera plantejada, encara que hi ha bastants tipus de crims comesos a RESIDENCE, el total no és gaire alt, sembla que la majoria de crims es cometen als carrers, junts a la classe STREET CRIME.

La tercera i ultima distribució plantejada, es ajuntar les classes segons les coordenades més comunes en les que apareixen els diferents tipus de crim. El criteri seguit ha sigut mirant la taula de mapes de calor de l'apartat anterior.

La clasee NON-RELEVANT CRIME és una classe que agrupa els crims que no tenien suficients dades com per fer un mapa de calor decent, irònicament té més dades que els crims de la "Zona 4".



5 Primeres regressions

S'ha treballat principalment amb un classificador Random Forest, ja que classificadors com el knn o una SVM trigaven molt a convergir, per exemple un knn amb les dades normalitzades trigava més de 20 minuts, llavors fer proves amb aquest tipus de classificador no era rentable, la SVM no arribava a convergir del tot, més de 2 hores fent el fit i encara no acabava, per tant es van rebutjar fer proves amb aquests classificadors, ja que a més no aportaven millors resultats que el Random Forest, que triga uns pocs minuts en acabar de fer l'entrenament.

5.1 Random Forest

Aquest ha sigut el classificador principal d'aquest treball. S'han fet proves amb les 3 agrupacions de crims plantejades al apartat anterior.

5.1.1 Agrupació 7 classes més comunes

Aquesta és la agrupació amb la precisió més baixa, tenint en compte que també és la que més classes té es pot deduïr també que és la que més difícil té les prediccions. Amb aquesta agrupació s'ha provat fer la classificació mantenint les columnes de LOCATION DESCRIPTION i sense matenirles. Els resultats són els següents.

S'ha mantingut la profunditat màxima de arbre de 15, ja que es va estimar que és el límit abans de l'overfitting, encara que inclús fent overfitting la precisió en el conjunt test no variava massa. A partir de 40 arbres la precisió no augmenta tampoc.

En el cas en el que es mantenien les columnes de localització obtenim una precisió del 48% i acaben resultant les següents matrius de confusió:

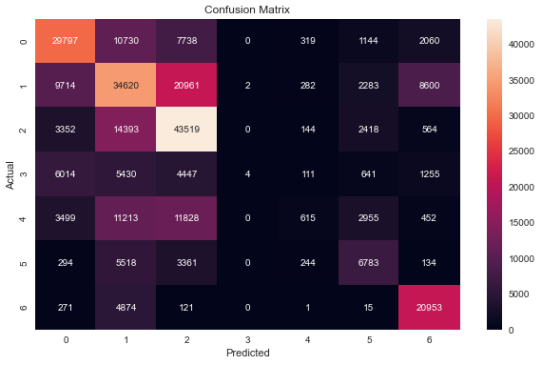


Figura 1: Matriu de confusió del test

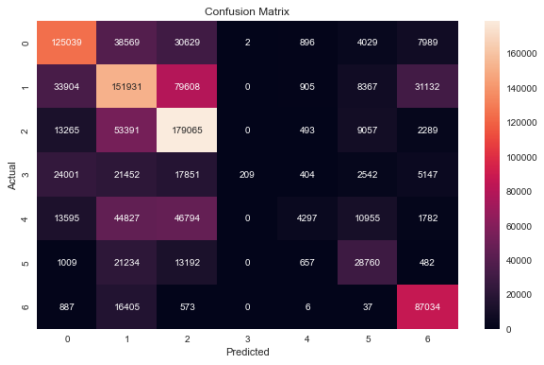


Figura 2: Matriu de confusió del train

Es pot apreciar com les matrius de confusió són bastant similars, s'ha de destacar les classes 3 i 4, que el model no les té pràcticament en compte a l'hora de predir.

Les matrius de confusió resultants d'aquest mateix model però sense la descripció de localització són similars, encara que la precisió en aquest cas es del 44%.

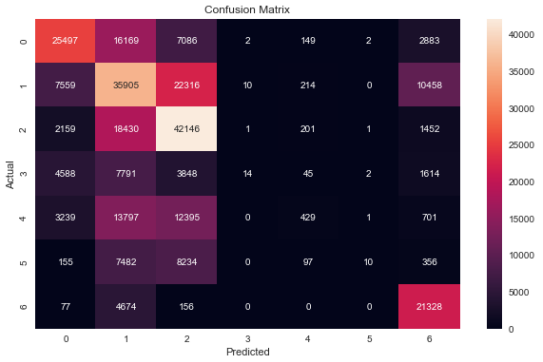


Figura 3: Matriu de confusió del test

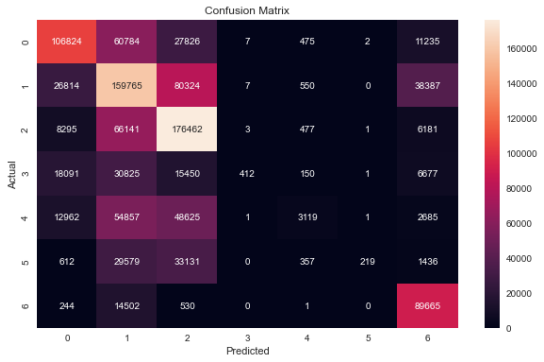


Figura 4: Matriu de confusió del train

És evident que les prediccions són similar, els temps d'execució també ho són, aquest 4% de diferencia en la precisió s'avaluarà més endavant en l'apartat de FEATURE SELECTION.

5.1.2 Agrupació per location description

En aquest cas, com tenir les columnes de localització no feia que l'entrenament trigués molt més, s'ha optat per fer només entrenaments amb totes les columnes a partir d'ara fins a la feature selection o la PCA.

En quant a dades, aquest és el millor model, ja que s'obté la precisió més alta (67%), pero hi ha un inconvenient que s'entén millor amb les matrius de confusió:

Es pot veure clarament que el predictor només prediu la classe 2 i la classe 0, les altres les deixa de costat, encara que quan prediu que un crim serà d'aquestes acostuma a encertar. Per tant es pot concluïr que realment la precisió no és més alta perqué estigui predint bé totes les classes, sinó que ho és perquè està predint que tots els crims són les dos classes més probables.

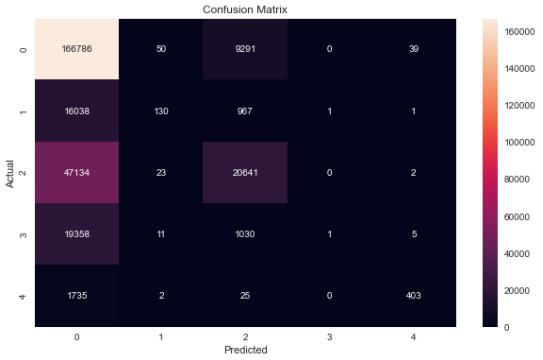


Figura 5: Matriu de confusió del test

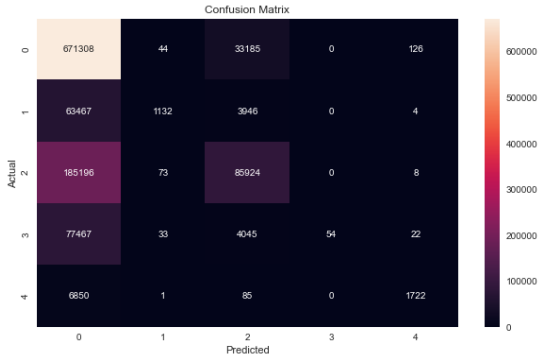


Figura 6: Matriu de confusió del train

5.1.3 Agrupació per regions al heatmap

Aquesta agrupació ha sigut la més oportuna per fer classificacions del tipus de crim, ja que obtenim una precisió similar a la del apartat anterior (66%) amb l'avantatge de que les prediccions no estàn tan esbiaxades.

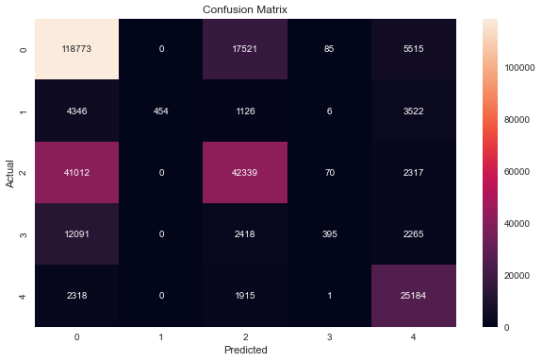


Figura 7: Matriu de confusió del test

Figura 8: Matriu de confusió del train

Es pot veure com ara hi ha 3 classes rellevants en la predicció la 0,2 i 4, i hi ha dues que queden més rebutjades, la 1 i 3. De la mateixa manera que anteriorment, quan el model prediu que el crim serà classe 1 o 3 acostuma a tenir raó. Sembla ser que aquesta distribució de les classes és la més adient, i per tant és la que es seguirà utilitzant en la resta del treball juntament a la agrupació de les 7 classes més comuns.

5.2 Catboost classifier

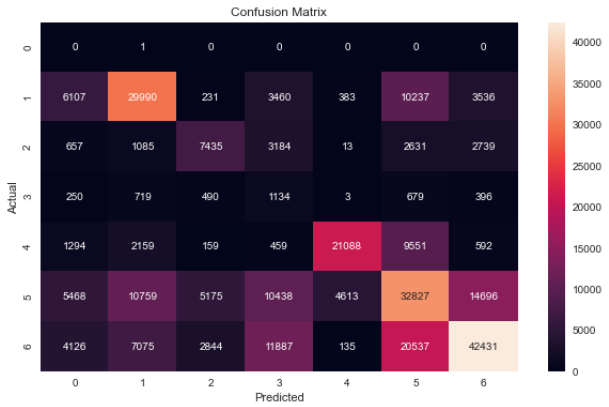
Un altre model que trigava relativament poc en fer l'entrenament de les dades era el Catboost classifier. Aquest classificador pot treballar amb classes categòriques, però com el dataset ja estava adaptat i la única dada categòrica ja havia estat transformada amb un OneHotEncoding. Aquest clasificador és bastant ràpid, fa un entrenament sencer en un minut i mig aproximadament i aconsegueix una accuracy molt similar al Random Forest.

Els hiperparàmetres utilitzats han sigut els següents:

- iterations=65 (a partir de 60 el model no pregressa massa)
- learning_rate=0.1 (valor estàndar)
- depth=10 (maxim de 16, a partir de 10 trigava bastant més)
- loss_function="MultiClass"(La nostra y es multiclasse)

5.2.1 Catboost amb les 7 classes més comunes

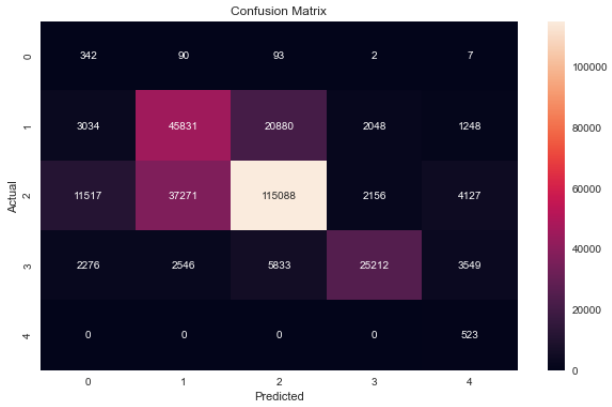
S'aconsegueix una accuracy del 48%, d'igual manera que el Random Forest amb aquesta agrupació.



Podem veue com aquí la diagonal està més marcada que en el Random Forest. Una cosa a destacar és la classe 0, que per alguna raó el model mai decideix escollir-la, sembla bastant curiós tenint en compte la gran quantitat de dades disposades.

5.2.2 Catboost amb agrupació per Heatmap

S'aconsegueix una precisió del 66%, de nou, igual que el Random Forest.



Aquesta matriu de confuusió s'assembla més a la que es generava amb el Random Forest. Cal destacar que aquest classificador triga una mica menys que el Random Forest.

6 Feature Selection

S'ha realitzat una Feature Selection utilitzant el mètode RandomForest amb el dataset de la agrupació per mapes de calor. La llista de atribut seleccionats és la següent: 'ID', 'Arrest', 'Domestic', 'Beat', 'X Coordinate', 'Y Coordinate', 'Latitude', 'Longitude', 'Month', 'Day', 'Hour', 'Minute'. Es pot observar com les columnes del OneHotEncoding no han sigut seleccionades, per tant concluïm que realment no són rellevants a l'hora de classificar.

Si es realitza una classificació random forest amb aquests atributs obtenim una precisió molt similar a la obtinguda en apartats anteriors (65%). Obtenim una mica d'overfitting, però la precisió del test no baixa.

Figura 9: Matriu de confusió del test

Figura 10: Matriu de confusió del train

Es veu que la matriu de confusió del test és similar a les anteriors però ara hi ha més error en les classes 1 i 3, això indica que l'error en les classes 0, 2 i 4 ha disminuït, ja que l'error general es manté respecte les anteriors classificacions.

7 PCA

Una PCA en aquest dataset pot ser perillosa perquè perdre el sentit de algunes columnes pot ser poc beneficiós. De totes maneres s'ha realitzat una PCA la qual se li ha demanat que representi un 90% de la desviació de les dades originals. El dataset resultant és de 24 columnes, reduïnt en 8 les 32 que teniem anteriorment. Pel que fa les regressions, la precisió torna a rondar el 66%, per tant no s'ha considerat seguir endavant amb la PCA, ja que obteniem una precisió similar amb el pes de perdre el sentit interpretatiu de les dades.

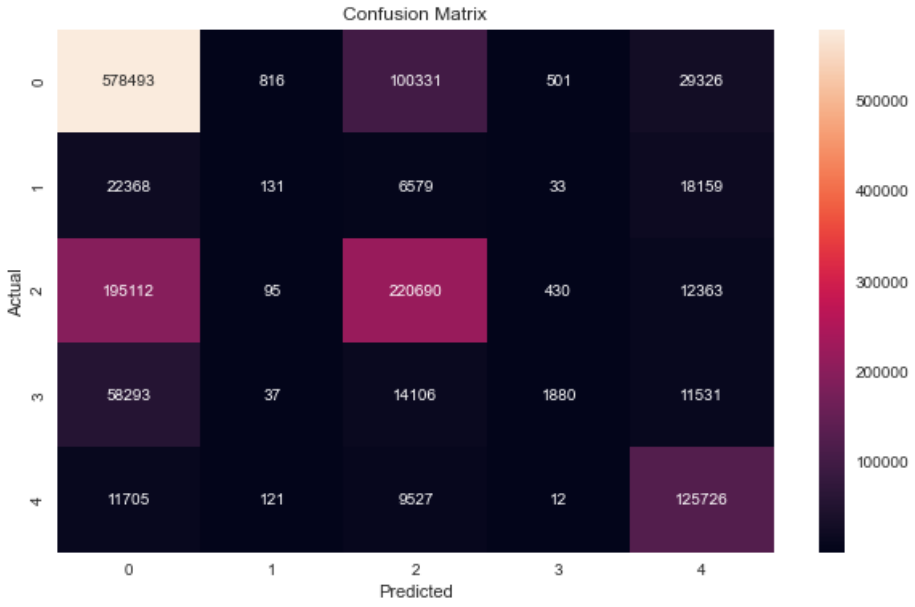
8 Cross-Validation

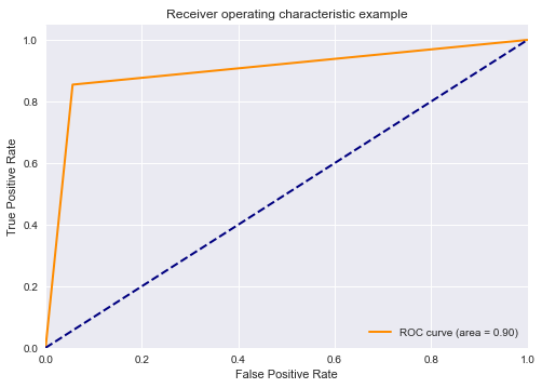
La Cross-Validation en aquest dataset ha presentat problemes sobretot a nivell de cost computacional, fer un entrenament de 5 folds podia extendre's a més de 25 minuts d'execució per acabar obtenint una precisió molt similar a la que ja teniem. Per temes de cost computacional només s'han fet proves amb 3 i 5 folds.

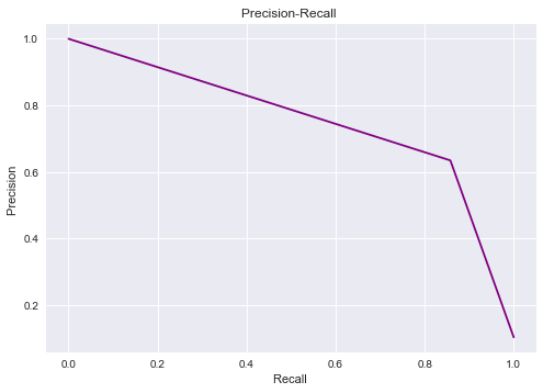
8.1 Random Forest amb 3 folds

8.1.1 Dataset sencer

L'accuracy obtinguda és de nou similar a les anterior 65%. El temps emprat en fer tota la cross-validation és d'uns 10 minuts.







Corbes Precision-Recall i ROC per a un OnevsAll de la classe 4

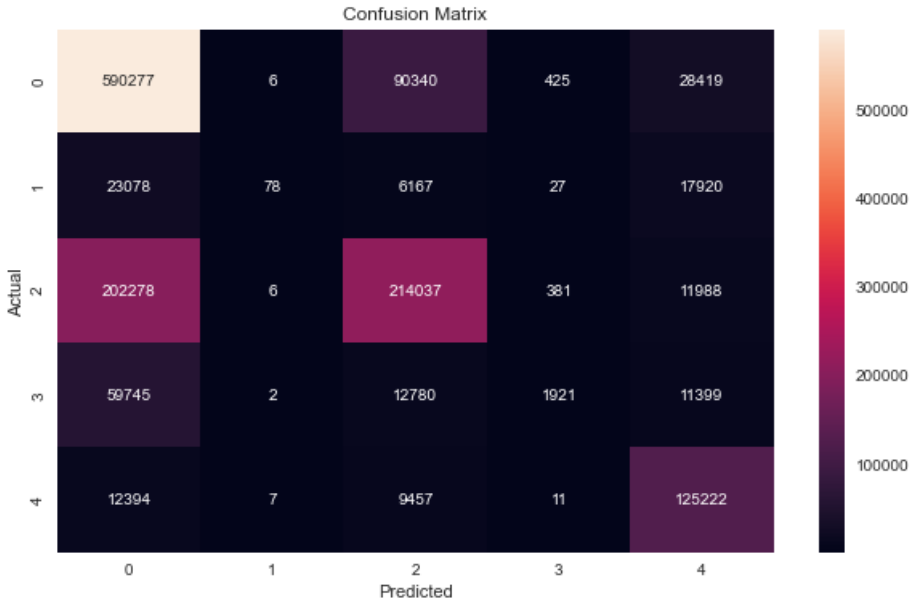
8.1.2 Dataset després de Feature Selection

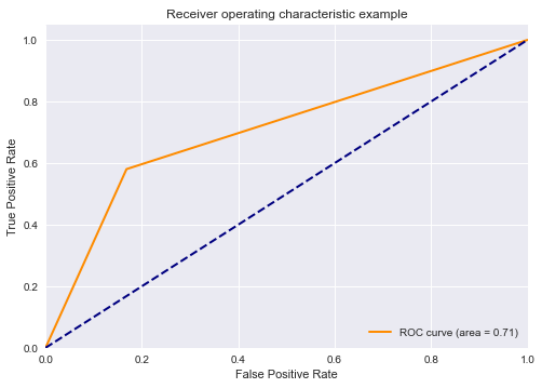
L'accuracy ara és inferior a les anteriors, d'un 56%. És possible que la raó d'això és que encara que no siguin columnes molt importants per fer un entrenament, però alhora d'explicar totes les dades sí que es perd informació important.

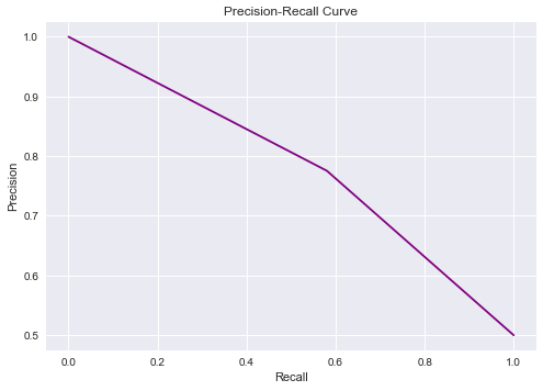
8.2 Random Forest amb 5 folds

8.2.1 Dataset sencer

S'obté una accuracy aproximada de 66%, el temps que triga en acabar de fer la validació és d'uns 25-30 minuts,







Corbes Precision-Recall i ROC per a un OneVSAll de la classe 0.

9 Hyperparameter Search

De nou, per temes de cost computacional s'ha deicidit que fer una cerca d'hiperparametres amb aquest dataset no té gaire sentit donat el temps disposat. Tenint en compte que una regressió normal pot arribar a trigar varis minuts, una cerca d'hiperparàmetres es podria allargar a varies hores fàcilment. De totes maneres, com el model utilitzat és un Random Forest, els paràmetres importants ja han estat avaluats anteriorment (a partir de 40 arbres aproximadament no varia i a partir de 15 de profunditat s'assoleix overfitting).

10 Conclusions

10.1 Conclusions del dataset

Amb tot el treball fet es poden concluïr algunes coses d'aquestes dades.

- KNN, SVM i classificadors similars no són adecuats per aquest dataset.
- Tant la localització en coordenades i en descripció com el temps són atributs molt importants en termes de predir el tipus de crim comés.
- No sembla que aquest dataset sigui bo per fer regressions o classificacions, ja que és difícil obtenir bons resultats sense forçar massa l'atribut objectiu.
- És un dataset interessant per fer un anàlisi.
- Sembla difícil superar un cert llindar de precisió, la raó d'això segurament sigui que moltes classes són molt difícils de diferenciar entre elles.

10.2 Conclusions personals

Crec que amb una mica més de temps podria haver fet algunes coses que m'han quedat pendents, com per exemple la cerca d'hiperparàmentres o altres models/atributs objectiu. Penso que el dataset és molt interessant però que justament per classificació/regressió no ho és tant ja que s'obtenen precisions molt baixes. He disfrutat bastant fent l'anàlisi de dades previ a la classificació, penso que podria haver indagat més en el tema de coordenades/temps si hagués tingut més temps també. En part també he perdut molt de temps esperant a que el codi s'acabés d'executar ja que amb tantes dades qualsevol comanda trigava bastant.