Dataset: San Francisco Crime Classification[1]

1 – Descripció del dataset. Per què és important i quina pregunta/problema pretén respondre

El conjunt de dades que he utilitzat conté dades sobre la delinqüència a la ciutat de San Francisco entre els anys 2003 i 2015. Les dades les he descarregat de la plataforma Kaggle, ja que formaven part d'una competició que va finalitzar el juny de 2016. No obstant això, les dades provenien del portal de dades obertes de la ciutat de San Francisco.

L'anàlisi de la criminalitat en aquesta ciutat és rellevant perquè San Francisco ha estat una de les ciutats amb taxa de delictes més gran dels EUA. La competició consistia en intentar esbrinar la tipologia delictiva a partir d'un moment i un lloc. Aquesta informació podia ser rellevant per un futur, per predir la delinqüència o per intentar aportar una informació més detallada als agents que s'adrecen a cobrir fets delictius. A més, també es proposava als participants realitzar visualitzacions de les dades.

Com es pot veure a l'execució del notebook ipython, la capçalera i als descriptors, el dataset té 9 camps i 878.049 registres.

Per una part tenim la informació sobre el **moment dels fets**, distribuïda en dos camps diferents:

- *Dates* indica els dies i hores dels fets i n'hi ha 389.257 valors únics.
- *DayOfWeek* com el seu nom indica, recull el dia de la setmana del fet.

Per una altra part, tenim informació sobre el **tipus del fet**:

- *Descript* recull les 39 categories o grups de delictes en què s'inclou cada fet.
- *Category* identifica a quina de les 879 tipologies de delictes correspon el fet de cada registre.

Pel que fa als fets, també tenim la informació sobre la **resolució** d'aquell delicte, amb un sol camp:

- *Resolution* com s'ha comentat mostra si s'ha resolt o com s'ha resolt el fet delictiu, amb 17 valors diferents.

Per últim, tenim quatre camps relacionats amb el **lloc dels fets**:

[1] https://www.kaggle.com/c/sf-crime

- *District* informa sobre en quin dels 10 districtes de la ciutat han passat els fets.
- *Address* conté la informació sobre l'adreça del lloc, i n'hi ha 23.228 valors únics.
- *X* i *Y* fan referència a les coordenades geogràfiques. Tot i tractar-se d'atributs diferents han d'analitzar-se de manera conjunta. Són els únics atributs de tipus numèric del dataset.

2– Integració i selecció de les dades d'interès a analitzar.

De l'anàlisi previ, cal destacar que les dades estan en format de text (*str*), per tant una primera acció que cal realitzar és passar-les a un format de data que es pugui analitzar. A més també s'ha considerat necessari dividir la informació en dos camps, un per al dia i un altre per l'hora, i afegir-ne una categoria per diferenciar entre nit, matí, tarda i vespre.

De les categories delictives hi ha algunes que no es consideren rellevants, ja sigui per no ser penals, o per tenir una menor rellevància, i s'han descartat del conjunt de dades.

Més enllà d'aquestes dues qüestions, el conjunt de dades no conté variables que no puguin ser explotables, per tant, la selecció és de tot el conjunt de variables. A més es tracta d'un únic arxiu, en conseqüència tampoc no cal realitzar una integració de diversos conjunts de dades.

No obstant això, pel que fa a la integració, per realitzar una anàlisi completa i acurada es podrien cercar conjunts de dades complementaris per integrar amb el de la delinqüència. Per una part, pel que fa a la informació cronològica, podria ser rellevant un calendari dels dies festius i, fins i tot, d'alguns esdeveniments rellevants de la ciutat (fires, congressos, esdeveniments esportius o concerts en què hi hagi grans concentracions de persones). També podria ser interessant, completar la informació geogràfica per districtes (o barris) amb conjunts de dades sobre la densitat de població, la renda per càpita, la distribució del vot o el nivell d'estudis. Per últim, una informació que seria més difícil d'aconseguir però que internament en un anàlisi policial es podria obtenir seria la relacionada amb els efectius, els patrullatges o les operacions policials, distribuïdes al llarg del temps i de la ciutat.

Malauradament no he pogut destinar temps a la recerca d'alguna d'aquestes dades per poder integrar-les al conjunt de dades.

3– Neteja de les dades.

3.1– Les dades contenen zeros o elements buits? Com gestionaries aquests casos?

En el conjunt de dades no hi ha elements buits i, tot i que no hi ha zeros, sí que un dels camps de text té valors que podriem considerar com a tals. Es tracta del camp *Resolution* en el qual el valor més freqüent és **None**. Malgrat que no hi ha una informació sobre el tancament del cas, el fet que encara resti obert, ja es pot considerar un indicador, per tant, cal mantenir la integritat d'aquest camp i no eliminar els registres que tenen aquest valor. Això no descarta que més endavant es pugui fer alguna anàlisi sense aquest indicador.

Una hipòtesi que em plantejo és que abans de publicar-se les dades, ja sigui a la plataforma Kaggle o al portal de dades obertes de la ciutat de San Francisco, ja s'hagi fet algun tractament d'aquests casos. Considero molt factible que aquells registres sense una informació geogràfica determinada, especialment sense una ubicació amb coordenades X-Y, hagi estat descartada i eliminada del conjunt de dades. També considero possible que en alguns altres casos, si no es tenia aquesta ubicació geogràfica precisa, però si es tenia una adreça, aquesta s'hagi establert a partir d'aquest punt.

3.2– Identificació i tractament de valors extrems

En els descriptors de les dades, entre els valors de la variable "Y" s'observa que el màxim és 90, quan la majoria de les dades són properes al 37. Això indica que clarament hi ha un o més registres amb un valor extrem, pel que fa a la ubicació geogràfica. Entre els valors de la variable "X", la distorsió no és tan gran, però sí que hi ha un o més registres amb el valor -120, quan la majoria dels valors són propers al valor -122. Atès que els objectius estan directament relacionats amb la ubicació geogràfica, els registres d'aquests valors extrems s'han d'ometre i obviar de cara a l'anàlisi.
Amb una ràpida visualització es pot observar com aquests valors extrems estan molt concentrats, per tant es podrien eliminar els registres que els contenen establint un marge ampli, encara que no es correspongui estrictament amb les coordenades geogràfiques de la ciutat.

Entre les variables categòriques, algunes tenen una freqüència molt baixa i, en certa mesura, també es podrien considerar valors extrems. Ara bé, atès que aquests camps, especialment els de *Category* i *Descript* són els objectius de la classificació. Més enllà de la primera tria que s'ha fet, descartant alguns valors, en aquest punt també s'ha optat per crear una nova variable *Descript2* que agrupi dels valors de la variable *Descript* amb una menor freqüència dins de la seva categoria (he escollit aquelles que representen menys d'un 5%).

4– Anàlisi de les dades.

4.1– Selecció dels grups de dades que es volen analitzar/comparar (planificació de les anàlisi a aplicar)

Com hem comentat anteriorment, la variable de l'adreça en aquests moments no ens és útil. De la mateixa manera que ja no ens calen les variables *DescriptFr*i *CategoryFr* que mostren, respectivament, les freqüències dels camps *Descript* i *Category*. A més, atès que la informació del moment del fet l'hem treballada i disgredada, tampoc no cal ni el camp original *Dates*, ni el camp *Days* amb la mateixa informació amb format de temps.

Les possibiltats de comparació i de creuament de dades són moltes però entre d'altres es podrien analitzar:

- Dies i franges horàries amb major nombre de fets
- Similituds o diferències en els districtes en funció de dia i la franja horària
- Similituds o diferències en les categories de fets de fets en funció de la franja horària
- Fets o categories de fets amb més resolucions
- Districtes on es resolen més delictes
- Estacionalitat dels delictes al llarg dels anys

4.2– Comprovació de la normalitat i homogeneïtat de la variància

No l'he pogut analitzar més enllà d'extreure alguna mètrica que es pot veure al document d'ipython.

4.3– Aplicació de proves estadístiques per comparar els grups de dades. En funció de les dades i de l'objetiu de l'estudi, aplicar proves de contrast d'hipòtesis, correlacions, regressions, etc.

No l'he pogut analitzar més enllà d'extreure alguna mètrica que es pot veure al document d'ipython.

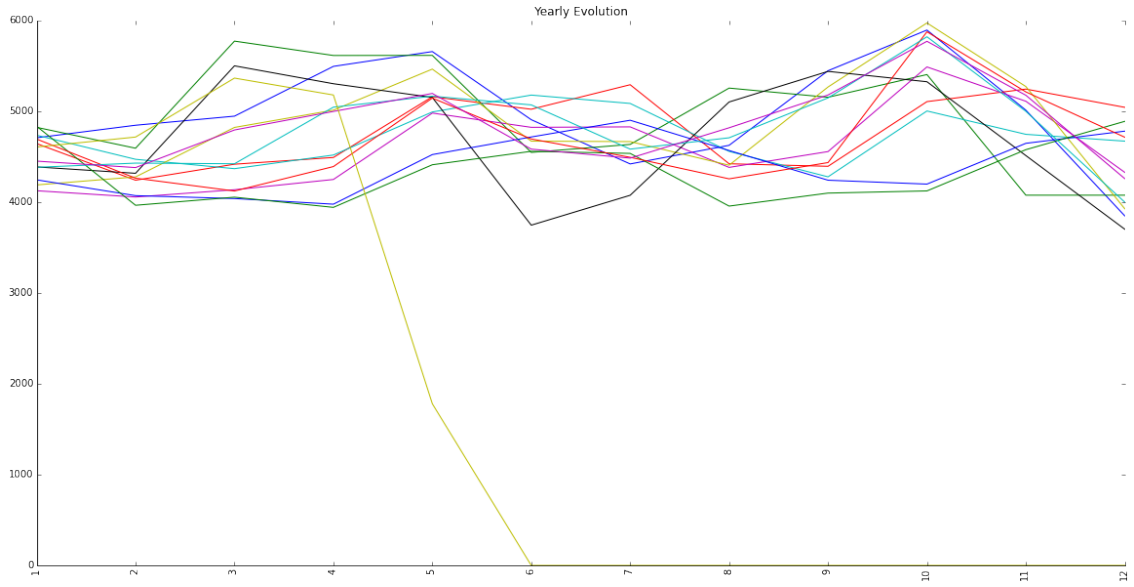
5 – Representació dels resultats a partir de taules i gràfiques

Veure les gràfiques al document ipython

6 – Resolució del problema. A partir dels resultats obtinguts, quines són les conclusions? Els resultats permeten respondre al problema?

Malauradament no pogut dedicar prou temps a la pràctica com per obtenir grans respostes que permetin solucionar el problema.

Potser un dels resultats més destacats és la manca d'una estacionalitat en els fets analitzats, malgrat que potser una anàlisi per tipologies delictives permetria matisar aquesta afirmació.



7– Codi:

El codi està disponible a:

https://github.com/hitnas/hitnas-TCVD_Prac2_Neteja_validacio