Progetto Machine Learning

Team 28: Boyuan Zhang, Stefano Fiorini, Federico Maria Viotti, Andrea Ongaro, Massimiliano Scardovelli Cracchiolo

Abstract
Il tasso di criminalità in una città è parte fondamentale per definire la qualità della vita. Gestire al meglio i crimini, riuscire a prevederli può essere fondamentale per cercare di ridurlo. In particolare l'obiettivo di questo report è quello di individuare, attraverso un algoritmo di Machine Learning, il miglior modo di classificare o prevedere due particolari attributi: la tipologia del crimine e la risoluzione del crimine. Il report, dopo un analisi esplorativa del dataset, espone i risultati dei classificatori utilizzati, confrontandoli e individuando quello migliore.
In aggiunta, attraverso l'association analysis, il report cerca di individuare, dato un crimine rilevante e il distretto, se esiste un'associazione tra essi.

Contents

1. Introduzione

San Francisco crime è il dataset scelto dal nostro gruppo di lavoro.
La città di **San Francisco** è situata all'estremità nord dell'omonima penisola. Essa si estende dalla costa occidentale degli Stati Uniti (con una superficie di 120,9 km^2) e costituisce la parte più occidentale della più vasta regione geografica e urbana chiamata *San Francisco Bay Area*.

La città è famosa per la ricchezza di etnie e culture molto diverse tra loro. La sua popolazione stimata nel 2015 è di 864.816 abitanti, stima che la colloca al dodicesimo posto fra le città più popolose degli Stati Uniti d'America e allo stesso tempo al secondo posto per densità di popolazione, dietro a New York. Fa parte di una vasta area metropolitana (circa 7 milioni di abitanti, la quinta dell'intero Paese), la San Francisco Bay Area, di cui è sempre stata il centro economico-finanziario, culturale e turistico, anche se ha ormai perso il primato di popolazione.

1.1 Dataset

Il dataset che verrà analizzato è composto da 878.049 righe e 9 attributi che riportano una complessiva descrizione dei crimini commessi nella città di San Francisco e dintorni dal 2003 al 2015. Il dataset contiene sia dati temporali (data e ora) che dati spaziali (distretto, indirizzo e coordinate geografiche); inoltre sono presenti una descrizione del reato, la categoria a cui esso appartiene e il provvedimento preso dalle autorità. Non sono presenti invece ne dati mancanti ne descrizioni delle persone coinvolte. Tutti gli attributi del dataset sono di tipo categoriale tranne le due coordinate spaziali (latitudine e longitudine). Si osserva che, probabilmente per motivi di sicurezza pubblica, non sono presenti nel dataset informazioni relative ad omicidi e/o crimini efferati.

1.2 Obiettivo dell'analisi

Come scopo dell'analisi di questo dataset si è pensato di creare un workflow in grado, attraverso l'utilizzo di diversi algoritmi di machine learning, di risolvere due differenti obiettivi: il primo è quello di riuscire a inferire la tipologia dei reati commessi a San Francisco sulla base delle informazioni spaziali e temporali; in secondo luogo si è deciso di provare a prevedere se il crimine commesso comporterà un arresto o meno da parte delle autorità.

2. Esplorazione dei dati

2.1 Frequenze

Per comprendere la distribuzione del nostro dataset, abbiamo deciso di visualizzare la distribuzione delle frequenze delle variabili: *'Category'*, *'PdDistrict'* e *'Resolution'*.

District Il *'Southern'* è il distretto che ha frequenza più elevata all'interno del dataset(157.182), con una percentuale del 18%. Al contrario, il distretto di *'Richmond'* ha una frequenza assoluta pari a 45.209 che corrisponde a una percentuale del 5%.

Category Per quanto riguarda la categoria del crimine riscontrata nel dataset, *'Larceny/Theft'* è il più frequente(174.900), con una percentuale del 20%. Un terzo dei reati è riconducibile a 2 categorie: *'Other Offenses'* e *'Larceny/Theft'*. Inoltre un 10% viene etichettato come *'Non-Criminal'*, pertanto non vi è associato alcun reato.

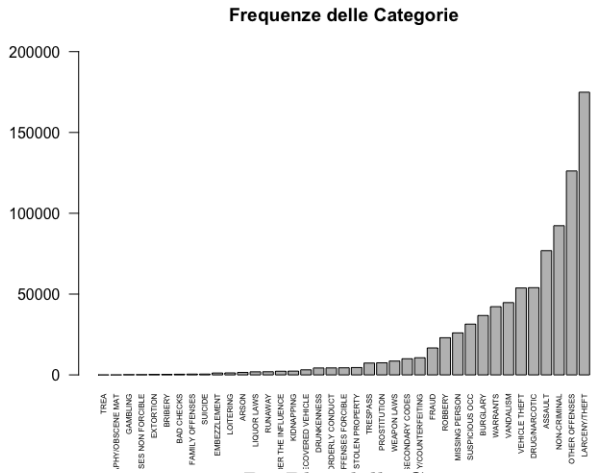


Figure 1. BarPlot delle Categorie

Resolution Più della metà dei dati è riconducibile a una *'None'* Resolution. Possiamo interpretare *'None'* come l'impossibilità da parte della Polizia di intraprendere determinate azioni nei confronti del trasgressore. *'Arrest/Booked'*, la seconda risoluzione più frequente, rappresenta il 29% dei dati.

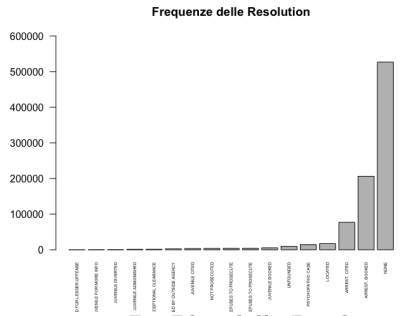
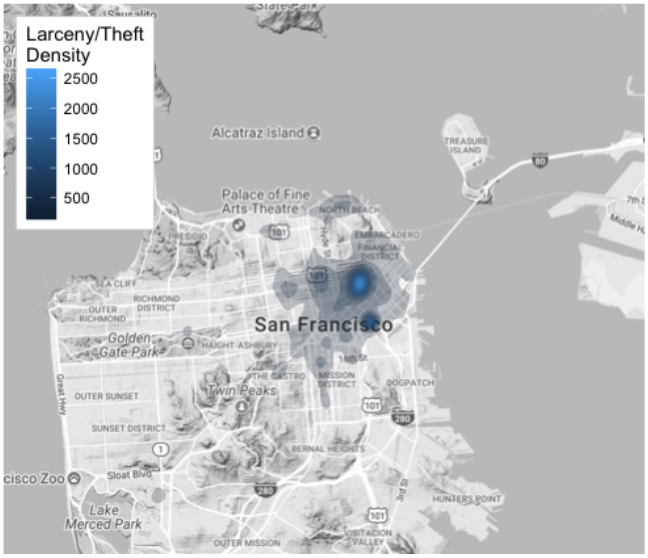


Figure 2. BarPlot delle Risoluzioni

2.2 Mappa delle densità

Rappresentiamo la densità della distribuzione della categoria più frequente sulla mappa di San Francisco, cioè il numero di reati registrati rispetto alle loro coordinate geografiche.

Larceny/Theft in San Francisco



3. Preprocessing

Una prima operazione sul dataset è stata la verifica della presenza di eventuali *missing values*, utilizzando un'appropriata funzione di R che ha dato un riscontro negativo. In seguito abbiamo trasformato la variabile *'Dates'* dividendola in più attributi (anno, mese, giorno e ora) così da poter utilizzare anche separatamente le informazioni temporali nella classificazione.
Il nuovo attributo *'Hour'*, in formato 'ora : minuti : secondi' è stato poi trasformato in numeri interi. Abbiamo poi raggruppato i records di questo attributo in 5 fasce orarie rispetto ad una eguale frequenza, utilizzando il nodo **auto-binner** di Knime. L'output del nodo sono delle fasce di tipo stringa che trasformiamo in tipo numerico.
L'attributo *'PdDistrict'*, che rappresenta il distretto della città in cui il singolo reato è stato commesso, è stato binarizzato così da renderlo utilizzabile dai classificatori che accettano in input dati continui o binarizzati.

Entrambi gli attributi *'Category'* e *'Resolution'*, le variabili target, sono state modificate appositamente sulla base degli obiettivi prefissati: è stata dunque creata una colonna *'res_binary'* data dalla binarizzazione dell'attributo *'Resolution'* in arrestato/non arrestato. L'attributo *'Category'*, composto da 40 diverse tipologie di reato è stato invece ridotto nella sua dimensionalità a 22 categorie, raggruppando i crimini sulla base di sogget-

tive considerazioni sulle similitudini tra i reati(denominando la nuova variabile come *'CatGroup'*).

Per continuare la nostra analisi e applicare i modelli predittivi di classificazione abbiamo effettuato una partitioning primaria del nostro dataset in training e test,suddividendoli al 67 %, mediante campionamento stratificato sulla variabile *'DayofWeek'* e impostando un seme randomico per la riproducibilità dell'esperimento. Successivamente abbiamo suddiviso il training set derivante dalla partizione primaria in altre due parti, seguendo la stessa metodologia, con il fine di validare i modelli.

4. Modelli

Per lo studio dell'attributo di *'Category'* sono stati usati tali classificatori:

- **Modelli Probabilistici:**Naive Bayes (NBY), Naive Bayes Tree (NBT) , Naive Bayes Multinomial (NBM)
- **Modelli Euristici:** Random Forest di Weka, Albero di regressione (j48) con 20 alberi

Per quanto riguarda l'attributo *'res_binary'*:

- **Modelli Probabilistici:** Naive Bayes Tree (NBT)
- **Modelli Euristici:** Random Forest di Weka, Albero di regressione (j48), Decision Tree, Random tree, Gradient Boosted Tree

I modelli sono stati scelti secondo il criterio del *no-best-model* e sono stati validati attraverso una *K-cross validation* con k = 10

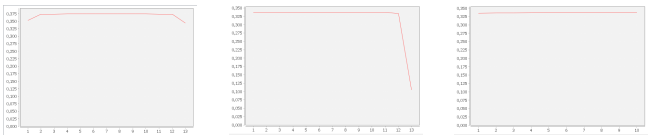
5. Category

5.1 Feature Selection

Scelti i modelli abbiamo deciso di applicare una *feature selection* su ogni singolo modello per vedere la combinazione di attributi migliore per raggiungere la massima accuratezza e per, a parità di essa, diminuire il peso computazionale utilizzando meno attributi possibili e eliminando quelli che non migliorano i modelli. Inoltre ciò è stato fatto per semplificare il modello e renderlo più interpretabile, per evitare la *curse of dimensionality* e ridurre l'*overfitting*.
Abbiamo applicato questa operazione sulla partizione primaria per fare in modo che una riduzione della quantità di dati su cui applicarla non ne precludesse la validità.
Abbiamo deciso di utlizzare il nodo chiamato **Backward feature Elimination** così da, partendo con tutte le feature del nostro dataset, rimuovere le meno significative ad ogni iterazione che migliora l'accuratezza, ripetendo il processo finchè non si arresti l'aumento della performance ad una rimozione delle variabili.

I processi di feature elimination hanno dato i seguenti risultati (Figure: 3)

Come si evince dai grafici aggregati la variazione di accuratezza al variare del numero di features è pressochè



(a) J48-NBayes-NBMultinomial

(b) NBtree-RandomForest

Figure 3. LineChart dei modelli

inesistente; solo in alcuni modelli l'utilizzo di una sola o tutte le variabili altera in maniera più o meno consistente la perfomance.

Alla luce di questa osservazione, abbiamo deciso dunque di optare per la selezione del minor numero possibile di variabili così da ridurre il tempo di training e semplificare l'interpretazione dell'analisi.

5.2 Validazione del modello

Con i risultati ottenuti dalla *feature selection* abbiamo implementato sulla partizione secondaria una *cross validation* per ogni modello. Questo approccio ci ha permesso di verificare quale sia l'algoritmo che raggiunge un'accuracy migliore riducendo i problemi di overfitting ma anche quelli di campionamento asimmetrico del training set, tipico della sua suddivisione in due parti(ovvero training e validation set).

La *cross validation* è una tecnica che suddivide in k-parti, di uguale numerosità, il dataset. Si implementano k iterazioni e in ognuna di esse la k-esima parte del dataset viene considerata validation dataset e la restante parte costituisce il training dataset. Abbiamo esposto i risultati di questa procedura su un BoxPlot (Figure: 4),

il quale mostra che tra tutti gli algoritmi selezionati, il modello di **weka j48** ha un'accuratezza complessiva maggiore(in media pari a 0.374).

Abbiamo quindi utilizzato tale modello per la predizione del nostro dataset utilizzando la partizione primaria:

Table 1. Risultati J48

Indices	Results
Correct Classified	251.418
Wrong Classified	336.874
Accuracy	42,737%
Error	57,263%
Cohen's kappa(k)	0,239

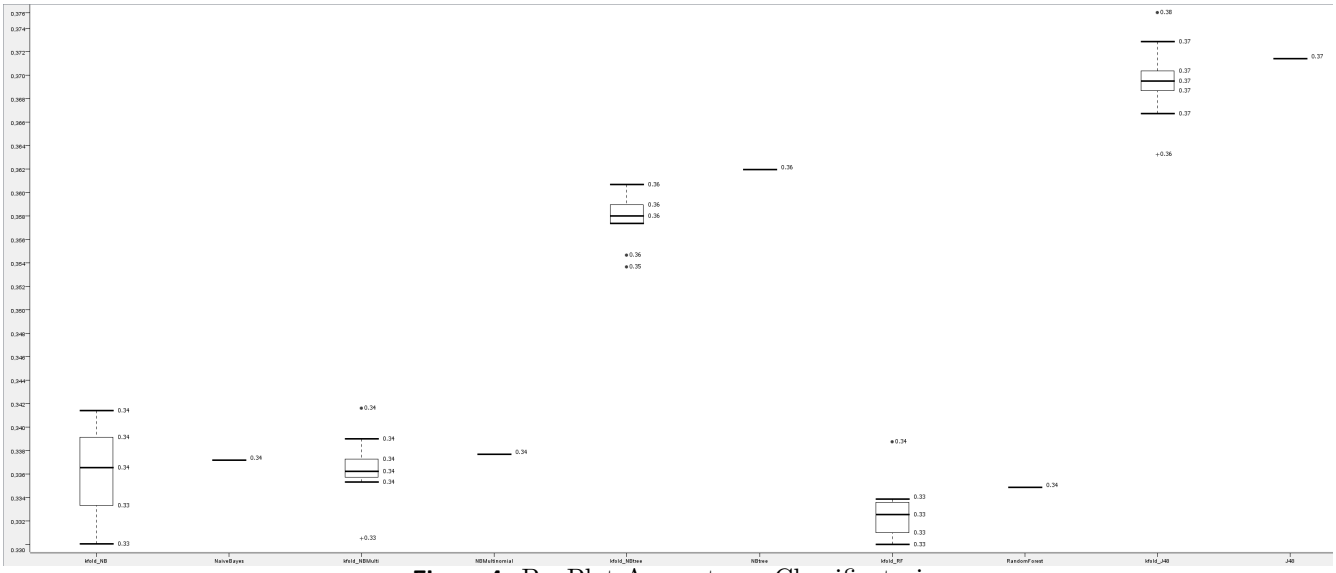


Figure 4. BoxPlot Accuratezza Classificatori

5.3 Intervalli di confidenza

Per dare una valenza statistica al risultato osservato nella *cross-validation* abbiamo deciso di comparare i due modelli con accuratezza maggiore attraverso il calcolo degli intervalli di confidenza degli errori sulla base della Distribuzione T-Student.

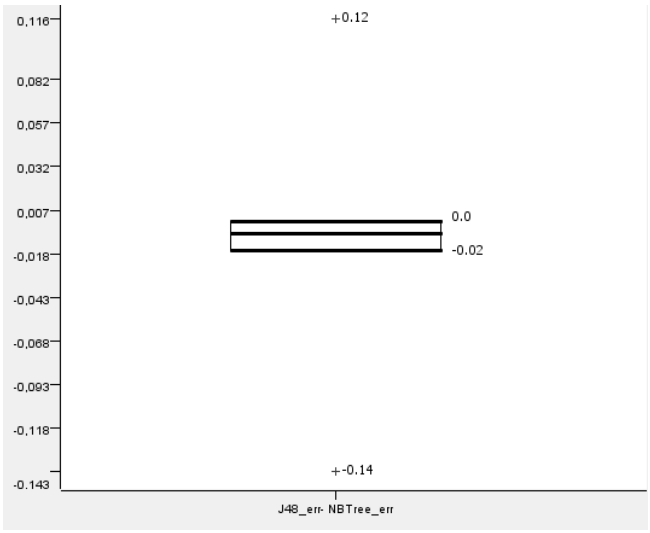


Figure 5. Intervallo di Confidenza

A livello di confidenza 0,95 i due modelli di classificazione non sono significativamente differenti perché lo 0 è contenuto nell'intervallo di confidenza.

6. Res_Binary

6.1 Feature Selection

Anche per la classificazione di questo attributo abbiamo scelto di utilizzare una feature selection(per gli stessi motivi della precedente analisi) ma, differentemente da quanto fatto per la variabile *'CatGroup'*, in questo caso si è scelto di adottare il metodo **filter** che evita gli alti costi computazionali e l'elevato rischio di *overfitting* del metodo *Wrapper*. Questo metodo seleziona gli attributi prima dell'apprendimento del classificatore; in particolare abbiamo utilizzato l'opzione **CFS**(*Correlation Feature Selection*), facente parte della categoria *multivariate*.

Per poter adottare questo criterio abbiamo dovuto escludere la variabile *'Y'* poiché la presenza di un valore doppio non permetteva la buona riuscita della procedura.

Questo procedimento ci ha restituito gli attributi *'X'* e *'TENDERLOIN'* che abbiamo dunque utilizzato per l'analisi.

6.2 Validazione del Modello

Dopo la *feature selection* abbiamo operato una *cross-validation* per validare i modelli scelti e verificare la loro accuratezza. Abbiamo esposto i risultati di questa procedura in un BoxPlot (Figure: 6)

I 6 modelli hanno restituito una accuracy simile, quello migliore però risulta essere il **Weka random Tree** .

6.3 Recall

Si deve però considerare il problema dello sbilanciamento della classe *'res_binary'*. Questo introduce il problema della *class imbalance*, ovvero la tendenza degli algoritmi di classificazione a dare più peso alla classe maggiormente presente nell'attributo a discapito dell'accuratezza dell'altra, che potrebbe essere quella di maggior interesse per l'analisi, generando problemi di *overfitting* e di interpretazione dei risultati.

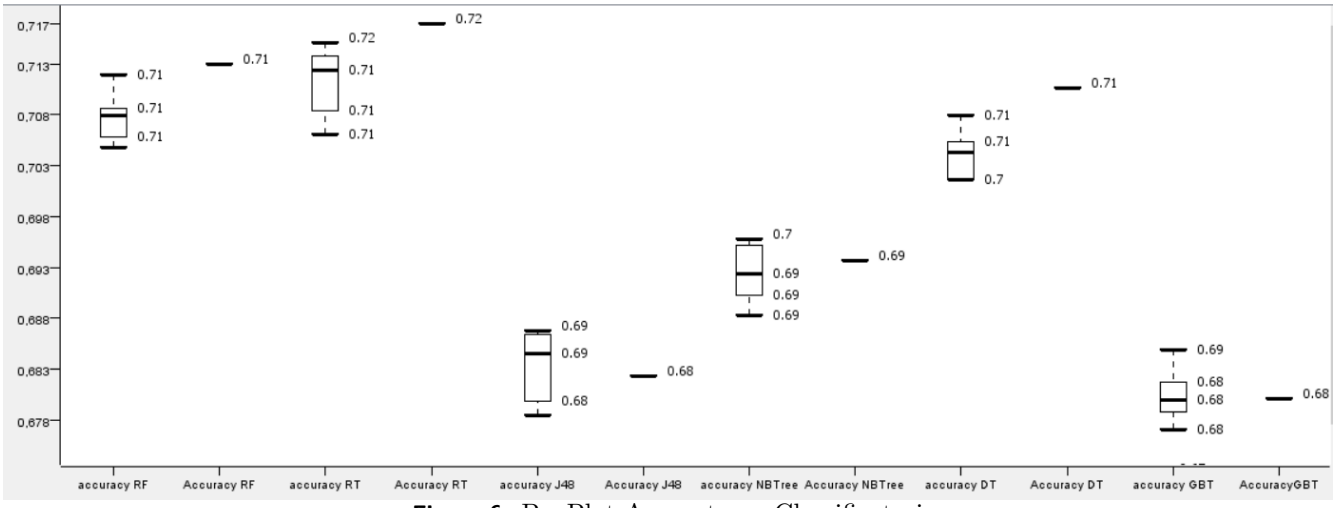


Figure 6. BoxPlot Accuratezza Classificatori

	Accuracy	Recall	Precision	F1
GBT	0,682	0,209	0,662	0,317
NBTW	0,700	0,334	0,650	0,441
RFW	0,721	0,493	0,637	0,555
DTL	0,722	0,487	0,642	0,554
RTW	0,723	0,481	0,645	0,551
J48W	0,685	0,252	0,640	0,362

Table 2. Risultati dei Classificatori

Nel nostro caso abbiamo un subset di 588.292 record in cui, riferendoci all'attributo *'res_binary'*, si osserva che il 64,57% dei record presenta il valore 0 mentre i restanti sono 1. Si può quindi dedurre che non ci sia un problema di sbilanciamento così significativo. In ogni caso è bene osservare la qualità dei modelli relativa alla classe più rara; le misure più utilizzate sono la precision (una misura dell'esattezza della classificazione), la recall (il numero di record classificati correttamente sul totale dei record positivi presenti nel dataset) e l'indicatore F1 (una media armonica tra precision e recall). Dando a 1(record positivo) il significato di 'arrestato' si può fare un'interessante osservazione: quanti possibili arresti riesco a prevedere così da fornire questa informazione al sistema carcerario al fine di conoscere il numero stimato di possibili nuovi carcerati? Per rispondere a questa domanda abbiamo deciso di utilizzare la recall. Questo indice ci permette di osservare che, nonostante l'algoritmo *Random Tree* dia un'accuratezza maggiore, esso non sia quello che ci da la recall più alta; è infatti l'algoritmo *Random Forest* a fornirci la recall maggiore(49,3%).

6.4 ROC Curve

Dalle curve ROC(Figure: 7) si evince che tutti i modelli utilizzati, per quanto riguarda il rapporto tra *True Positive Rate* e *False Positive Rate*, sono buoni modelli in quanto le aree sottese a esse sono sensibilmente maggiori dell'area sottesa alla retta rappresentante il modello casuale ($y = x$ con y e x rispettivamente TPR e FPR).

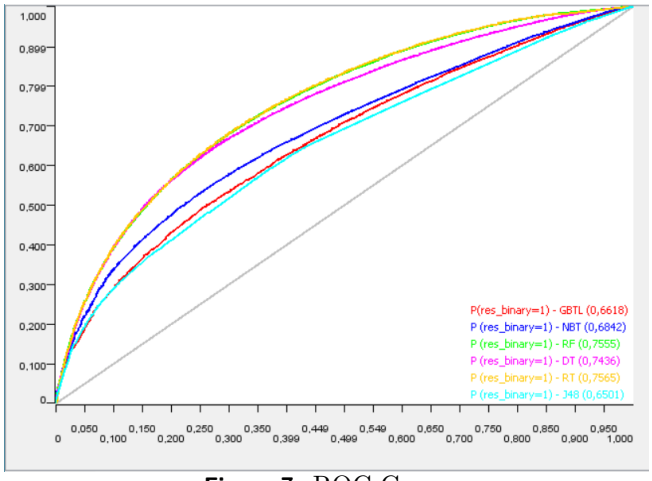


Figure 7. ROC Curve

In particolare le curve migliori sono quelle dei modelli Random Tree e Random Forest mentre il peggiore è J48

7. Association Analysis

Come ultima analisi effettuata sul dataset abbiamo voluto verificare la regola di associazione tra:

- *CatGroup* e *PdDistrict*;
- *Resolution* e *PdDistrict*;
- *CatGroup, PdDistrict e Resolution*;

Per individuare queste relazione abbiamo utilizzato la libreria *ARules* di R e, sfruttando l'algoritmo *Apriori*(impostando come supporto minimo 0,001 e come confidenza minima 0,30 per i primi due casi, mentre per l'ultimo caso 0,75), siamo arrivati ai seguenti risultati:

Utilizzando l'indice *lift*, il quale fornisce una misura della dipendenza tra le variabili considerate nella regola, siamo giunti alla conclusione che le uniche regole associative significative, rispetto alla prima e alla seconda associazione, sono:

Table 3. Risultati AA

Componenti					
LHS	RHS	Supporto	Confidenza	Lift	Count
CG=10	PdD=Mission	0,004	0,485	3,551	3.629
CG=13	PdD=Tenderloin	0,021	0,301	3,234	18.238

- CG:**M**[1] ⇒ Distretto **Mission**
- CG:**P**[2]⇒ Distretto **Tenderloin**.
- Distretto **Tenderloin** ⇒ Res: **Arrest, Booked**

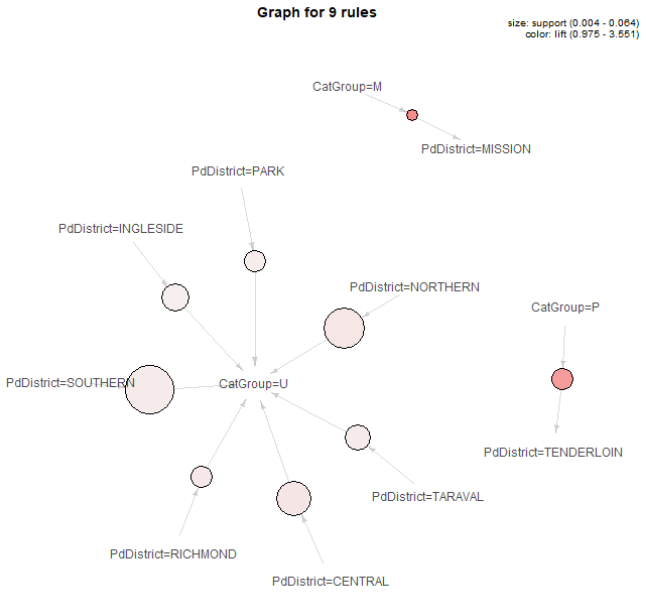


Figure 8. Grafo Association:CG e Distretto

Una spiegazione di queste regole può essere data approfondendo la conoscenza del dominio. Il quartiere *'Tenderloin'* è noto per essere uno dei più pericolosi della città, quindi per la seconda e terza regola individuata abbiamo una corrispondenza con il dominio. Il quartiere *'Mission'* invece è a prevalenza latina ed è considerato la zona della città bohémien, il che non fornisce una spiegazione concreta alla regola associativa ottenuta. Pertanto quest'ultima può avere una valenza informativa a differenza delle altre. Per la terza regola associativa non abbiamo riscontrato evidenze significative aggiuntive.

8. Conclusione

Lo scopo di questo elaborato è stato quello di analizzare i dati relativi ai crimini commessi a San Francisco dall'1/1/2003 al 13/5/2015. Si è voluto, in prima analisi, classificare le tipologie di crimine avvenute sulla base delle informazioni spaziali e temporali, preventivemente aggregate per ridurne il numero troppo alto per poter essere processato e, in seconda anlisi, sono state adottate tecniche di Machine Learning per prevedere la risoluzione dei crimini nelle due possibilità arrestato/non arrestato; sono state successivamente effettuate delle analisi associative per spiegare particolare fenomeni che si potrebbero manifestare all'interno dei confini della città. Per completare l'analisi sono stati scelti più algoritmi di classificazione in una logica *'no-best-model'* e i risultati sono stati validati e raggiunti per mezzo di procedimenti di *cross validation* e *feature selection* con l'obiettivo di ridurre *overfitting* e costo computazionale e di aumentare la facilità di interpretazione di tutto il lavoro.

Le maggiori difficoltà incontrate nel progetto sono state:

- **Preprocessing**: il dataset era abbastanza grande e le variabili, pressochè tutte qualitative, sono state quindi opportunamente modificate perchè avevano un formato che non ne consentiva l'analisi.
- **Calcoli**: la grandezza del dataset e alcune tecniche, come quelle di *feature selection*, hanno generato una difficoltà di calcoli per le macchine utilizzate che ci ha portato a scegliere di implementare alcune parti dell'analisi mediante l'utilizzo congiunto del software di riferimento (Knime) e del linguaggio R, per nostra esperienza migliore a livello di costo computazionale.
- **Features**: le variabili del dataset, come prima riportato pressochè tutte qualitative, non hanno facilitato l'obiettivo di classificazione limitando la scelta dei classificatori. Inoltre, nel caso della classificazione delle tipologie di crimine, essendo queste in numero elevato non è stato possibile superare la soglia di accuratezza del 50%.

Sicuramente i dati e il tema affrontato possono suggerire due ulteriori spunti di analisi:

- Effettuare una *cluster analysis* raggruppando i distretti rispetto alle tipologie dei reati, effettuando una segmentazione della città.
- Si potrebbe applicare una serie storica così da prevedere quando e dove si verificherà un crimine.

[1] Prostitution
[2] Driving under the influence, Drug/Narcotic, Drunkness

References

[1] Wikipedia:
https://en.wikipedia.org/wiki/San_
Francisco.

[2] Wikipedia:
https://en.wikipedia.org/wiki/Mission_
District,_San_Francisco.

[3] Wikipedia:
https://en.wikipedia.org/wiki/Tenderloin,
_San_Francisco

[4] Quora:
https://www.quora.com/
What-is-the-best-way-to-understand-the-terms-precision-and-recall

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

[6] Librerie R:
Arules:https://cran.r-project.org/web/
packages/arules/index.html
GGmap:https://cran.r-project.org/web/
packages/ggmap/index.html
GGplot2:https://cran.r-project.org/web/
packages/ggplot2/index.html