



**UNIVERSITÀ
DEGLI STUDI DI BARI
ALDO MORO**

DIPARTIMENTO DI INFORMATICA

**Corso di Laurea Magistrale in Informatica
Data Mining**

**Association Rule Mining from Spatial Data
for Crime Analysis**

Prof. Michelangelo Ceci

Christopher Piemonte

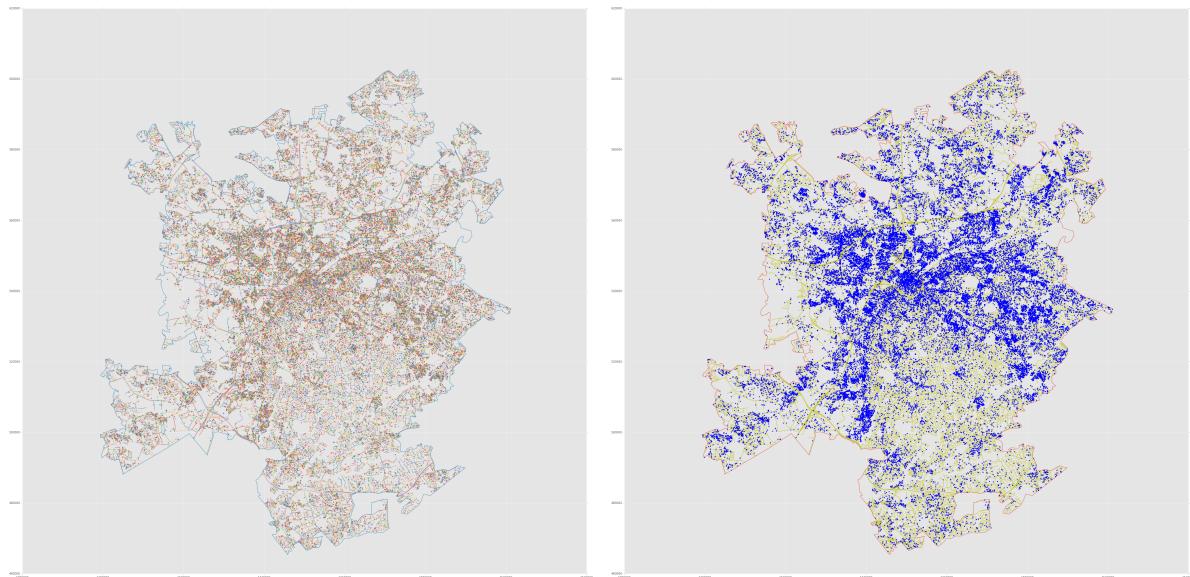
659723

a.a. 2016/2017

1 - INTRO	1
1.1 - CRISP-DM	3
2 - BUSINESS UNDERSTANDING	4
2.1 - Obiettivi primari	4
2.1 - Data Mining Task	4
2.3 - Misure di successo	4
2.4 - Project Plan	5
3 - DATA UNDERSTANDING	6
3.1 - Block Groups	6
3.2 - Blocks	9
3.3 - Boundary	11
3.5 - Business	12
3.6 - Streets	13
3.7 - Crime Count	13
3.8 - Crimes	14
3.9 - Point of Interest POI	17
4 - DATA PREPARATION	19
4.1 - Proposizionalizzazione	19
4.2 - Vicinato	20
4.2.1 - KDtree	20
4.2.2 - Sentence Embedding	20
4.3 - Selezione colonne	21
4.4 - Discretizzazione	23
4.5 - Missing Values	25
5 - MODELING	27
5.1 - Regole di associazione	27
5.2 - Relim	28
5.3 - Configurazioni	28
5.3.1 - Sampled	28
5.3.2 - Entire	30
6 - EVALUATION	31
6.1 - Sampled Dataset	31
6.2 - Entire Dataset	36
6.3 - Conclusioni	38

1 - INTRO

Obiettivo del presente caso di studio è quello di analizzare i dati relativi alla distribuzione spaziale di crimini commessi, nell'arco di un anno, nella città di Charlotte - NC, negli Stati Uniti d'America. Sono presenti dati raccolti da circa 60000 crimini, oltre a dati relativi al censimento dei quartieri e ad informazioni sulle attività commerciali.



Il campione a disposizione è composto da shapefile contenenti sia dati in forma tabellare che informazioni spaziali riguardo a:

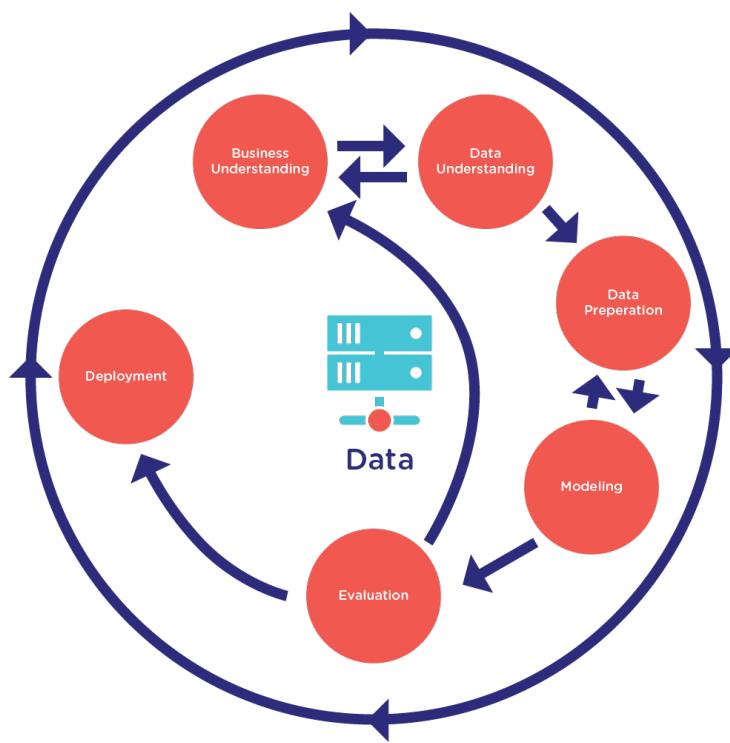
- **POI**: negozi di alcolici, centri commerciali, hotel/motel, parcheggi, walmart
- **Business**: informazioni riguardo le attività commerciali e la loro locazione.
- **blocks**: dati raccolti da ogni famiglia nell'ultimo censimento disponibile in data 2010.
- **Block groups**: costituiscono un raggruppamento di Blocks, e contengono dati raccolti da un campione di famiglie nell'ultimo censimento disponibile in data 2010.
- **crimes**: informazioni riguardo i crimini commessi.

Si intende, pertanto, applicare algoritmi per la scoperta di regole di associazione fra le variabili, in modo cercare associazioni tra le informazioni disponibili. È importante notare come le regole di associazione non descrivono relazioni di tipo causa-effetto,

ma piuttosto l'implicazione evidenziata può essere interpretata come: quando accade una cosa, allora accade anche un'altra.

1.1 - CRISP-DM

La tipologia di processo adottata per l'estrazione di conoscenza è basata sulla standardizzazione aziendale del processo di KDD (Knowledge Discovery in Databases), ovvero il CRISP-DM. Questo mira a sviluppare un processo indipendente dall'industria, dai tools e dalle applicazioni. In seguito saranno descritte le fasi del processo e le iterazioni affrontate.



La sequenza delle fasi non è obbligatoria, è possibile infatti andare avanti e indietro tra le diverse fasi dipendentemente dall'output di ogni fase. Le frecce indicano il cammino più frequente. Il cerchio esterno evidenzia la natura ciclica del processo che può continuare anche dopo il deployment dell'elaborato.

2 - BUSINESS UNDERSTANDING

2.1 - Obiettivi primari

Obiettivo principale è la scoperta di regole di associazione fra i crimini e i dati a disposizione, soprattutto i dati provenienti dai censimenti. In particolare si è interessati alle regole riguardanti le tipologie di crimine in modo da capire come agire per prevenirli. Infatti regole interessanti potrebbero portare alla stanziazione di manovre, attuazioni di misure, tassazioni, incentivi o ancora a ricollocamento delle pattuglie.

2.1 - Data Mining Task

Le regole di associazione sono uno dei metodi per estrarre relazioni nascoste tra i dati. Inizialmente introdotte per la scoperta di regolarità all'interno delle transazioni registrate nelle vendite dei supermercati (market basket analysis), tale informazione veniva utilizzata come base per le decisioni riguardanti le attività di marketing, come ad esempio le offerte promozionali o il posizionamento dei prodotti negli scaffali.

La scoperta di regole si divide in due fasi: l'estrazione degli itemset frequenti e la generazione di regole.

2.3 - Misure di successo

Dopo aver esposto gli obiettivi si forniscono le misure che si dovranno rilevare per il conseguimento degli obiettivi.

Come al solito nella scoperta di regole di associazione siamo interessati ad associazioni con alto supporto ed alta confidenza. È necessario notare come la particolarità del dataset sotto analisi indebolisca il valore del supporto, data la numerosità di crimini riportati e la grande varietà di tipologie di crimine.

2.4 - Project Plan

Gli obiettivi specificati precedentemente saranno perseguiti tramite il processo di KDD. Per la scoperta di regole sarà necessario proposizionalizzare i vari file esplicitando eventuali relazioni spaziali, come ad esempio l'appartenenza ad un quartiere o le attività commerciali nelle vicinanze, pulire il dataset risultante dai missing values e discretizzare. Si procederà seguendo le fasi del processo CRISP-DM una dopo l'altra ciclicamente.

3 - DATA UNDERSTANDING

Il dataset è formato da 12 shapefile, ognuno dei quali formato da file di tipo:

- .shp - il file che conserva le geometrie;
- .shx - il file che conserva l'indice delle geometrie;
- .dbf - il database degli attributi.
- .sbn e .sbx - indici spaziali;
- .prj - il file che conserva l'informazione sul sistema di coordinate
- .shp.xml - metadato dello shapefile;

In seguito si descrivono i vari shapefile, la loro dimensione, gli attributi presenti ed il loro significato.

3.1 - Block Groups

CLT_BlockGroups_Attr

Un Census **Block Group** è una unità geografica usata dal Census Bureau degli Stati Uniti che si colloca tra il Census Tract (distretto, più piccolo della contea) ed il Census Block (isolato).

È la più piccola unità geografica per la quale il Bureau pubblica dati campionati, ovvero dati raccolti sola da una frazione delle famiglie. Tipicamente ogni Block Group ha una popolazione tra 600 e 3000 persone. Ogni Block Group ha un codice FIPS di 12 cifre, il quale è anche l'identificatore.

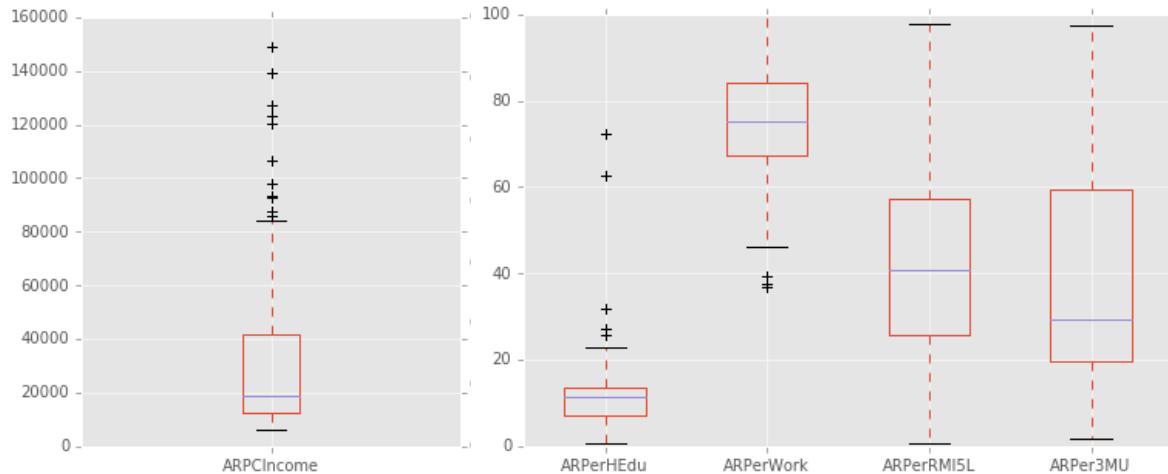
Questo numero influenza l'identificatore dei Census Blocks, i quali aggiungono 3 cifre alle 12 del Block Group.

L'ultimo numero dell'identificatore è il numero di quel Block Groups in quella contea, quindi per esempio il Block Group 2 contiene tutti i Census Blocks da 2000 a 2999.

Dimensione: 485 x 18

Attributi	Descr	Type	Missing Values
GEOID10	ID 12 cifre	Int	-
STATEFP10	ID stato appartenenza	Int	-
COUNTYFP10	ID contea appartenenza	Int	-
TRACTCE10	ID tract appartenenza	Int	-
BLKGRPCE10	ID del block group in quella tratta (una cifra)	Int	-
ALAND10	% di superficie terrestre in metri quadri	Int	-
AWATER10	% di superficie coperta da acqua in metri quadri	Int	-
INTPTLAT10	latitudine del centroide	Real	-
INTPTLON10	longitudine del centroide	Real	-
ARPCIncome	reddito	Real	3 su 485
ARPerHEdu	% istruzione	Real	8 su 485
ARPerWork	% occupazione	Real	3 su 485
ARPerRMI5L	% di pop. che ha affittato e traslocato in meno di 5 anni	Real	37 su 485
ARPer3MU	% di case 3 vani o più	Real	126 su 485

ARPCIncome è un valore numerico reale, mentre gli altri sono valori espressi in percentuale. I Boxplot mostrano la loro distribuzione ed in seguito sono riportati alcuni valori statistici:



	ARPCIncome	ARPerHEd	ARPerWork	ARPerRMI5L	ARPer3MU
count	19106.000000	18550.000000	19106.000000	18785.0000	17515.0000
mean	32143.145295	13.654647	75.291682	41.441861	39.024404
std	29527.596815	13.254983	12.646551	21.316652	27.910962
min	6203.000000	0.554017	36.974790	0.721371	1.809211
25%	12375.000000	7.037037	67.435897	25.912409	19.542421
50%	18560.000000	11.480602	75.030750	40.944882	29.477021
75%	41499.000000	13.636364	83.985765	57.195572	59.402985
max	149122.000000	72.421525	100.000000	97.619048	97.300771

3.2 - Blocks

CLT_Blocks_Attr

Un Census **Block** è la più piccola unità geografica usata dal Census Bureau per la compilazione del 100% dei dati (dati raccolti da tutte le case e non solo da un campione di queste). I Block sono raggruppati in Block Groups. In media ci sono 39 Block per ogni Block Group.

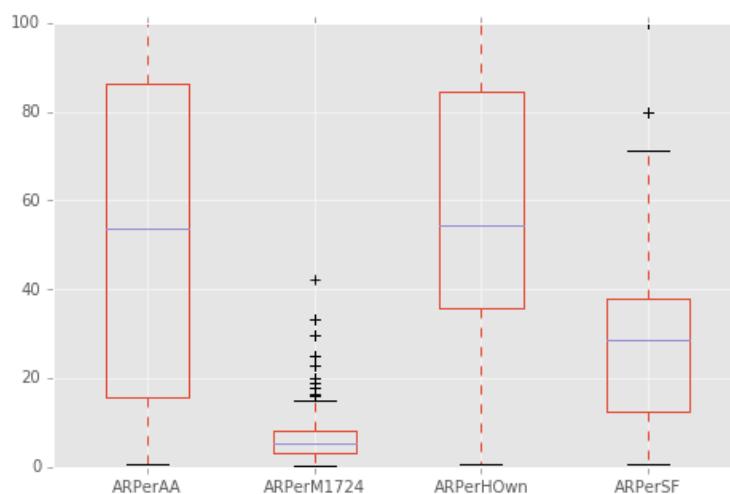
I Census Block hanno un identificativo di 4 cifre, di cui la prima cifra indica il Block Group di appartenenza.

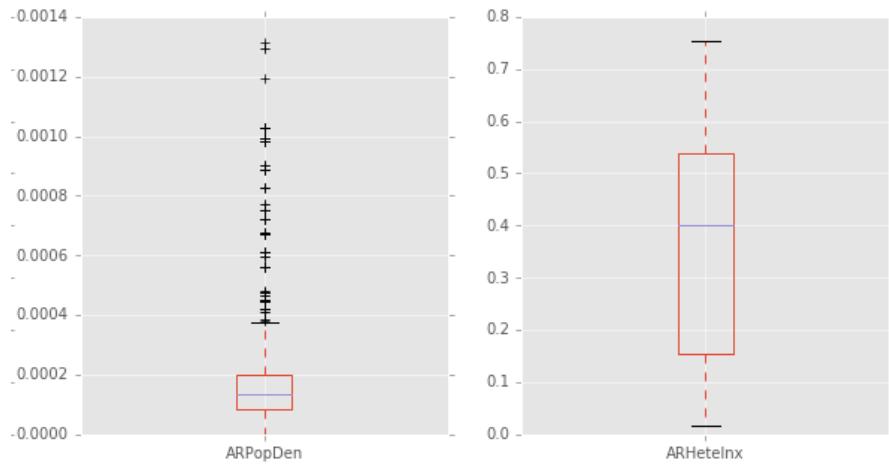
Ad esempio il Block 3019 si trova nel Block Group 3.

Questi sono tipicamente divisi da strade, in città corrispondono di solito agli isolati, mentre nelle zone rurali, dove ci sono meno strade, possono essere delimitati da altre caratteristiche.

La popolazione all'interno di un Block può variare sensibilmente.

Dimensione: 8354 x 12





Attributi	Descr	Type	Missing Values
BLOCKID10	ID di 15 cifre	Int	-
BLOCKCE	ultime 4 cifre di BLOCKID10	Int	-
STATEFP10	ID stato appartenenza	Int	-
COUNTYFP10	ID contea appartenenza	Int	-
TRACTCE10	ID tract appartenenza	Int	-
ARPopDen	Densità di popolazione	Real	2045 su 8354
ARPerAA	% di popolazione Afro Americana	Real	3272 su 8354
ARHetelnx	Indice di eterogeneità delle razze	Real	2962 su 8354
ARPerM1724	% di pop. maschile tra 17 e 24 anni	Real	3683 su 8354

ARPerHOwn	% di popolazione che possiede casa	Real	2523 su 8354
ARPerSF	% di famiglie con genitori single	Real	2895 su 8354

	ARPopDen	ARPerAA	ARHeteInx	ARPerM1724	ARPerHOwn	ARPerSF
count	1.33790 0e+04	11536.0	11443.00	10020. 0	12238.000	12062.00
mean	1.73208 3e-04	51.646927	0.368186	5.9234 97	57.136425	27.69665
std	1.63568 9e-04	34.419660	0.203971	4.0476 35	30.217798	16.83713
min	3.69900 0e-07	0.490196	0.015872	0.3717 47	0.719424	0.628931
25%	8.25085 0e-05	15.789474	0.154481	3.2258 06	35.714286	12.50000
50%	1.33489 6e-04	53.846154	0.402042	5.2631 58	54.545455	28.54251
75%	1.99881 1e-04	86.486486	0.538108	8.3333 33	84.375000	38.00000
max	1.31153 3e-03	100.00000 0	0.752447	42.177 914	100.000000	100.00

3.3 - Boundary

CLT_Boundary

Contiene le linee che formano i confini della città di Charlotte, NC.

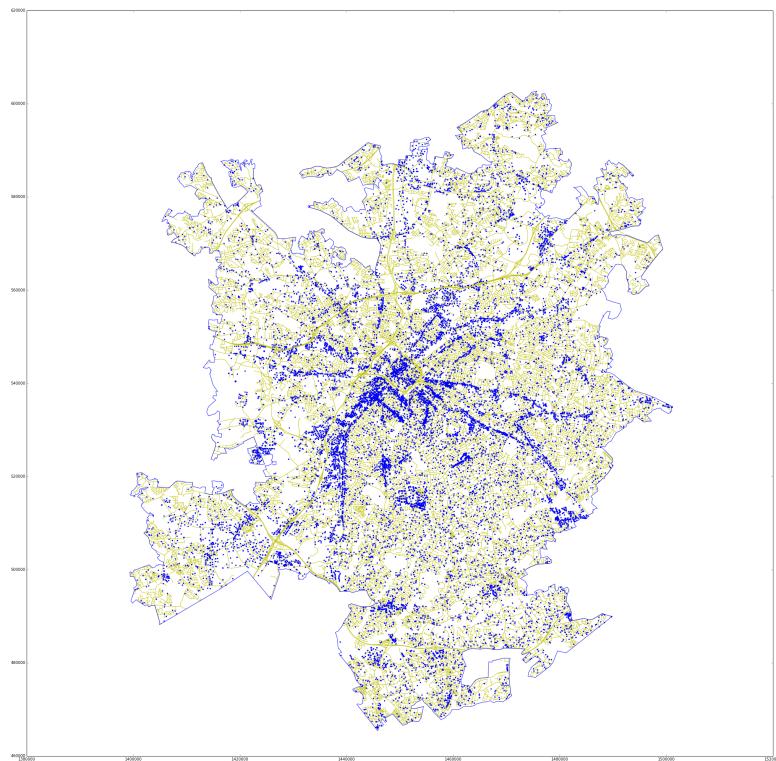
3.5 - Business

CLT_Business_Attri

Contiene informazioni riguardanti le attività commerciali, la loro posizione spaziale e attributi come il trade name, indirizzo, Block di appartenenza, descrizione. Vengono riportati solo gli attributi noti.

Dimensione: 24375 x 80

Attributi	Descr	Type
TradeName	Nome dell'attività	String
X_Coord	Coordinata x	Real
Y_Coord	Coordinata y	Real
NUM_ADDR	Numero civico	Int
year_	Anno (?)	Int
CntCode	Codice contea	Int
legalCode	Nome dell'attività (?)	String
tradeNm	Nome dell'attività	String
PhyStr1	Indirizzo	String
taxCity	Città	String
taxState	Stato	String
taxZip	CAP	Int
Block_id	Block di appartenenza	Int
WHOLESTNAM	Indirizzo	String



3.6 - Streets

CLT_Streets

Contiene informazioni spaziali delle strade della città di Charlotte, NC.

3.7 - Crime Count

CLT_Streets_crmCntAUNCC_Business

Statistica di quanti crimini sono stati segnalati in ogni strada.

3.8 - Crimes

CrimeIncident_CMPD_2010

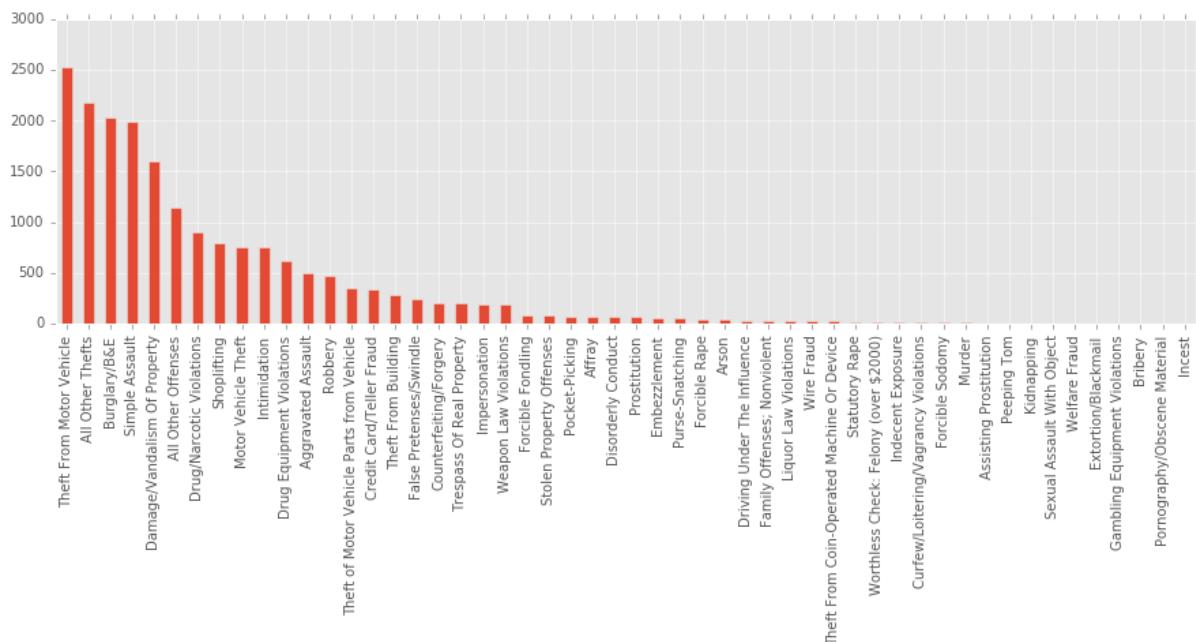
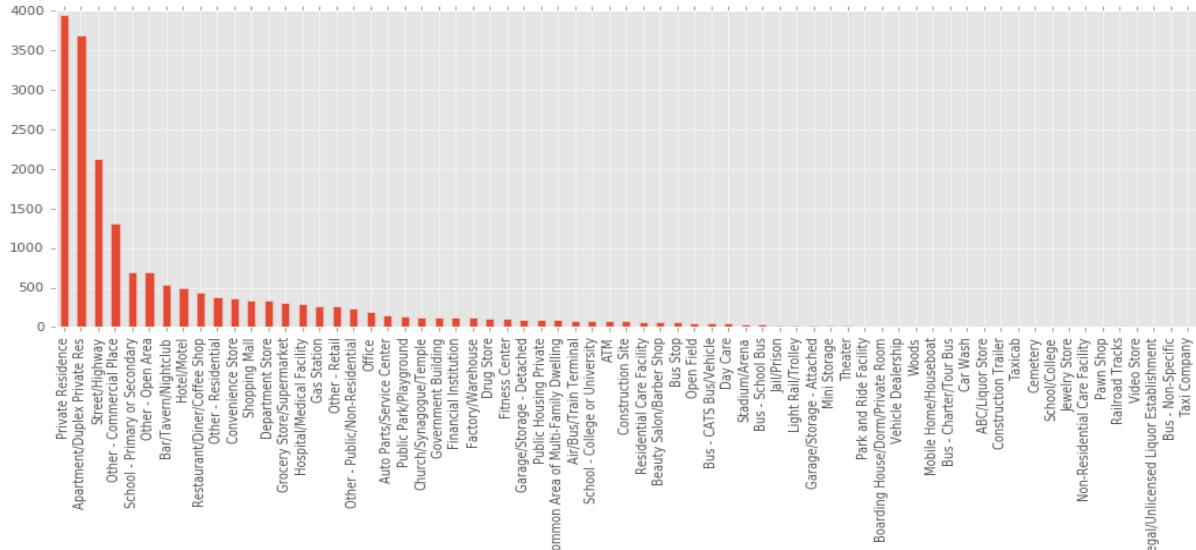
Contiene informazioni spaziali sui crimini avvenuti nel 2010 nella città di Charlotte, NC. Gli attributi rilevanti sono descritti di seguito:

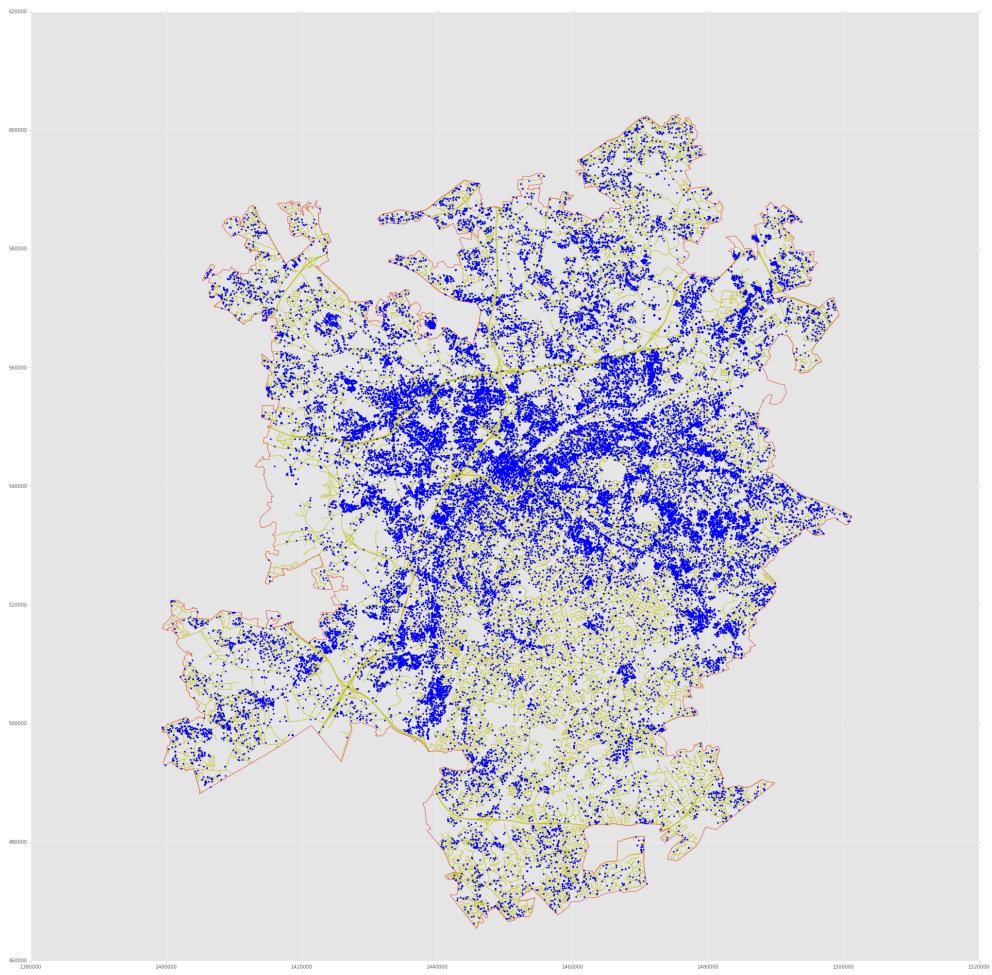
Dimensione: 67594 x 53

Attributi	Descr	Type
Block_No	Block dove è avvenuto il crimine	Int
NIBRSclass	Tipologia di crimine. 52 valori possibili	Categorical
Attempt	Esito: <ul style="list-style-type: none">• Attempted• Completed	Categorical
Case_Status	Esito <ul style="list-style-type: none">• Close/Cleared• Closed/Leads Exhausted• Further Investigation• Inactive	Categorical
Clearance_	Esito della chiamata: <ul style="list-style-type: none">• Exceptionally Cleared - By Death of Offender• Exceptionally Cleared - Cleared by Other Means• Exceptionally Cleared - Located	Categorical

	<p>(Missing Persons and Runaways only)</p> <ul style="list-style-type: none"> ● Exceptionally Cleared - Prosecution Declined by DA ● Exceptionally Cleared - Victim Chose not to Prosecute ● Normal Clearance - Cleared by Arrest ● Normal Clearance - Cleared by Arrest by Another Agency ● Open ● Open - Cleared, Pending Arrest Validation ● Unfounded 	
Place1	<p>Luogo dove è avvenuto il crimine:</p> <ul style="list-style-type: none"> ● Commercial Place ● Open Area ● Public/Non-Residential ● Residential ● Retail 	Categorical
Place2	<p>Luogo dove è avvenuto il crimine (più dettagliato (67 possibili valori).</p>	Categorical
Location_T	<p>Luogo dove è avvenuto il crimine</p> <ul style="list-style-type: none"> ● Indoors ● Other ● Outdoors ● Parking Deck ● Parking Lot 	Categorical
Report_Dat	<p>È composto da data più un numero progressivo che indica quante</p>	Date

	segnalazioni sono già avvenute esempio: 01/01/2010 0005	
Street_Nam	Strada dove è avvenuto il crimine	String





3.9 - Point of Interest POI

POI_AlcoholicDrinkingPlaces_CLIP

Contiene informazioni spaziali sulle attività di vendita di alcolici.

POI_HotelsMotels

Contiene informazioni spaziali sulle attività di vendita di alcolici.

POI_Malls

Contiene informazioni spaziali sui centri commerciali.

POI_ParknRideFacilities

Contiene informazioni spaziali riguardo strutture per il parcheggio auto ed eventuale fermata mezzi pubblici.

POI_WALMART

Contiene informazioni spaziali riguardo i punti vendita Walmart.



4 - DATA PREPARATION

In seguito all'esplorazione dei dati è emerso il bisogno di unire i vari file in una unica tabella in modo da poter applicare algoritmi di scoperta di regole associative classici. L'analisi dei crimini verterà sulla omonima tabella e si aggiungeranno attributi e questi prelevati dalle altre. In particolare gli attributi interessanti sono quelli relativi al censimento dei Blocks e dei Block Groups, è necessario quindi aggiungere le colonne alla tabella dei crimini in base al Block dove sono avvenuti.

4.1 - Proposizionalizzazione

Analizzando l'attributo **Block_No** della tabella crimini è emerso che questo non era riconducibile agli attributi **BLOCKID10** e **GEOID10** delle tabelle, rispettivamente, di Blocks e Block Groups. Un'alternativa è quella di scoprire il Block del crimine attraverso la via usando la tabella **Business**, infatti questa possiede sia la via che il Block di appartenenza.

Non è stato possibile però aggiungere il **BLOCKID10** ad ogni crimine, inoltre è stato riscontrato che non ogni **BLOCKID10** aggiunto era presente nella tabella dei **Block**.

La tabella **Crimes** inizialmente di 67595 righe è stata quindi ridotta a 50455 righe, ovvero le uniche per le quali è stato trovato il Block corrispondente. In seguito alla aggiunta degli attributi relativi al censimento delle tabelle **Block** e **Block Groups**, ed alla eliminazione delle righe non provviste di tali dati, le righe della tabella sono così diventate 19106.

Gli attributi aggiunti sono: **ARPCIncome**, **ARPerHEdu**, **ARPerWork**, **ARPerM15L**, **ARPerRM15L**, **ARPer3MU** della tabella **Block Group** e **ARPopDen**, **ARPerAA**, **ARHeteInx**, **ARPerM1724**, **ARPerHOwn**, **ARPerSF**, della tabella **Block**.

4.2 – Vicinato

Sono stati aggiunti i 5 più vicini poi/business per ogni crimine in modo da poter includere anche questo aspetto spaziale all'interno della scoperta di regole di associazione.

4.2.1 – Kdtree

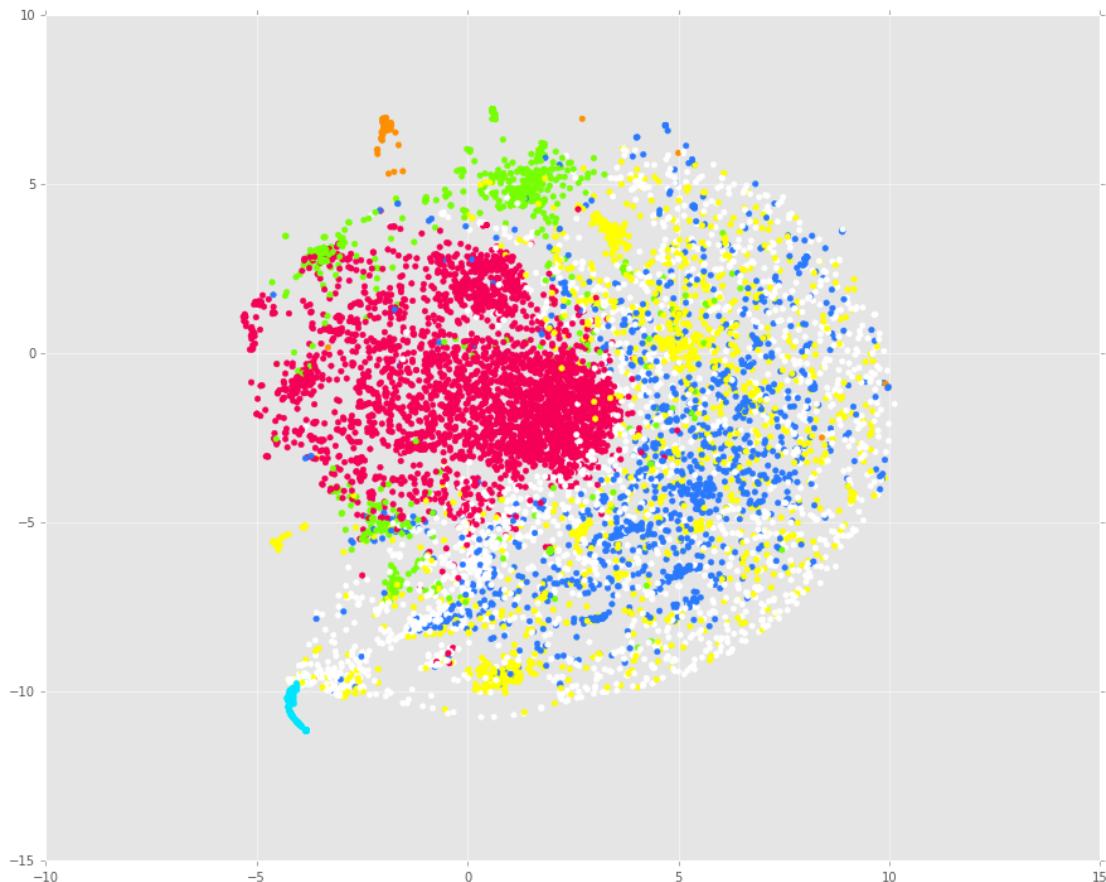
È stato utilizzato un k-d Tree come struttura dati per cercare rapidamente il vicinato di ogni punto. Questo è un albero binario dove ogni nodo è un punti nello spazio K-dimensionale. Ogni nodo non foglia può essere visto come un iperpiano che divide lo spazio in due parti. I punti a sinistra di questo iperpiano sono presenti nel sotto-albero sinistro mentre i punti a destra nel sotto-albero a destra. La direzione dell'iperpiano è scelta nel modo seguente: ad ogni nodo nell'albero viene associata una delle k dimensioni, con l'iperpiano perpendicolare all'asse di quella dimensione.

4.2.2 - Sentence Embedding

Però se per i POI provenienti dalle tabelle **Malls**, **ParkNRideFacilities**, **Walmart**, **AlcoholicDrinkingPlaces** e **HotelsMotels** è possibile usare il nome della tabella come generalizzazione dello specifico punto, nelle istanze della tabella **Business** sono disponibili solo i nomi commerciali (trade name) delle attività, non è presente alcun attributo che specifichi il tipo di attività come ad esempio alimentari, ristorante, abbigliamento ecc. Questo sarebbe stato utile al fine di trovare regole più generali e non specifiche alla singola attività commerciale.

È stato pensato di usare tecniche di NLP per raggruppare i negozi della tabella **Business** in cluster ed usare questi come classe di appartenenza. In particolare è stato utilizzato un algoritmo di sentence embedding per ricavare un vettore dal trade name dei negozi. L'algoritmo utilizzato, **Doc2Vec**, modifica Word2Vec per l'apprendimento non supervisionato per rappresentazioni continue di blocchi di testo, come frasi, paragrafi o interi documenti.

I risultati comunque risultano ancora non utilizzabili in quanto la valutazione della bontà dei cluster e l'interpretazione di questi necessita di ulteriori sviluppi.



4.3 - Selezione colonne

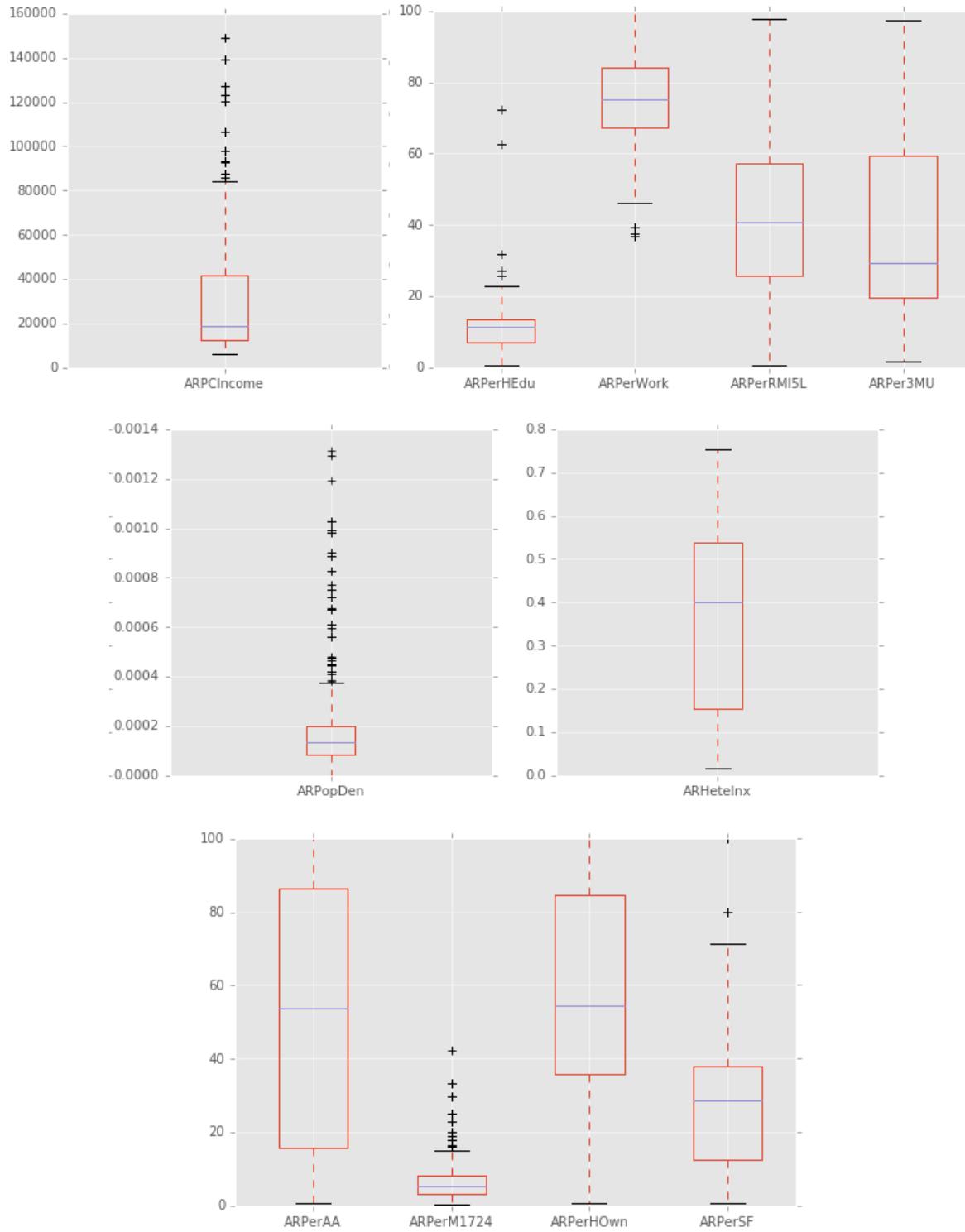
Sono stati rimossi tutti gli attributi ritenuti non rilevanti ai fini dell'analisi dalla tabella **Crimes**, risultata delle fasi precedenti. Gli attributi rimasti sono:

BLOCKID10	ID di 15 cifre del Block
First_POI	Primo POI più vicino
Second_POI	Secondo POI più vicino
Third_POI	Terzo POI più vicino
Fourth_POI	Quarto POI più vicino

Fifth_POI	Quinto POI più vicino
ARPCIncome	Reddito
ARPerHEdu	Percentuale di istruzione
ARPerWork	Percentuale di occupazione
ARPerRMI5L	Percentuale di pop che ha affittato e traslocato in 5 anni
ARPer3MU	Percentuale di case con 3 vani o più
ARPopDen	Densità di popolazione
ARPerAA	Percentuale di popolazione Afro Americana
ARHeteInx	Indice di eterogeneità delle razze
ARPerM1724	Percentuale di popolazione maschile tra 17 e 24 anni
ARPerHOwn	Percentuale di popolazione che possiede casa
ARPerSF	Percentuale di famiglie con genitori single
NIBRSclass	Tipo di crimine
Attempt	Esito
Case_Status	Stato
Clearance_	Esito
Place1	Luogo
Place2	Luogo più dettagliato
Location_T	Luogo
Report_Dat	Data con numero di chiamata
Street_Nam	Indirizzo

4.4 - Discretizzazione

Gli attributi relativi al censimento dei **Block** e dei **Block Groups** sono valori reali, e risulta quindi necessaria una fase di discretizzazione per poterli utilizzare in fase di modellazione, in quanto algoritmi di scoperta di regole di associazione accettano come input solo valori categorici.



La fase di discretizzazione scelta è la equal-depth in base ai quantili. Come primo tentativo è stato pensato di discretizzare in più intervalli gli attributi con più varianza, in seguito, per aumentare le regole in fase di modellazione e per migliorare l'interpretabilità, è stata preferita una divisione ad intervalli uguali per tutti gli attributi. Di seguito sono riportati come i valori numerici provenienti dagli attributi sono stati discretizzati. La **q** rappresenta il quantile di appartenenza, mentre la **t** il totale degli intervalli.

Attributo	Valori assunti
ARPCIncome	inc-q/t
ARPerHEdu	edu-q/t
ARPerWork	empl-q/t
ARPerRMI5L	lt5y-q/t
ARPer3MU	3mu-q/t
ARPopDen	popden-q/t
ARPerAA	afro-q/t
ARHeteInx	hetero-q/t
ARPerM1724	youngm-q/t
ARPerHOwn	own-q/t
ARPerSF	sinpar-q/t

4.5 - Missing Values

I missing values sono presenti solo nelle tabelle **Block** e **Block Groups** negli attributi relativi al censimento. In particolare, sono molto presenti nella tabella **Block** dove arrivano a circa $\frac{1}{3}$ del totale.

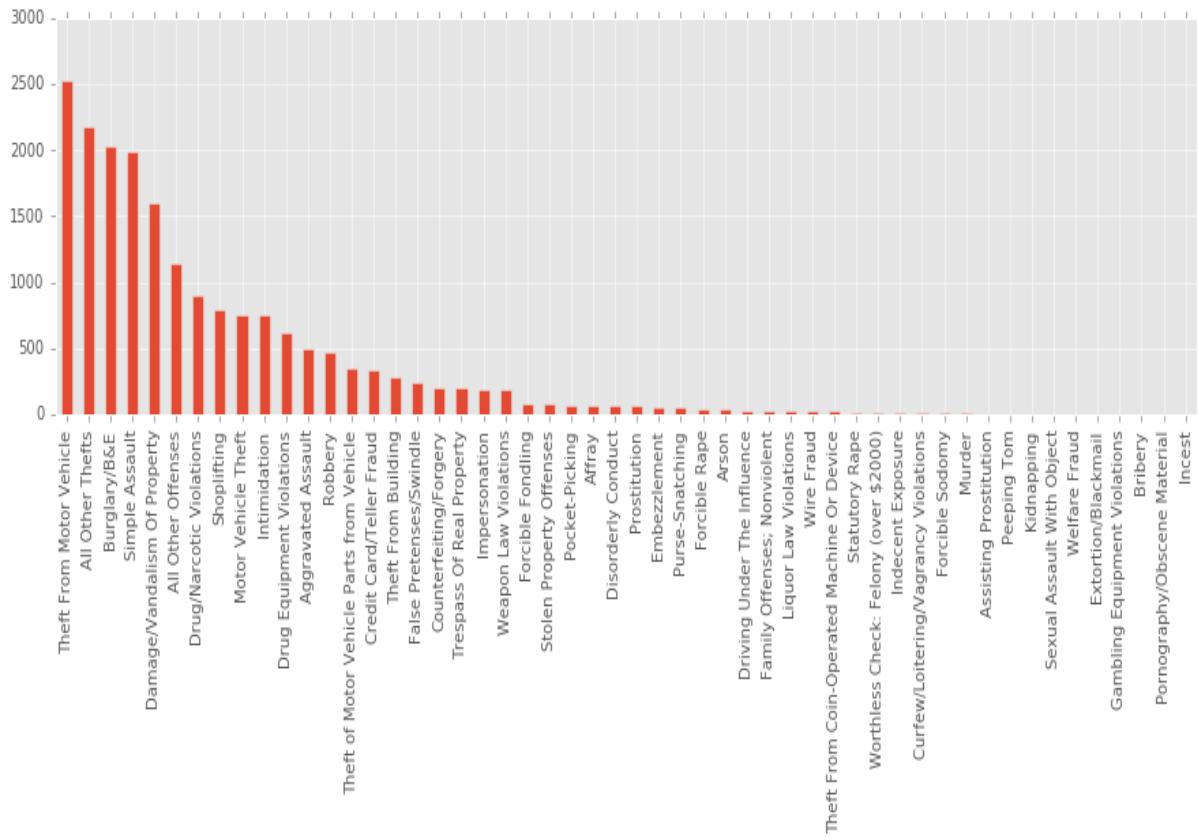
Per gli attributi **ARPCIncome**, **ARPerHEdu**, **ARPerWork**, **ARPerRMI5L** è stato scelto di eliminare le righe contenenti i missing values, in quanto risultavano solo poche istanze.

Per l'attributo **ARPer3MU** invece è stato preferito la rimozione dell'intera colonna, scelta motivata dalla presenza di numerosi missing values e dalla difficile interpretazione dell'attributo nelle eventuali regole scoperte.

Infine per gli attributi **ARPopDen**, **ARPerAA**, **ARHeteInx**, **ARPerM1724**, **ARPerHOwn**, **ARPerSF** è stato deciso di predirli sulla base degli altri attributi attraverso una versione ottimizzata dell'algoritmo CART e la Gini impurity per misurare la qualità degli split.

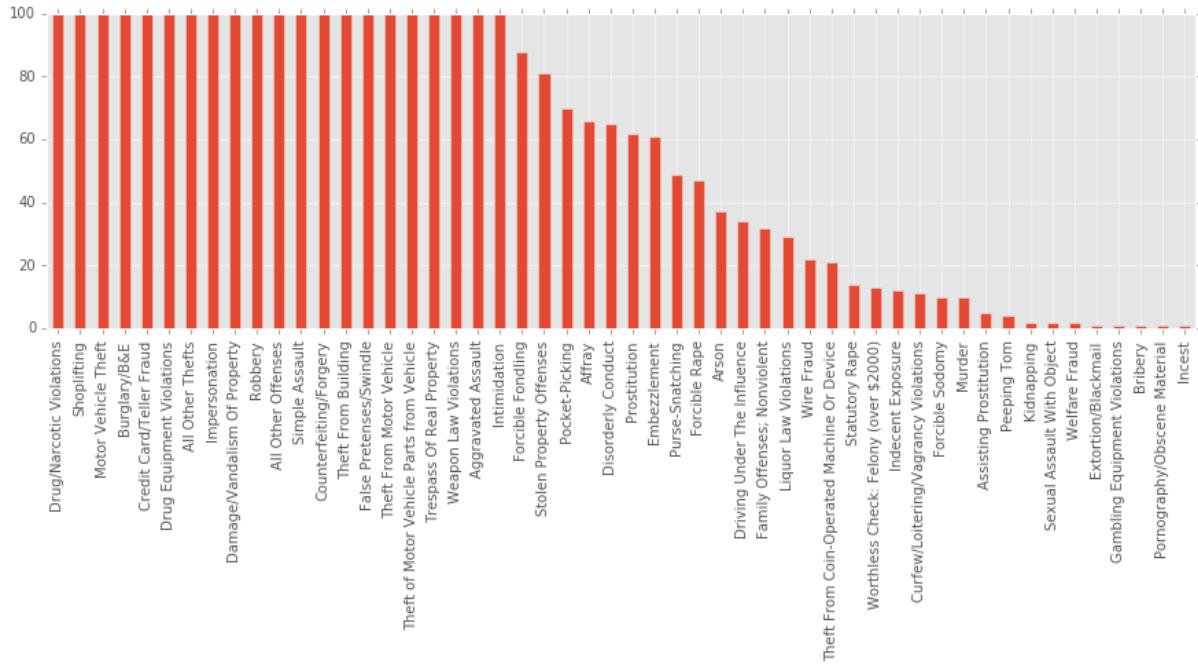
sampling

Per un migliore bilanciamento del dataset è stato effettuato un campionamento stratificato.



La distribuzione dei tipi di crimine all'interno del dataset non è omogenea, quindi è stato preso un campione casuale di 100 crimini per ogni tipo di crimine. Dove il tipo di crimine aveva meno di 100 elementi, sono stati presi tutti.

Il numero di righe rimaste è così sceso a 2954.



5 - MODELING

5.1 - Regole di associazione

Le regole di associazione sono uno dei metodi per estrarre relazioni nascoste tra i dati. Inizialmente introdotte per la scoperta di regolarità all'interno delle transazioni registrate nelle vendite dei supermercati (market basket analysis), tale informazione veniva utilizzata come base per le decisioni riguardanti le attività di marketing, come ad esempio le offerte promozionali o il posizionamento dei prodotti negli scaffali.

La scoperta di regole si divide in due fasi: l'estrazione degli itemset frequenti e la generazione di regole. Diversamente da algoritmi di Sequence Mining, questi non considerano l'ordine degli items all'interno di una transazione. Algoritmi di Association Rule Mining possono processare solo valori categorici, ed inoltre assumono l'esistenza di una unica tabella relazionale (single table assumption) dove le tuple sono transazioni. Le regole di associazione possono essere interpretate differentemente in base al contesto. In generale una regola del tipo $X \Rightarrow Y$ indica che se si verifica X allora con una certa confidenza si verifica anche Y . La relazione di implicazione \Rightarrow non è simmetrica, quindi non vale il contrario.

Una regola è definita come un'implicazione nella forma $X \Rightarrow Y$ dove $X, Y \subseteq I$ e $X \cap Y = \emptyset$. L'insieme di oggetti (o itemsets) X e Y vengono chiamati rispettivamente antecedente e conseguente della regola.

Il **supporto** di un itemset X è la quantità di volte in cui esso appare nel dataset ovvero $p(X)$. Se il supporto è maggiore di una soglia minima **minsupp** questo è detto itemset frequente. Il **supporto** di una regola $X \Rightarrow Y$ è $p(X \cup Y)$.

La **confidenza** di una regola $X \Rightarrow Y$ è la quantità di volte in cui appare Y in tutte le transazioni dove X è presente, ovvero $p(Y|X)$.

5.2 - Relim

Relim (Recursive Elimination) è un algoritmo per la scoperta di itemset frequenti, fortemente ispirato da algoritmi come FP-growth e H-mine. Lavora senza costruire un prefix-tree o altre strutture dati, processando le transazioni direttamente. Il suo punto di forza non è la velocità ma la semplicità di implementazione.

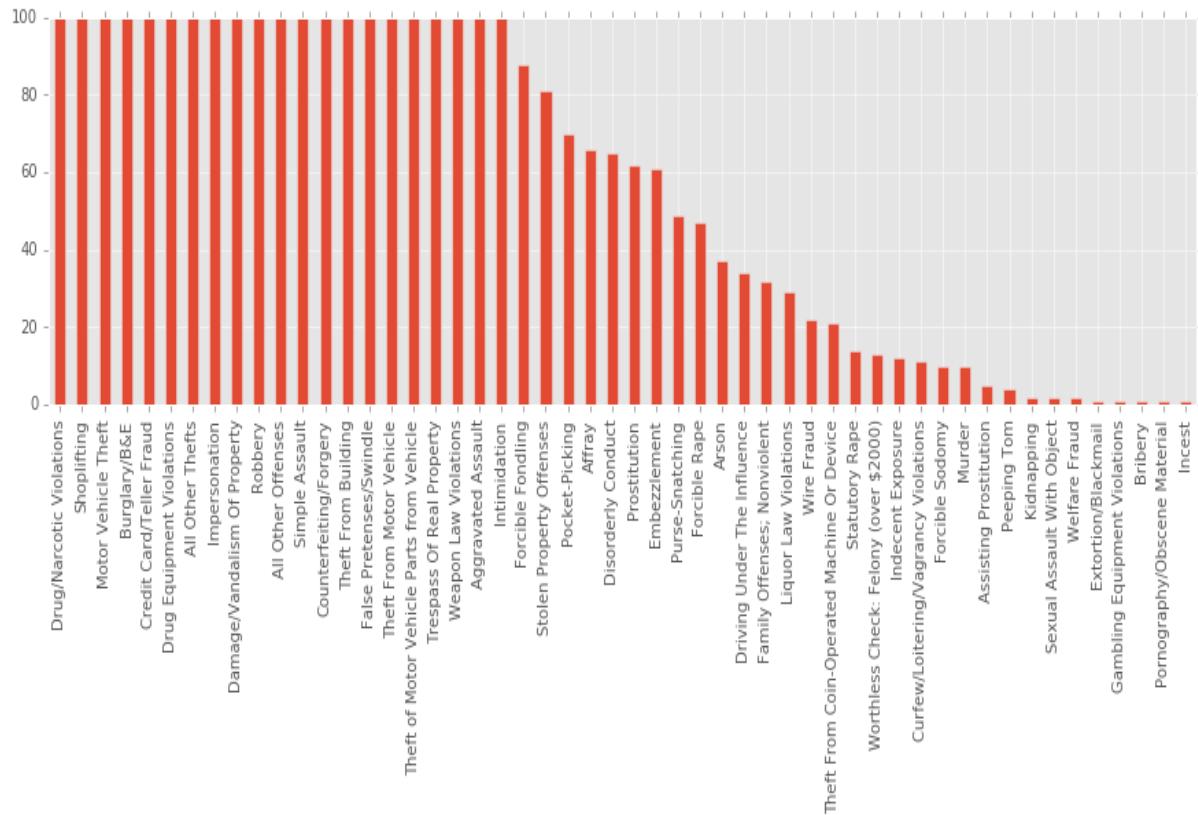
5.3 - Configurazioni

Gli attributi rimossi in questa fase sono **BLOCKID10 - Street_Nam - ARPerRMI5L - ARPer3MU - ARHeteInx - ARPerHOwn - Attempt - Case_Status - Place1 - Clearance_ - Location_T**, e gli attributi relativi ai **POI**.

Dimensione: 2954 x 10

5.3.1 - Sampled

Con il dataset campionato il **minsupp** dovrà essere necessariamente minore di 100.



Senza predire i valori mancanti rimangono solo **ARPCIncome, ARPerHEdu, ARPerWork, NIBRSclass, Place2, Report_Dat**

Dimensione: 2954 x 6

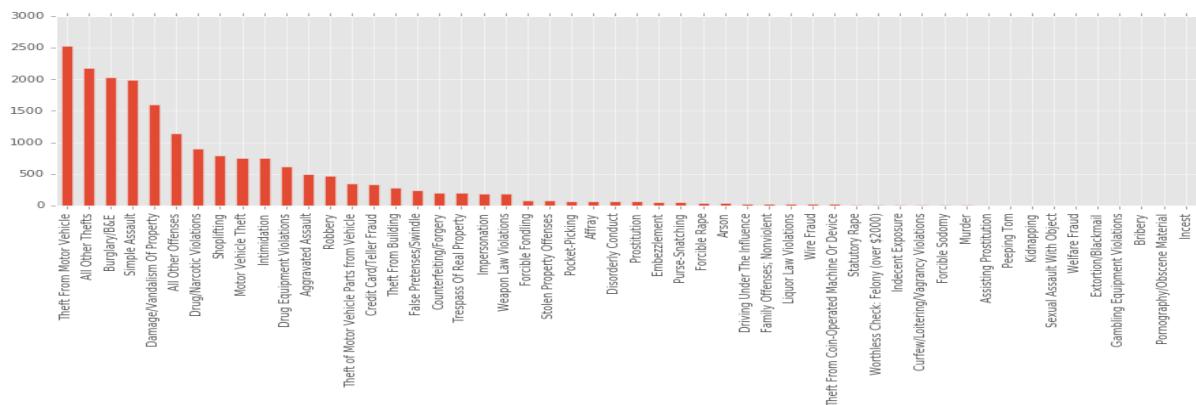
# bins	mv	min supp	min conf	freq itemset	time item	rules	time rule	rules crimes	rules crimeCon
3	si	50	0.8	2639	418 ms	3490	93.2 ms	5	0
3	si	10	0.8	28607	897 ms	81271	2.07 s	32897	459
3	si	50	0.6	2639	340 ms	7154	147 ms	11	0
3	si	10	0.6	28607	963 ms	130556	2.72 s	50183	2643
3	no	50	0.8	227	75.6 ms	20	1.43 ms	2	0
3	no	10	0.8	2570	134 ms	753	35.9 ms	375	12
3	no	50	0.6	227	76.1 ms	62	1.63 ms	4	0
3	no	10	0.6	2570	120 ms	1547	34.5 ms	717	37
7	si	50	0.8	1839	248 ms	7635	111 ms	0	0
7	si	10	0.8	20967	684 ms	131609	2.33 s	39212	0
7	si	50	0.6	1839	270 ms	10563	123 ms	0	0
7	si	10	0.6	20967	727 ms	172968	2.55 s	53140	3669
7	no	50	0.8	237	73.7 ms	27	1.23 ms	0	0
7	no	10	0.8	2248	132 ms	612	25.2 ms	276	0
7	no	50	0.6	237	70.8 ms	42	1.45 ms	0	0
7	no	10	0.6	2248	128 ms	1097	24.3 ms	489	14

5.3.2 - Entire

Dimensione: 19106 x 10

Dimensione senza predire i missing values: 19106 x 6

# bins	mv	min supp	min conf	freq itemset	time item	rules	time rule	rules crimes	rules crimeCon
3	si	200	0.8	5822	4.77 µs	7105	238 ms	673	0
3	si	70	0.8	24349	4.05 µs	41256	1.4 s	13175	859
3	si	200	0.6	5822	3.81 µs	13321	342 ms	1462	1
3	si	70	0.6	24349	3.81 µs	71907	1.87 s	23496	1468
3	no	200	0.8	574	4.05 µs	78	7.54 ms	32	0
3	no	70	0.8	2154	3.81 µs	405	26 ms	165	4
3	no	200	0.6	574	5.01 µs	185	15.8 ms	81	1
3	no	70	0.6	2154	3.81 µs	1003	92.1 ms	415	16
7	si	200	0.8	3438	3.81 µs	11235	203 ms	282	0
7	si	70	0.8	17086	4.05 µs	73441	1.54 s	16728	1379
7	si	200	0.6	3438	4.05 µs	16355	255 ms	459	1
7	si	70	0.6	17086	3.81 µs	101352	1.75 s	23332	1461
7	no	200	0.8	455	4.05 µs	47	2.8 ms	5	0
7	no	70	0.8	2008	3.81 µs	311	21.3 ms	99	8
7	no	200	0.6	455	4.05 µs	86	2.91 ms	8	1
7	no	70	0.6	2008	3.81 µs	540	21.5 ms	170	12



6 - EVALUATION

Si procede con una valutazione qualitativa e riscontro con i requisiti di business. Si è scelto la configurazione che ha restituito un numero di rilevante di regole con crimini nel conseguente, e si sono valutate qualitativamente le più interessanti, considerando il supporto e la confidenza, raggruppandole per tipo di crimine.

Maggiore enfasi sarà posta alle regole con solo il tipo di crimine nel conseguente.

6.1 - Sampled Dataset

CREDIT CARD/TELLER FRAUD

out of 100

Ant	Cons	Supp	Conf
ATM, afro-3/3, edu-3/3, empl-1/3, \Rightarrow youngm-3/3	Credit Card/Teller Fraud	10	0.91

Ant	Cons	Supp	Conf
ATM, afro-3/3, empl-1/3, youngm-3/3 \Rightarrow	Credit Card/Teller Fraud	10	0.91

Ant	Cons	Supp	Conf
ATM', 'afro-3/3', 'youngm-3/3 \Rightarrow	Credit Card/Teller Fraud	10	0.91

Ant	Cons	Supp	Conf
ATM, edu-3/3, youngm-3/3 \Rightarrow	Credit Card/Teller Fraud	10	0.91

Ant	Cons	Supp	Conf
ATM, popden-1/3, youngm-3/3 \Rightarrow	Credit Card/Teller Fraud	10	0.91

Le regole evidenziano come il 10% le frodi di carte di credito si verifichino presso sportelli per il prelievo automatico di denaro. Inoltre questi si verificano in zone con bassa densità di popolazione, con elevata popolazione maschile tra 17 e 24 anni, elevata popolazione afro americana, alta istruzione e bassa occupazione.

SHOPLIFTING

out of 100

Ant	Cons	Supp	Conf
Shoplifting	⇒ sinpar-1/3	61	0.61

Ant	Cons	Supp	Conf
Shoplifting, Shopping Mall	⇒ inc-7/7	11	0.85

Ant	Cons	Supp	Conf
Shoplifting, Shopping Mall, edu-4/7, empl-2/7	⇒ inc-7/7	11	1.0

Ant	Cons	Supp	Conf
Shoplifting, Shopping Mall, empl-2/7	⇒ inc-7/7	11	0.92

Il taccheggio sembra preferire i centri commerciali, con alto reddito e con concentrazione di famiglie con genitori single.

DISORDERLY CONDUCT

out of 65

Ant	Cons	Supp	Conf
'Disorderly Conduct	⇒ sinpar-3/3	50	0.76

Ant	Cons	Supp	Conf
School - Primary or Secondary, edu- 6/7, empl-1/7, inc-1/7	⇒ Disorderly Conduct	23	0.67

Ant	Cons	Supp	Conf
youngm-2/3, School - Primary or Secondary, inc-1/3, edu-3/3, afro-3/3	⇒ Disorderly Conduct	23	0.72

Ant	Cons	Supp	Conf
sinpar-3/3, School - Primary or Secondary, afro-3/3, youngm-2/3, empl-1/3, inc-1/3, popden-1/3	⇒ Disorderly Conduct	23	0.72

Ant	Cons	Supp	Conf
youngm-2/3, School - Primary or Secondary, afro-3/3, popden-1/3	⇒ Disorderly Conduct	23	0.72

Rientrano i casi di disturbo della quiete pubblica, ubriachezza, o cattiva condotta. Questa è solita verificarsi nelle scuole, in quartieri con basso reddito, molte famiglie con genitori single. Inoltre la bassa occupazione, l'elevata concentrazione di popolazione giovane contribuiscono al crimine.

WEAPON LAW VIOLATIONS

out of 65

Ant	Cons	Supp	Conf
School - Primary or Secondary, empl- 6/7	⇒ Weapon Law Violations	12	0.71

Circa il 20% delle violazioni sulle armi avvengono nelle scuole.

PROSTITUTION

out of 62

Ant	Cons	Supp	Conf
Prostitution	sinpar-3/3	50	0.81

Ant	Cons	Supp	Conf
afro-3/3, youngm-2/3, popden-2/3, empl-2/3, inc-2/3, sinpar-3/3, Gas Station	Prostitution	12	0.8

Ant	Cons	Supp	Conf
Gas Station, afro-3/3, edu-2/3, popden-2/3, sinpar-3/3, youngm-2/3	Prostitution	13	0.81

Ant	Cons	Supp	Conf
Gas Station, popden-2/3, sinpar-3/3, youngm-2/3	Prostitution	13	0.81

Ant	Cons	Supp	Conf
Gas Station, afro-3/3, edu-2/3, empl-2/3, popden-2/3	Prostitution	12	0.71

I crimini relativi alla prostituzione sono correlati alle stazioni di benzina e a quartieri con elevate popolazione afro americana, famiglie con genitori single ed elevata popolazione maschile fra i 17 e 24 anni.

FORCIBLE FONDLING

out of 88

Ant		Cons	Supp	Conf
afro-3/3, sinpar-2/3, Private Residence, youngm-3/3, inc-1/3, popden-1/3	⇒	Forcible Fondling	13	0.65

Ant		Cons	Supp	Conf
Private Residence, sinpar-2/3, afro-3/3, popden-1/3, empl-1/3	⇒	Forcible Fondling	12	0.71

Riguarda i crimini a sfondo sessuale, le regole evidenziano come il 15% di questi avvengono in residenze private e in quartieri a basso reddito, bassa occupazione, con molti genitori single, bassa densità di popolazione.

COUNTERFEITING/FORGERY

out of 100

Ant		Cons	Supp	Conf
Financial Institution, popden-2/3	⇒	Counterfeiting/Forgery	17	0.81

Ant		Cons	Supp	Conf
Financial Institution	⇒	Counterfeiting/Forgery	30	0.71

Ant		Cons	Supp	Conf
inc-2/3, Financial Institution, sinpar-3/3	⇒	Counterfeiting/Forgery	10	0.91

Ant	Cons	Supp	Conf
inc-2/3, Financial Institution	⇒ Counterfeiting/Forgery	10	0.91

I reati di falsificazione/contraffazione avvengono nelle istituzioni finanziarie e nelle zone mediamente popolate.

6.2 - Entire Dataset

SHOPLIFTING

out of 787

Ant	Cons	Supp	Conf
Department Store, inc-3/3	⇒ Shoplifting	180	0.73

Ant	Cons	Supp	Conf
Department Store, empl-3/3	⇒ Shoplifting	171	0.77

Ant	Cons	Supp	Conf
Department Store, afro-1/3	⇒ Shoplifting	186	0.71

Ant	Cons	Supp	Conf
Department Store, afro-1/3, inc-3/3	⇒ Shoplifting	179	0.73

Ant	Cons	Supp	Conf
afro-1/7, Department Store, sinpar-2/7, empl-6/7	⇒ Shoplifting	78	0.84

Anche nel dataset intero il taccheggio si concentra in grandi magazzini, in quartieri ad alto reddito, con bassa percentuale di popolazione afro americana e genitori single e ad alta occupazione.

THEFT FROM MOTOR VEHICLE

out of 2523

Ant	Cons	Supp	Conf
Theft From Motor Vehicle, edu-1/7, empl-7/7	⇒ inc-7/7	305	0.93

Ant	Cons	Supp	Conf
Theft From Motor Vehicle, empl-7/7, inc-7/7	⇒ edu-1/7	305	0.93

Ant	Cons	Supp	Conf
Theft From Motor Vehicle, edu-1/7, inc-7/7	⇒ empl-7/7	305	0.93

Il crimine di furto da motoveicolo risulta solo nelle regole scoperte il dataset intero, in quanto in questo risulta essere il crimine più frequente. Dove si verificano crimini di questo tipo risultano essere quartieri con bassa istruzione, alto reddito e alta occupazione.

ALL OTHER THEFTS

out of 2173

Ant	Cons	Supp	Conf
All Other Thefts, inc-1/7	⇒ empl-1/7	280	0.82

I furti si verificano in zone a basso reddito e bassa occupazione.

SIMPLE ASSAULT

out of 1996

Ant	Cons	Supp	Conf
Simple Assault, inc-1/7	⇒ empl-1/7	306	0.81

Dove avvengono i crimini di aggressione, se il reddito è basso, risulta esserci bassa occupazione.

6.3 - Conclusioni

Il dataset analizzato presentava pochi attributi e molte istanze, di conseguenza le regole scoperte risultavano con valori di supporto bassi. Questo è dovuto anche al fatto che il processo mirava a scoprire relazioni fra i dati e le tipologie di crimine, i quali risultavano numerosi e distribuiti non uniformemente. Tuttavia alcune regole sono risultate interessanti per i requisiti di business. È necessario notare come i filtri applicati alle regole, ovvero selezionare solo le regole con tipologia di crimine nel conseguente, diminuivano di molto l'output dell'algoritmo. Queste però erano quelle maggiormente rilevanti in quanto mostravano dove, al verificarsi delle condizioni presenti nell'antecedente della regola, le varie tipologie di crimine sono più probabili a manifestarsi. Questo era l'obiettivo del processo di KDD, che ha iterato più volte le varie fasi cercando di estrarre più conoscenza possibile dai dati a disposizione.