POLITECNICO DI TORINO

Corso di Laurea Magistrale
in Ingegneria Gestionale

Tesi di Laurea Magistrale

Caratterizzazione, esplorazione ed analisi di dati relativi alla sicurezza urbana mediante tecniche di data mining

Relatore
prof. Silvia Anna Chiusano
Co-Relatore
prof. Tania Cerquitelli

Candidato
Federica Serra

Aprile 2018

Sommario

1. Introduzione 1

2. Esplorazione degli open data sulla sicurezza urbana 3

2.1 Open data della città di Torino 3

2.2 Open data in Italia 7

2.3 Open data in Europa e nel Mondo 9

2.3.1 Contesto Europeo 10

2.3.2 Contesto extra-europeo 12

3. Introduzione alle tecniche di data mining 14

3.1 Data Mining 14

3.2 Regole di associazione 15

3.3 Rapidminer 17

3.3.1 Il Processo e l'algoritmo FP-Growth 18

3.4 Il classificatore associativo 20

3.4.1 Descrizione del classificatore associativo 20

3.4.2 WEKA e L3 21

3.4.3 Costruzione del processo con il classificatore associativo 22

4. Pre-processing del dato 26

4.1 Arricchimento dati e formule excel 26

4.1.1 Arricchimento ulteriore con aggiunta festività 33

4.2 Tassonomia 34

4.3 Divisione del dataset 35

4.4 Costruzione del processo con Rapidminer 35

5. Analisi delle regole di associazione 44

5.1 Tipologia di estrazione delle regole di associazione generate 44

5.2 Analisi del dataset utilizzando la categoria come conclusione o premessa 46

5.3 Analisi del dataset utilizzando la sottocategoria come conclusione o premessa 55

5.4 Circoscrizioni e quartieri 61

5.5 Analisi dei dataset divisi per anno e per sottocategoria 62

6. Analisi dei risultati con il classificatore associativo 68

6.1 Analisi quantitativa delle prestazioni del classificatore associativo 68
6.1.1 Matrice di confusione, precisione e richiamo 88
6.2 Analisi qualitativa delle regole di classificazione estratte..... 90
7. Conclusioni 96
Bibliografia e sitografia 98

1. Introduzione

Al giorno d'oggi, uno tra gli argomenti principali discussi è sicuramente quello relativo agli open data. Il futuro del mondo digitale sarà principalmente costituito da "dati open". Precisamente il tema trattato è quello delle "smart cities". Vengono chiamate "città intelligenti", infatti, sono aree urbane in cui sono collezionati dati non sensibili e non personali (ad esempio tramite sensori elettronici oppure dai cittadini stessi). Questi dati sono collezionati per puro scopo di conoscenza, per fornire informazioni utili da usare in futuro o anche per ottimizzare l'efficienza di un servizio (ad esempio i trasporti urbani). Tramite questi dati, è possibile monitorare e tenere sotto controllo ciò che sta accadendo nella città e come la città si sta sviluppando.
Nell'ambito relativo alle città, una delle cose da non sottovalutare nella vita di tutti i giorni è la sicurezza del cittadino, e più nel dettaglio la percezione della sicurezza che ogni cittadino ha. Si sono cercati modi di usare la tecnologia per rendere la vita cittadina più sicura e più "smart". Infatti, una "città intelligente" può essere più preparata a rispondere a qualsiasi evento possa succedere; inoltre, migliora la qualità della vita e soprattutto quella percepita. Ad esempio l'individuazione di zone che risultano pericolose possono essere non frequentate spesso. Questo è un aspetto molto importante per un cittadino: sapere ciò che sta accadendo in città potrebbe fargli evitare certe zone in certi periodi, sapere che un luogo è considerato costantemente pericoloso negli anni potrebbe evitargli ad esempio di comprare o affittare casa in quel luogo.
I dati open potrebbero essere utili ai cittadini stessi che li utilizzano per prendere decisioni migliori nella vita di tutti i giorni. Questi dati hanno il vantaggio di poter essere accessi, essere utilizzati ed essere ulteriormente condivisi da tutti. Possono essere quindi sia esplorati liberamente sia analizzati come si voglia. Infatti, uno dei loro scopi principali, è quello di predire gli andamenti futuri guardando al passato.
L'utilizzo e l'analisi di dati open è fatto solamente a scopo di beneficio, anche se bisogna stare molto attenti alla condivisione di dati che possono essere sensibili poiché, al giorno d'oggi, si potrebbe essere soggetti ad attacchi informatici maligni oppure una semplice violazione della privacy nell'utilizzo di dati sensibili. Questi dati inoltre potrebbero anche aiutare a risolvere problemi che ancora non si è in grado di prevedere.
In questa tesi, sarà svolta una ricerca approfondita su dati riguardanti la sicurezza urbana in città diverse appartenenti a tutto il mondo. Questi dati sono stati esplorati e sono stati confrontati tra loro; si è notato cosa avessero in comune, cosa li differiva, cosa li rende particolari. Una delle caratteristiche che li accomuna è il volume dei dati, cosa negativa che può essere tramutata in positiva. Negativa perché non è possibile analizzare direttamente il set di dati a disposizione, si potrebbe esplorarlo e caratterizzarlo ma non si può arrivare ad una conclusione non sbagliata, salvo nel più fortunato dei casi. Il problema che i dati siano "Big Data", si è risolto con l'integrazione delle tecniche di data mining e attraverso strategie di Business Intelligence. Quest'ultima si preoccupa principalmente di trasformare i dati e le informazioni in conoscenza, attraverso l'uso di

software che, partendo da grandi masse di dati, tramite l'utilizzo di un algoritmo, forniscono indici facilmente comprensibili da una mente umana. Infatti sono chiamati anche "sistemi per il supporto alle decisioni".
Nello specifico si sono esplorati gran parte dei dati provenienti da tutto il mondo. Come obiettivo di questa tesi, verranno caratterizzati solamente i dati relativi alla città di Torino e ne verrà analizzato un campione, quello relativo alle segnalazioni inviate da cittadini al Contact Center della Polizia Municipale di Torino negli anni.
L'analisi di questi avverrà attraverso tecniche di data mining, cercando di trarre conclusioni focalizzandosi sulla correlazione tra i risultati, e non sull'algoritmo.
È interessante notare che i dati analizzati non sono quelli originari scaricati dal sito del comune di Torino, ma sono stati "puliti", convertiti, organizzati e arricchiti prima di essere processati.
In questa tesi, per l'analisi, si è deciso di usare due tecniche: una basata sulle regole di associazione e l'altra basata su un classificatore associativo. La prima è un metodo in grado di far scoprire, attraverso l'analisi di indici, relazioni interessanti tra due o più attributi, che inizialmente non sono risultate ovvie. La seconda ha come obiettivo l'associazione di ogni segnalazione ad una categoria specifica nel modo più accurato possibile. Infine si sono analizzati i risultati ottenuti. Nel primo caso, secondo diverse tipologie di regole di associazione ritenute opportune e interessanti sia statisticamente, attraverso l'analisi degli indici ottenuti, sia rilevante per il contenuto, rispondendo a domande del tipo "dato un luogo, cosa succede in quel luogo? Cosa è accaduto in un certo periodo di tempo in particolare? Come si sposta la pericolosità relativa ad un certo luogo nel tempo?". Nel secondo caso, oltre a dare un'interpretazione alle regole associative estratte dal punto di vista del contenuto informativo, è stata svolta un'analisi dei parametri dell'algoritmo di classificazione per trovare la configurazione più opportuna e scoprire qual è l'impatto sulle performance.
Infine, nell'ultimo capitolo, verranno trattate le conclusioni e quali potrebbero essere gli sviluppi futuri di questo progetto.

2. Esplorazione degli open data sulla sicurezza urbana

Questo capitolo riporta una caratterizzazione generale di overview sugli open data relativi alla sicurezza urbana e verranno descritti generalmente gli open data relativi alla sicurezza urbana utilizzati. È stata svolta una ricerca sugli open data disponibili in Italia, in Europa e successivamente in tutto il mondo, con particolare riferimento ai dati sulla sicurezza urbana. Per l'attività oggetto di questa tesi sono stati scelti come caso di studio di riferimento i dati relativi alla sicurezza urbana per la città di Torino, disponibili sul sito AperTO [1].

2.1 Open data della città di Torino

In questa sezione verranno presi in considerazione dataset disponibili sul sito AperTO, dove sono presenti gli open data della città di Torino, localizzata geograficamente come mostra la figura 2.1.

Figura 2.1 - Torino in Italia, tratta da http://www.imetravel.com/ita/italia.htm

I principali dataset resi disponibili di nostro interesse sono principalmente tre:

- Segnalazioni al Contact Center della Polizia Municipale [2]
- Violazioni ai Regolamenti Comunali [3]
- Violazioni al Codice della Strada [4]

Le segnalazioni al contact center della polizia municipale sono comunicazioni dei cittadini arrivate al contact center della Polizia Municipale di Torino, con l'indicazione del tipo di violazione segnalata, luogo dove è avvenuto il reato, data e ora di quando è avvenuta la segnalazione. Inoltre, queste segnalazioni sul sito di AperTO, sono complete da gennaio 2012 fino a luglio 2017, con aggiornamento semestrale.
Il dataset è caratterizzato da 7 attributi:

- *Categoria criminologa*: è la categoria che descrive il tipo di avvenimento o di reato. Le tre opzioni possibili per la scelta della macrocategoria sono: Allarme sociale, Convivenza Civile, Qualità Urbana.
- *Sottocategoria criminologa*: è la sottocategoria più specifica che meglio descrive la macrocategoria.
  Le sottocategorie associate ad "Allarme sociale" sono:
  - ✓ Atti di Vandalismo
  - ✓ Altro

  Le sottocategorie associate a "Convivenza Civile" sono:
  - ✓ Aggregazioni giovanili
  - ✓ Comportamenti molesti
  - ✓ Disturbi animali
  - ✓ Disturbi Cani
  - ✓ Disturbo da locali
  - ✓ Rumori molesti
  - ✓ Uso improprio di parti comuni
  - ✓ Altro

  Le sottocategorie associate a "Qualità Urbana" sono:
  - ✓ Decoro e degrado urbano
  - ✓ Veicoli abbandonati
  - ✓ Altro
- *Circoscrizione*: è un numero da 1 a 10 che rappresenta le 10 circoscrizioni di Torino.
- *Località*: è la via della città di Torino dove è avvenuto l'avvenimento o il reato (senza civico).
- *Area Verde*: è un campo booleano. Indica se il luogo dove è avvenuto l'avvenimento o il reato è o non è un'area verde.
- *Data*: è la data in cui è avvenuto l'avvenimento e anche la data della segnalazione alla Polizia Municipale in formato gg/mm/aaaa.
- *Ora*: è l'ora in cui è avvenuto l'avvenimento presupponendo che coincide con l'ora della segnalazione alla Polizia Municipale in formato hh.mm.

Il secondo dataset riguarda le violazioni ai regolamenti comunali e le segnalazioni sono complete da gennaio 2011 a giugno 2017, anche queste aggiornate semestralmente. Questo dataset è caratterizzato da 12 attributi:

- *Anno*: corrisponde all'anno in cui è avvenuta la violazione al regolamento comunale.
- *Mese*: corrisponde al mese in cui è avvenuta la violazione e anche la segnalazione ed è rappresentato da un campo numerico in cifre da 1 a 12 (1 per Gennaio, 2 per Febbraio, ecc).
- *Giorno*: corrisponde al giorno in cui è avvenuta la violazione del regolamento comunale e anche la segnalazione e corrisponde ad un numero da 1 a 31.
- *Tipologia verbale*: riporta l'autorità che ha rilevato e verbalizzato l'infrazione, che può essere:
  - ✓ A.S.L.
  - ✓ Carabinieri
  - ✓ Guardia di finanza
  - ✓ Guardie ecologiche
  - ✓ Min. Interno
  - ✓ Polizia Municipale
  - ✓ Questura
  - ✓ Uff. lavori pubblici
- *Tipo sanzione*: è un campo booleano e si riferisce al tipo della sanzione, che può essere:
  - ✓ Amministrativa, principalmente per l'A.S.L., i Carabinieri, la Guardia di Finanza, le Guardie ecologiche, il Min. Interno, la Polizia Municipale, la Questura e gli Uffici lavori pubblici.
  - ✓ Tributaria, per Polizia Municipale e per la Questura.
- *Via 1*: indica la via dove è avvenuta la violazione.
- *Numero civico*: è riferito al precedente campo e indica il numero civico della via dove è avvenuta la violazione.
- *Descrizione sanzione*: è la descrizione riferita al tipo della sanzione che è stata assegnata con il relativo articolo.
- *Descrizione paragrafo*: campo riferito al precedente, è il paragrafo dove è scritta nel dettaglio la sanzione.
- *Descrizione capitolo*: campo riferito al precedente, è il capitolo dove è possibile trovare la sanzione.
- *Descrizione Prontuario*
- *Numero verbali*: è un campo numerico che indica, se maggiore di 1, che sono state rilevate più infrazioni e quindi redatti più verbali, non necessariamente al medesimo soggetto.

Il terzo dataset, che è il più grande di tutti, è sulle violazioni al codice della strada con segnalazioni da gennaio 2011 a giugno 2017, anche queste aggiornate semestralmente, ed è caratterizzato da:

- *Anno*: corrisponde all'anno in cui è avvenuta la violazione al codice della strada.

- *Reg*: è un campo che identifica il regolamento usato che è il "Regolamento di esecuzione e di attuazione del codice della strada".
- *Articolo*: è l'articolo del regolamento che si riferisce nello specifico alla violazione commessa.
- *Classe*: rappresenta la classe del veicolo ovvero il tipo di veicolo con cui è stata commessa l'infrazione.
- *Data infrazione*: è la data in cui è stata commessa l'infrazione nel formato gg/mm/aaaa.
- *Loc1*: è la via della città o la località dove è stata commessa l'infrazione (senza civico)
- *Numciv1*: è il numero civico riferito alla via segnalata nel precedente campo.
- *Bisinternolettera*: è un campo che determina l'interno del civico e può essere rappresentato come numero o lettera o serie di numeri e lettere (es. 31/A)
- *Ora infrazione*: è l'ora in cui è stata commessa l'infrazione che presupponiamo coincida con l'ora della segnalazione in formato hh:mm.
- *Sanzione accessoria*: è un campo che rappresenta la sanzione aggiuntiva assegnata (se necessario), ad esempio oltre alla multa si avrà una sanzione accessoria che può essere il ritiro della patente.
- *Tipo infrazione*: rappresenta il tipo di infrazione commessa che può essere ad esempio: cintura di sicurezza, limiti di velocità, guida senza patente, sosta, precedenza…ecc.
- *Numero Verbali*: è un campo numerico che indica, se maggiore di 1, che sono state rilevate più infrazioni e quindi redatti più verbali, non necessariamente al medesimo soggetto.
- *Sanzioni*: è un campo che rappresenta il numero di sanzioni che è sempre maggiore del numero di verbali. In pratica in ogni riga del dataset avrò un totale di sanzioni, ma non per tutte le sanzioni è stato fatto il verbale.

Questa è la descrizione e la caratterizzazione di tutti gli attributi dei tre dataset che saranno tenuti in considerazione per la città di Torino. Successivamente sono state contate le segnalazioni appartenenti ai tre diversi dataset con un totale di:

- 12.571 per le segnalazioni alla Polizia Municipale
- 46.096 per le violazioni ai regolamenti comunali
- 4.764.683 per le violazioni al codice della strada

ricordando che per gli ultimi due dataset è stato considerato un anno in più, ovvero il 2011.
Per effettuare un'analisi più approfondita e interessante si è pensato di arricchire i dati aggiungendo:

- la circoscrizione (qualora non sia già presente)
- una fascia oraria, rappresentata in tabella 1:

Mattino	Dalle ore 5 alle ore 13
Pomeriggio	Dalle ore 13 alle ore 18
Sera	Dalle ore 18 alle ore 23
Notte	Dalle ore 23 alle ore 5

Tabella 1 – Fascia Oraria

- suddivisione in stagioni, rappresentata in tabella 2:

Primavera	Dal 21 marzo al 20 giugno
Estate	Dal 21 giugno al 20 settembre
Autunno	Dal 21 settembre al 20 dicembre
Inverno	Dal 21 dicembre al 20 marzo

Tabella 2 – Suddivisione in stagioni

- giorno della settimana, ovvero lunedì, martedì e così via.
- giorno feriale o festivo (considerando come festivi le domeniche, 1 gennaio, 6 gennaio, Pasqua e Pasquetta, 25 aprile, 2 giugno, 24 giugno (San Giovanni, patrono di Torino), 15 agosto, 1 novembre, 25 e 26 dicembre).

2.2 Open data in Italia

Dopo aver caratterizzato gli open data della città di Torino, che saranno quelli che verranno analizzati nello specifico con tecniche di data mining, si è svolta una ricerca generale sul numero di open data presenti in Italia.
Una statistica conferma che, attraverso il grafico in figura 2.2, gli open data in Italia si stanno sempre ampliando di anno in anno.

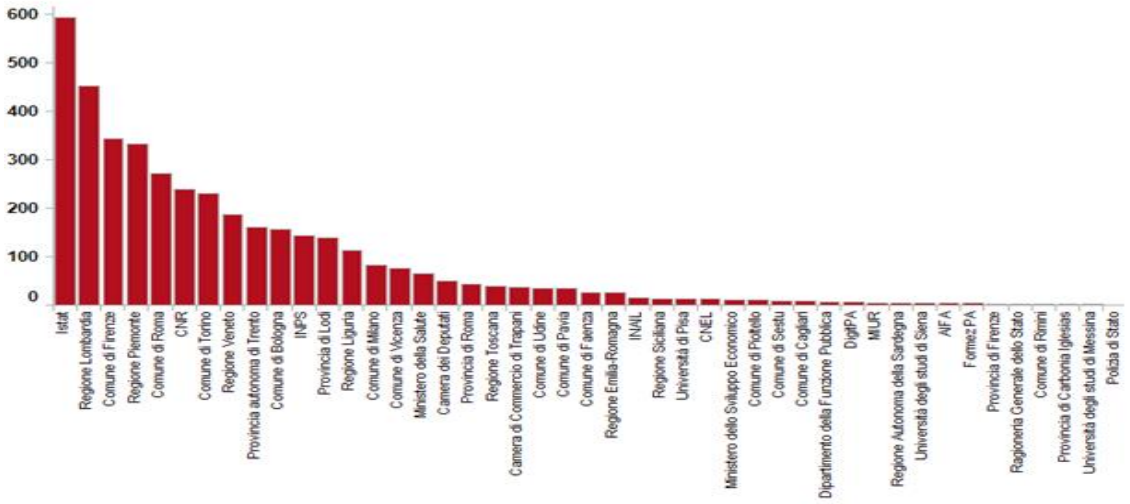


Figura 2.2 - Open data in Italia, tratta da https://www.dati.gov.it/content/quanti-sono-open-data-italia

Dal grafico in figura 2.2, è possibile vedere che le regioni o i comuni che hanno maggiori open data in Italia sono i dati Istat [5], succeduti dalla regione Lombardia e come terzo in classifica il comune di Firenze. Il comune di Torino è classificato come settimo e ha pubblicato quasi la metà dei dati che ha pubblicato l'Istat, mentre la Regione Piemonte si classifica al quarto posto con un po' più della metà dei dati pubblicati dall'Istat. Partendo da questo istogramma, si sono identificate le città italiane con più open data pubblicati e si è fatta un'analisi relativa alla sicurezza urbana, ovvero si sono cercati solamente i dati relativi alla sicurezza tra tutti gli open data di queste città italiane.
Prima città in cui si è svolta la ricerca è stata Roma che si trova al quinto posto nella classifica generale degli open data, nonché capitale d'Italia, attraverso il sito del comune di Roma [6] relativo agli open data. Si sono trovati varie tipologie di dati, in particolare dati su ciclisti coinvolti negli incidenti stradali. Questi dati sono molto dettagliati poiché vi è la data, l'ora, il luogo, la natura dell'incidente, se è uno scontro frontale o laterale, un ribaltamento, un tamponamento (singolo o multiplo) oppure se è stato investito un pedone, il tipo della strada e a quante carreggiate è, se il fondo stradale al momento in cui è avvenuto l'incidente era asciutto o bagnato, se la pavimentazione della strada è asfaltata o meno, se la segnaletica non c'era oppure era orizzontale o verticale, se le condizioni atmosferiche di quel giorno erano serene o c'era pioggia, se le condizioni di traffico erano normali, scarse o intense, se la visibilità era buona, sufficiente o ottima, se l'illuminazione era sufficiente (in ore notturne), quanti sono stati il numero di feriti, di morti e di illesi, inoltre sono indicate la latitudine e la longitudine in cui è avvenuto l'incidente, il tipo di veicolo (in questo caso la bicicletta) con la marca e modello e il tipo di persona se era conducente o passeggero, e tutte le informazioni legate alle persone incidentate con il tipo di lesione riscontrata, se è deceduto e se ha utilizzato il caschetto di sicurezza.
Un altro dataset importante sulla sicurezza che troviamo nel sito del Comune di Roma è quello sugli "incidenti stradali, veicoli coinvolti, persone coinvolte e pedoni coinvolti".

Gli attributi sono identici a quelli descritti precedentemente riguardo i ciclisti con in più informazioni del tipo se è stata usata la cintura di sicurezza, se l'airbag è esploso o inesploso.
A differenza dei dati che abbiamo per Torino, questi dati sono molto più dettagliati, hanno molti più attributi ma quelli relativi ai ciclisti non sono molti (relativi solo all'anno 2017 divisi per mese) mentre i secondi descritti sono parecchi (dal 2006 al 2017). Possiamo inoltre dire che la città di Roma ha inserito sul sito open data molto prima (2006) rispetto alla città di Torino, quindi il numero di dati è maggiore e, sicuramente da notare, i dati pubblicati sono molto dettagliati e approfonditi.
La seconda città italiana di cui si è fatto il relativo studio degli open data è Milano che si trova al quattordicesimo posto della classifica generale degli open data in Italia. Sul sito del comune di Milano [7] sono presenti dati relativi alla mobilità, in particolare incidenti stradali e persone infortunate suddivisi per mese, mese e veicoli coinvolti, mese e natura dell'incidente, mese e zona di decentramento e mese e cerchia cittadina. I dati non sono recentissimi e dopo un po' di tempo non sono più stati aggiornati, ma i dati più vecchi si riscontrano nel 2001, e i più recenti nel 2014. Però la grande differenza rispetto ai dati precedenti è che questi dati sono strutturati come delle statistiche e non come dei singoli avvenimenti come visto precedentemente per Torino e per Roma. Infatti il dataset è composto principalmente dall'anno e dal mese e poi vi è una statistica che riguarda quanti incidenti sono avvenuti, quanti morti, quanti feriti in base a quanti veicoli sono stati coinvolti in quell'incidente stradale, e ne tiene traccia fino a sette veicoli coinvolti. Alcuni dati hanno anche il conteggio di morti e feriti per una determinata zona della città, altri sono suddivisi per la natura dell'incidente quindi ad esempio in quanti incidenti la natura era il tamponamento, tenendo sempre il conteggio dei feriti e dei morti.
La terza città in cui si è svolta la ricerca sugli open data è Firenze. Il comune di Firenze si classifica al terzo posto nella classifica generale sugli open data. Dalle ricerche effettuate sul sito del comune di Firenze [8], nella sezione open data, si è constatato che i dati relativi alla sicurezza sono praticamente inesistenti e sono solamente statistiche e non dati approfonditi, nonostante questa città si classifichi al terzo posto! Principalmente, di quelle poche informazioni sulla sicurezza riscontrate, si trovano dati relativi ai sinistri suddivisi per via, che sono delle statistiche che sono state aggiornate dal 2009 al 2011. In questo dataset gli attributi maggiormente rilevanti sono per ogni indirizzo, quanti sinistri sono avvenuti nel 2009, 2010 e 2011 (tre anni), quanti morti, quante lesioni, quanti contusi e quanti danni. Quindi, in conclusione, i dati presenti sono solo conteggi o statistiche.
Questa è la ricerca effettuata in Italia.

2.3 Open data in Europa e nel Mondo

Guardando al di fuori dei confini italiani, si è cercato di trovare principalmente open data relativi alle città più importanti e famose in Europa e nel mondo. È stato trovato un documento che tratta un argomento in cui hanno discusso le nazioni appartenenti al G8 avvenuto nel 2013 [9]. I leader di 8 nazioni del mondo hanno firmato un accordo che

prevede di pubblicare rispettivamente i propri open data. Da alcune statistiche a gennaio 2015 (dopo 2 anni) la situazione era questa, rappresentata in tabella 3:

Canada	214.033
Stati Uniti	137.601
Regno Unito	20.505
Francia	13.967
Giappone	12.800
Germania	9.799
Italia	9.031
Russia	2.424

Tabella 3 – Open data nei paesi del G8

Nella seconda colonna è indicato il numero di dataset sul portale nazionale a Gennaio 2015 relativo al paese riportato nella prima colonna.
Dopo aver analizzato questo documento, si sono considerate nell'analisi città appartenenti a questi paesi a seconda della posizione in classifica che occupano. Infatti le ricerche successive saranno fatte principalmente in Francia e Regno Unito per quanto riguarda il contesto europeo, e in Canada e negli Stati Uniti per quanto riguarda il contesto extra-europeo.

2.3.1 Contesto Europeo

Figura 2.3 – Europa, tratto da https://cms-assets.packlink.com/media/it/2015/10/europa1.png

Partendo dal contesto europeo, come raffigurato in figura 2.3, la prima città di cui si è fatta la ricerca è stata Londra [10]. Per quanto riguarda Londra, i dati principali si hanno sulle segnalazioni ai vigili del fuoco (London Fire Brigade) che sono i dati appartenenti al dataset più dettagliato che si possa trovare riguardante la città di Londra. In questo dataset si ha il codice univoco dell'incidente, la data e l'ora della chiamata, il luogo (e anche il nome del quartiere), se l'incidente era un servizio speciale, un incendio o un falso allarme, la categoria che può essere ad esempio barca, residenziale, non residenziale, aree aperte, veicoli di strada/ferrovia, il tipo di proprietà per esempio se è un appartamento o una casa, quanti piani ha, se è una banca, un aeroporto, se è una stazione delle ambulanze, una galleria d'arte, un bed & breakfast, una scuola, un ponte, una macchina, un casinò, una cattedrale, e così via! Come informazione aggiuntiva vi è l'ora di quando è arrivato sul posto il primo camion dei vigili del fuoco, quindi il tempo di attesa di quest'ultimo, e anche del secondo camion, se c'è stato. Successivamente, per la città di Londra si hanno altri dataset che indicano il numero di crimini e tasso di criminalità dovuto ai trasporti pubblici, tram e metropolitana (solo statistiche), mentre è stato possibile trovare un altro dataset con il numero di reati e tasso di criminalità nell'area della polizia metropolitana, questi ultimi sono divisi per mese e indicano il volume e il tasso, quindi quanti incidenti con i bus sono stati segnalati. Queste statistiche vanno dal 2009 fino a giugno 2017.
Successivamente il Paese preso in considerazione è la Francia [11]. La Francia, secondo le statistiche del G8, è al quarto posto con ben 13.967 open data, gli unici dati riguardo alla sicurezza che sono stati riscontrati sono quelli relativi alla descrizione e ubicazione degli incidenti di sicurezza del sistema ferroviario SNCF [12]. Non sono statistiche, quindi si hanno segnalazioni complete e approfondite, però gli unici attributi a nostra disposizione sono la data, la localizzazione, il tipo e il commento.
Detto ciò, non si sono più svolte ricerche nel contesto europeo, ma si sono ampliati gli orizzonti per vedere quali e quanti open data si sarebbero riscontrati nel contesto extra-europeo.

2.3.2 Contesto extra-europeo

Figura 2.4 – Mondo, tratta da http://meteo.repubblica.it/img/mondo.png

Tra le città del mondo, in figura 2.4, in cui si è svolta la ricerca, in primis, troviamo Canada e Stati Uniti.
Per quanto riguarda il Canada, le città più famose (e più grandi) che si sono trovate sono Montreal, Toronto, Vancouver [13], Calgary, Halifax e Nuova Scozia, Surrey [14] ed Edmonton [15]. Prendendone una per fare un esempio, per la città di Montreal, la città più popolosa della provincia del Quebec, si ha un dataset sugli interventi dei pompieri di Montreal [16] tenuti dal 2005 al giorno d'oggi. Questi dati sono dettagliati e per esempio gli attributi sono il codice univoco dell'incidente, la data e l'ora in cui è avvenuto, il tipo dell'incidente, la descrizione, latitudine e longitudine e il numero di unità coinvolte. Per citare un altro esempio, prendiamo in considerazione la città di Toronto [17], capoluogo della provincia dell'Ontario e centro più popoloso del Canada, nella sezione "public safety" del sito web dedicato agli open data si hanno dati relativi ad esempio a tutti gli incidenti a cui sono accorsi i vigili del fuoco, incidenti dove sono stati necessari dei paramedici con attributi del tipo: priorità dell'incidente e il numero delle unità arrivate nel luogo dell'incidente. Poi, per la città di Calgary [18], la città più grande della provincia canadese dell'Alberta, si hanno all'incirca gli stessi dataset riferiti in particolare alle chiamate ai vigili del fuoco e agli incidenti generali suddivisi per anno ma, al contrario delle altre due città canadesi, questi sono solamente statistiche o conteggi. Anche per Halifax [19], capitale della Nuova Scozia, si hanno solo statistiche principalmente sui crimini e incidenti in generale, questi ultimi suddivisi in base ai crimini commessi, al reato commesso.
Infine, la ricerca è stata svolta nelle città degli Stati Uniti d'America. In questo paese di possono trovare tantissimi open data, poiché gli USA sono veramente grandi come

estensione geografica e hanno moltissime città che possono essere prese in considerazione.
La ricerca principale è stata svolta su New York [20], città più popolosa degli Stati Uniti situata nell'omonimo stato di New York, dove i dati in particolare si riferiscono a incidenti di veicoli che sono stati segnalati al New York Police Department (polizia di New York), reclami, incidenti in generale, segnalazioni al New York Fire Department (vigili del fuoco) ad esempio per l'evacuazione di edifici, e l'ufficio di emergenze.
La ricerca si è poi spostata nella città di San Francisco [21], quarta città della California per il numero di abitanti, dove le segnalazioni vertono su incidenti al dipartimento di polizia di San Francisco e vigili del fuoco; poi Los Angeles [22], la città più popolosa della California, a differenza delle altre città, questa è toccata dall'Oceano Pacifico quindi si sono trovate segnalazioni al porto di Los Angeles per l'evacuazione delle strade per un possibile tsunami. Sono state successivamente prese in considerazione altre città famose come Las Vegas [23] (città più grande dello stato del Nevada), Seattle [24] (città più popolosa dello stato di Washington), Denver [25] (capitale del Colorado), Chicago [26] (la città più grande dell'Illinois e terza per popolazione di tutti gli Stati Uniti), Houston [27] (città più popolosa della stato del Texas e quarta per popolazione di tutti gli Stati Uniti), Orlando [28] (sesta città più grande della Florida), Detroit [29] (nello stato del Michingan e diciottesima per popolazione di tutti gli Stati Uniti d'America), Philadelphia [30] (città più importante della Pennsylvania e sesta per popolazione di tutti gli Stati Uniti), Washington [31] (la capitale degli Stati Uniti d'America, ventiquattresima per popolazione) e Boston [32] (capoluogo e città più grande del Massachussets).
Non vengono analizzati gli attributi dei dataset per ogni città perché all'incirca coincidono con le prime tre città citate degli Stati Uniti d'America e sono tantissime!
Per concludere, la ricerca è stata svolta anche su altre città extra-europee che non appartenevano a paesi che fanno parte del G8 e si sono trovati dati relativi alla sicurezza a Tokyo [33], capitale del Giappone, a Mosca [34], capitale e città più popolosa della Russia, a Canberra [35], più grande città e capitale australiana. La maggior parte di questi dati, però, sono scritti in giapponese o in russo, quindi non si potrà fare un'analisi a meno che non si conosca la lingua però l'analisi potrebbe essere fatta su quelli australiani che potrebbero essere presi come campione (siccome la lingua parlata in Australia è l'inglese).
Quindi, volendo concludere questa parte che riguarda un'overwiew generale degli open data relativi alla sicurezza urbana si è arrivati a dire che non molte città hanno open data dettagliati, la maggior parte sono conteggi di avvenimenti o statistiche per anno o mese; non tutte le città cercate hanno open data inerenti alla sicurezza (ad esempio a Parigi [11] in Francia vi sono altri tipi di dati come elenco wi-fi pubblici, uso mensile dei wi-fi pubblici, elenco tribunali, posizione cestini per la raccolta differenziata). In merito all'analisi dei dati, è preferibile utilizzare dati più approfonditi rispetto a statistiche o conteggi, ma anche quest'ultimi possono essere tenuti in considerazione.

3. Introduzione alle tecniche di data mining

In questo capitolo verrà fatta un'introduzione generale alle tecniche di data mining. Saranno inoltre descritte più in dettaglio le tecniche di data mining utilizzate in questa tesi per l'analisi dei dati sulla sicurezza urbana. In particolare, saranno introdotte le regole di associazione, con i relativi indici utilizzati per valutare la qualità delle regole estratte quali supporto, confidenza, lift. Nel terzo paragrafo di questo capitolo è presentato Rapidminer, il software utilizzato per l'estrazione delle regole di associazione, e gli operatori adoperati all'interno di esso, con la spiegazione dell'algoritmo utilizzato (FP-Growth). Nel quarto paragrafo, invece, verrà introdotto il classificatore associativo, utilizzato alternativamente alle regole di associazione per l'analisi, sul medesimo dataset. Inoltre, sarà presentato Weka, il software utilizzato per la classificazione e l'algoritmo L3.

3.1 Data Mining

Il Data Mining [36] è il processo di scoperta di relazioni, pattern (una struttura, un modello, o, in generale una rappresentazione sintetica dei dati), ed informazioni precedentemente sconosciute e potenzialmente utili, all'interno di grandi basi di dati. È quindi anche l'insieme di tecniche e metodologie che hanno per oggetto l'estrazione di una informazione o di una conoscenza a partire da grandi quantità di dati.
I vantaggi del Data Mining sono molteplici. In primo luogo i dati che possono essere trattati sono eterogenei: possono essere dati quantitativi, qualitativi, testuali, immagini e suoni; si applica a qualunque fonte di dati. Un altro motivo principale è che non richiede ipotesi a priori da parte del ricercatore (non serve la conoscenza approfondita dei dati ad esempio).
Vi è la possibilità di elaborare un numero elevato di osservazioni e anche di variabili. Esistono algoritmi ottimizzati per minimizzare il tempo di elaborazione e per una più chiara visualizzazione dei risultati. Infine, la semplicità di interpretazione del risultato.
Perché fare data mining?

- La quantità dei dati memorizzata su supporti informatici è in continuo aumento (ad esempio Pagine Web, sistemi di e-commerce, dati relativi ad acquisti/scontrini fiscali, transazioni bancarie e relative a carte di credito…)
- L'hardware diventa ogni giorno più potente e meno costoso.
- La pressione competitiva è in continua crescita: la risorsa informazione è un bene prezioso per superare la concorrenza.
- I dati prodotti e memorizzati crescono a grande velocità (GB/ora): ad esempio sensori posti sui satelliti, telescopi, microarray che generano espressioni genetiche, simulazioni scientifiche che producono terabyte di dati…

- Le tecniche tradizionali sono inapplicabili alle masse di dati grezzi
- Il Data mining può aiutare gli scienziati a classificare e segmentare i dati e a formulare ipotesi.

Il data mining risulta utile se fatto principalmente su grandi set di dati. Molte delle informazioni presenti sui dati non sono direttamente evidenti e le analisi guidate dagli uomini possono richiedere settimane per scoprire informazioni utili e magari larga parte dei dati non vengono di fatto mai analizzati.
È molto facile confondersi: è stato detto precedentemente che il data mining è l'estrazione di informazioni potenzialmente utili dai dati. Ad esempio la ricerca di un cognome sull'elenco telefonico per risalire al numero di telefono di quella determinata persona, non è un esempio di data mining poiché non si trae nessuna conclusione. Si può invece dire che fare un'analisi dei cognomi delle persone più comuni in una certa zona è data mining.
Che cosa è il data mining? Oltre all'esempio precedente dei cognomi più comuni, un altro esempio può essere fare una ricerca nel web su una parola chiave e classificare i documenti trovati secondo un criterio semantico (per esempio "corriere": nome di giornale, professione, ecc.). Oppure scoprire chi sono i clienti che hanno maggiore propensione di acquisto su certi prodotti.

Il data mining viene utilizzato per cercare correlazioni tra più variabili relativamente ai singoli individui; ad esempio conoscendo il comportamento medio dei clienti di una compagnia telefonica cerco di prevedere quanto spenderà il cliente medio nell'immediato futuro.
Quindi le attività tipiche del data mining sono principalmente due:

- Predizione: nello specifico, utilizzare alcune variabili per predire il valore incognito o futuro di altre variabili. Per la predizione dei dati si usano le tecniche di classificazione, le tecniche di regressione e la "Deviation Detection" ovvero un'analisi delle anomalie.
- Descrizione: trovare pattern interpretabili dall'uomo che descrivano i dati. Per la descrizione dei dati si usano principalmente tecniche di Clustering e regole di associazione.

E' importante osservare che non esiste una tecnica "superiore" alle altre, ma ogni tecnica è riferita a determinati obiettivi e tipologie di dati da analizzare. Spesso i migliori risultati per trasformare i dati in informazioni si ottengono attraverso la combinazione di diverse tecniche di analisi.

3.2 Regole di associazione

Le regole di associazione [36], prevalentemente utilizzate in questa tesi, sono uno dei metodi per estrarre relazioni nascoste tra i dati.

Sono state introdotte per la scoperta di regolarità all'interno delle transazioni registrate nelle vendite dei supermercati. Per fare un esempio, la regola {cipolle, patate} ⇒ {hamburger} individuata nell'analisi degli scontrini di un supermercato indica che il se il cliente compra insieme cipolle e patate è probabile che acquisti anche della carne per hamburger. Tale informazione può essere utilizzata come base per le decisioni riguardanti le attività di marketing, come ad esempio le offerte promozionali o il posizionamento dei prodotti negli scaffali.
Le regole di associazione hanno come scopo quello di trovare associazioni interessanti e relazioni di correlazione in grandi insiemi di transazioni. Il dominio applicativo di queste regole sono le grandi collezioni di dati che possono essere raccolti con grande facilità se esiste un concetto di "transazione" (ad esempio: scontrini di supermercato).
Le regole di associazione sono una sorta di "implicazioni". La regola X ⇒Y viene interpretata come: "nelle transazioni in cui compare X compare anche Y"; X è detto corpo o rule body, Y è detta testa o rule head.
Le regole di associazione sono caratterizzate principalmente da due misure statistiche: supporto, e confidenza.

- Il supporto indica la percentuale di transazioni che contengono entrambe X ed Y. È l'indicazione di quanto frequentemente l'itemset appare nel dataset. Infatti si sono analizzate regole con il supporto maggiore, ovvero nel trovare l'itemset che compare di più nel dataset.
  Se la regola è X ⇒ Y, il supporto sarà:
  $$\frac{\#\{X, Y\}}{|T|}$$
  Ad esempio se l'itemset {hamburger, patate} ha un supporto pari a 20%, allora esso comparirà nel dataset per il 20% delle transazioni.
- La confidenza indica, date le transazioni che contengono X, qual è la percentuale di transazioni che contengono Y. È un'indicazione di quanto spesso la regola è stata trovata vera.
  Se la regola è X ⇒ Y, la confidenza sarà:
  $$\frac{\mathrm{supp}(X \sqcup Y)}{\mathrm{supp}(X)}$$
  Ad esempio se la regola {hamburger, patate} ⇒ {cipolle} ha una confidenza pari a 1, allora per il 100% delle transazioni contenenti hamburger e patate la regola è corretta.

In questa tesi, per un approfondimento maggiore, non si è tenuto conto solamente del supporto e della confidenza come metodi per l'analisi dei dati. È stata aggiunta una nuova misura statistica: il lift.

- Il lift è una misura delle performance di un modello: esso dice in che misura le due ricorrenze che sto prendendo in considerazione dipendono l'una dall'altra. Se lift è pari a 1 allora le ricorrenze sono indipendenti, mentre quando è minore di uno si avrà una correlazione negativa tra le due ricorrenze e quando è maggiore

di 1 la correlazione tra le ricorrenze è positiva. In questo elaborato si sono prese solamente le correlazioni positive poiché, dato il dataset descritto, le correlazioni negative non avrebbero avuto senso.
Se la regola è $X \Rightarrow Y$, il lift sarà:

$$\frac{\text{supp}(X \sqcup Y)}{\text{supp}(X) \cdot \text{supp}(Y)}$$

Ad esempio se la regola {hamburger, cipolle} ⇒ {patate} ha un lift pari a 1,25, allora vuol dire che vengono acquistati hamburger e cipolle, allora la probabilità che vengano acquistate anche le patate cresce di 1,25 volte.

Il problema di estrarre regole di associazione è definito come il problema di estrarre tutte le regole con un supporto superiore al parametro min_sup, una confidenza superiore al parametro min_conf, e un lift strettamente maggiore di 1, come si vedrà successivamente. Le regole che soddisfano questi vincoli sono dette regole "forti" e avranno un maggior peso nell'analisi svolta.

3.3 Rapidminer

RapidMiner [37], logo in figura 3.1, è un software molto completo per analizzare dati in grandi in quantità.

Figura 3.1 – Logo del software Rapidminer, tratta da https://1xltkxylmzx3z8gd647akcdvov-wpengine.netdna-ssl.com/wp-content/uploads/2016/06/rapidminer-logo-retina.png

RapidMiner, precedentemente noto come YALE (Yet Another Learning Environment), è stato sviluppato a partire dal 2001 da Ralf Klinkenberg, Ingo Mierswa e Simon Fischer presso l'Unità di Intelligenza Artificiale dell'Università Tecnica di Dortmund, Germania. A partire dal 2006, il suo sviluppo è stato guidato da Rapid-I, società fondata da Ingo Mierswa e Ralf Klinkenberg nello stesso anno. Nel 2007, il nome del software è stato cambiato da YALE a RapidMiner e nel 2013, la società si è spostata da Rapid-I a RapidMiner.
RapidMiner è un software che offre procedure di data mining e machine learning che comprendono: caricamento e trasformazione dei dati (estrazione, trasformazione), preelaborazione e visualizzazione dei dati, analisi predittiva e modellazione statistica, valutazione e implementazione. RapidMiner è scritto nel linguaggio di programmazione Java. RapidMiner fornisce una GUI, *Graphical User Interface* (interfaccia grafica), per progettare ed eseguire flussi di lavoro analitici. Questi flussi di lavoro sono chiamati "Processi" in RapidMiner e sono costituiti da più "Operatori". Ogni operatore esegue una singola attività all'interno del processo e l'output di ciascun operatore costituisce l'input

di quello successivo. In alternativa, il motore può essere richiamato da altri programmi o utilizzato come API. RapidMiner fornisce schemi di apprendimento, modelli e algoritmi e può essere esteso utilizzando gli script R e Python.
La funzionalità RapidMiner può essere estesa con plugin aggiuntivi resi disponibili tramite RapidMiner Marketplace. Il Marketplace di RapidMiner offre agli sviluppatori una piattaforma per creare algoritmi di analisi dei dati e pubblicarli nella comunità. In questo elaborato sono stati aggiunti a Rapidminer 4 plugin aggiuntivi:

- Parallel Processing Extension
- Weka Extension
- Wordnet Extension
- Text mining Extention

3.3.1 Il Processo e l'algoritmo FP-Growth

Sono elencati di seguito i principali operatori utilizzati per la creazione del processo con una breve descrizione, tratti dalla documentazione ufficiale di Rapidminer [38].

Read excel: Questo operatore legge un ExampleSet dal file Excel specificato. La tabella dei dati presente nel foglio del file Excel deve avere un formato tale che ogni riga sia un esempio e ogni colonna rappresenti un attributo. La prima riga del foglio Excel è spesso utilizzata per i nomi degli attributi che possono essere indicati da un parametro. I valori dei dati mancanti in Excel devono essere indicati da celle vuote o da celle contenenti solo "?".

Select Attributes: Questo operatore seleziona quali attributi di un ExampleSet dovrebbero essere conservati e quali attributi dovrebbero essere rimossi. Questo viene utilizzato nei casi in cui non siano necessari tutti gli attributi di un ExampleSet. Spesso c'è bisogno di selezionare gli attributi prima di applicare alcuni operatori. Ciò è particolarmente vero per i set di dati grandi e complessi. L'operatore Select Attributes consente di selezionare in modo appropriato gli attributi richiesti. Sono disponibili diversi tipi di filtri per rendere facile la selezione degli attributi. Solo gli attributi selezionati verranno considerati ovvero consegnati dalla porta di uscita e il resto verrà rimosso.

Discretize by User Specification: Gli operatori di discretizzazione possono essere utilizzati per modificare il valore degli attributi numerici agli attributi nominali. Questo operatore mappa gli attributi numerici selezionati in classi specificate dall'utente. Gli attributi numerici selezionati verranno modificati in attributi nominali o ordinali. I valori numerici sono mappati alle classi in base alle soglie specificate dall'utente nel parametro classes. L'utente può definire le classi specificando il limite superiore di ogni classe o, come in questo caso, l'unico valore assunto dalla classe. Se una classe viene chiamata '?', I valori numerici che rientrano in questa classe saranno sostituiti da valori sconosciuti negli attributi risultanti.

Text to Nominal: Questo operatore modifica il tipo di attributi di testo selezionati in nominale. Inoltre, associa tutti i valori di questi attributi ai corrispondenti valori nominali. Ogni valore di testo viene semplicemente utilizzato come valore nominale del nuovo attributo. Se il valore è mancante nell'attributo text, mancherà anche il nuovo valore.

Numerical to Binomial: Questo operatore modifica il tipo degli attributi numerici selezionati in un tipo binomiale (chiamato anche binario). Esso inoltre mappa anche tutti i valori di questi attributi ai corrispondenti valori binomiali. Gli attributi binomiali possono avere solo due valori possibili, vale a dire "true" o "false". Se il valore di un attributo è tra il valore minimo e massimo definito, diventa "falso", altrimenti "true". I valori minimi e massimi possono essere specificati rispettivamente dai parametri min e max. Se il valore manca, il nuovo valore manca. I limiti predefiniti sono entrambi impostati a 0,0, quindi solo 0,0 viene mappato a "falso" e tutti gli altri valori vengono mappati in modo "true" per impostazione predefinita.

Nominal to Binominal: Questo operatore modifica il tipo di attributi nominali selezionati in un tipo binomiale. Inoltre, mappa tutti i valori di questi attributi ai valori binomiali (true e false). Ad esempio, se viene trasformato un attributo nominale con il nome "costi" e possibili valori nominali "bassi", "moderati" e "alti", il risultato è un insieme di tre attributi binomiali "costi = bassi", "costi = moderati " e " costi = alti". Solo il valore di uno di questi attributi è vero per un esempio specifico, il valore degli altri attributi è falso. Dal dataset originale dove l'attributo "costi" aveva valore "bassi", nel nuovo dataset questi esempi avranno l'attributo "costi = bassi" impostato su "true", il valore di "costi = moderati" e “costi = alti” impostati sul valore “false”.

FP-Growth: Questo operatore calcola in modo efficiente tutti gli itemset frequenti dal dataset specificato utilizzando la struttura di dati dell'albero FP. È obbligatorio che tutti gli attributi dell'input ExampleSet siano di tipo “binominal”, infatti saranno necessari i 3 blocchi aggiunti precedentemente “Text to Nominal”, “Numerical to Binominal” e “Nominal to Binominal”.
In parole semplici, gli itemset frequenti (Frequent Pattern) sono gruppi di elementi che appaiono spesso nell’insieme dei dati.
Il numero di transazioni (ovvero righe totali del dataset) è solitamente assunto come molto grande.
Il problema degli itemset frequenti è quello di trovare insiemi di elementi che appaiono insieme almeno sopra una certa soglia. Questa soglia è definita dai criteri di "supporto minimo".
Il fatto di trovare itemset frequenti è spesso visto come la scoperta di "regole di associazione".
È stato scelto l’algoritmo FP-Growth per questa analisi, tuttavia esistono anche molti altri algoritmi di estrazione di itemset frequenti, ad esempio l'algoritmo Apriori. Uno dei principali vantaggi di FP-Growth rispetto ad Apriori è che utilizza solo due scansioni di dati ed è quindi spesso applicabile anche su grandi set di dati.

Questo operatore ha due modalità di lavoro di base:

- trovare almeno il numero specificato di itemset frequenti con il supporto più alto senza tenere conto del "supporto minimo".
- trovare tutti gli elementi con un supporto più grande del supporto minimo specificato.

Create Association Rule: Come si è detto in nel paragrafo precedente, l'operatore FP-Growth trova gli itemset frequenti e poi sono necessari operatori come "Create association rule" che utilizzano questi itemset frequenti per l'estrazione delle regole di associazione.
Le regole di associazione sono dichiarazioni if / then che aiutano a scoprire relazioni tra dati apparentemente non correlati. Un esempio di regola dell'associazione potrebbe essere "Se un cliente acquista uova, ha l'80% di probabilità di acquistare anche latte".
Una regola di associazione ha due parti, un antecedente (if) e un conseguente (then) combinato con l'antecedente.
Le regole di associazione vengono create analizzando i dati per gli itemset if / then frequenti e utilizzando i criteri di supporto e confidenza per identificare le relazioni più importanti. Il supporto è un'indicazione di quanto frequentemente gli articoli compaiono nel database. La confidenza indica il numero di volte in cui le affermazioni if / then sono state vere.
Tali informazioni possono essere utilizzate come base per le decisioni relative ad attività di marketing quali, ad esempio, prezzi promozionali o posizionamenti di prodotti. Oltre all'esempio sopra riportato, le regole di associazione dell'analisi del mercato sono utilizzate oggi in molte aree di applicazione, tra cui l'utilizzo del Web.

3.4 Il classificatore associativo

In questo capitolo verrà introdotto il classificatore associativo [39], che è un metodo che è stato utilizzato alternativamente alle regole di associazione per l'analisi, sul medesimo dataset.

3.4.1 Descrizione del classificatore associativo

Il classificatore associativo trova un modello per l'attributo di classe che esprima il valore dell'attributo in funzione dei valori degli altri attributi.
Quindi, in poche parole, ogni record è composto da un insieme di attributi, di cui uno esprime la classe di appartenenza del record. Lo scopo del classificatore associativo è associare il record ad una classe nel modo più accurato possibile. Infatti, viene utilizzato un test set per individuare l'accuratezza del modello.
In primis, il dataset viene suddiviso in:

- Training set, usato per costruire il modello
- Test set, usato per testare e validare il modello.

I classificatori vengono usati ad esempio nel campo pubblicitario, vengono raccolti dati su prodotti usati dai clienti e vengono lanciati prodotti simili a quelli; oppure per l'individuazione di frodi, ad esempio tenendo traccia delle precedenti transazioni di una carta di credito per delineare un profilo del possessore e così via.
Il problema che sorge usando modelli di apprendimento automatico è che non si conoscerà il rendimento fino a quando non si testeranno le sue prestazioni su un set di dati indipendente ovvero quelli che non è stato utilizzato per addestrare il modello.
In questo caso si è adottata la 10-fold cross validation, che è la tecnica utilizzata in questa tesi, per predire le prestazioni del modello creato.
Utilizzando questa tecnica, i dati in input verranno divisi in 10 parti: 9 parti su 10 saranno usate per creare il modello e la decima parte verrà predetta. In base a quanto è accurato il modello, si avrà una stima di quanto si è stati corretti a predire l'ultima parte. Quindi ci sarà una fase di "addestramento" del modello e una fase di test.
Più il modello è stato addestrato bene, più sarà in grado di predire i dati nella fase di test: questa stima viene riassunta in un indice che è l'accuratezza del modello, e viene espressa in percentuale.
L'accuratezza, perciò, è la probabilità che il modello azzecchi correttamente l'etichetta di classe.
Inoltre, si terrà conto ancora di altri due indici: il richiamo e la precisione. Essi servono a valutare la qualità del classificatore, infatti vanno a valutare la bontà del modello separatamente per ogni classe.

3.4.2 WEKA e L3

Weka [40] è un software open source rilasciato sotto GNU General Public License in grado di fare data mining utilizzando la piattaforma Java, attraverso una raccolta di algoritmi di apprendimento automatico che possono essere applicati direttamente a un set di dati.

Figura 3.2 – Logo del software Weka, tratto da https://upload.wikimedia.org/wikipedia/commons/0/07/Weka_%28software%29_logo.png

Il nome del software deriva da un uccello trovato solo sulle isole della Nuova Zelanda, come dal logo rappresentato in figura 3.2, incapace di volare e con una natura curiosa.

Utilizzando Weka, si è introdotto un plugin in cui è proposto un nuovo classificatore associativo che si basa su un approccio di sfoltimento delle regole di associazione generate.
Questo classificatore è stato chiamato L3 [41], che sta per Live e Let Live. Il nome sta ad indicare che questa tecnica consente di ridurre il set di regole generate eliminando quelle dannose cioè quelle che classificano erroneamente i dati di allenamento. L3 consente di rappresentare gruppi di regole di grandi dimensioni e la forma compatta proposta in L3 non fa avvenire la perdita di informazioni.
In questa tesi sono state considerate solamente le regole di livello 1, ovvero quelle di alta qualità.
Inoltre, è possibile raggiungere soglie di supporto molto basse e utilizzare dataset di grandi dimensioni.

3.4.3 Costruzione del processo con il classificatore associativo

Per quanto riguarda la costruzione del processo con il software Weka, sono state fatte delle piccole modifiche al dataset di input, poiché fosse leggibile dal software e fosse possibile inoltre applicare l'algoritmo L3, così da estrarne le regole principali, chiamate anche regole di primo livello, sia per la categoria che per la sottocategoria, facendone un'analisi qualitativa. Inoltre, sarà fatta un'ulteriore analisi quantitativa delle prestazioni del classificatore associativo in termini di accuratezza.
Innanzitutto, partendo dai dataset originali, siccome in Weka gli attributi devono essere tutti nominali per essere processati, tramite la funzione di Excel "Trova e Sostituisci", si sono sostituite principalmente le cifre che denominavano la circoscrizione, il mese e l'anno: per fare un esempio la circoscrizione da 1 diventa c1, da 2 diventa c2 e così via, e si è utilizzata la lettera "c" per la circoscrizione, la lettera "m" per il mese e la lettera "a" per l'anno (solo nel dataset intero, non quello frammentato per anni perché vi era già la suddivisione in anni). Quindi i valori che può assumere la circoscrizione sono: c1, c2, c3, c4, c5, c6, c7, c8, c9 e c10; i valori che può assumere il mese sono: m1, m2, m3, m4, m5, m6, m7, m8, m9, m10, m11 e m12; i valori che può assumere l'anno sono: a2012, a2013, a2014, a2015 e a2016.
Inoltre, si sono sostituite tutte le località che avevano accenti (contrassegnati come apici singoli) con uno spazio; la categoria è stata sostituita con la prima parola della categoria stessa; si sono spostate le colonne della categoria e della sottocategoria come rispettivamente ultima e penultima colonna, poiché in Weka l'etichetta di classe deve risultare sempre come l'ultimo campo.
Successivamente, i dataset di input sono stati dati in pasto al software Weka, dopo averli trasformati in un certo formato chiamato .arff, che è il formato standard di Weka. Questo formato è composto da un'intestazione (@relation), che contiene il titolo del file e i nomi

degli attributi (@attribute) con elencati tutti i valori che essi possono avere e dal corpo dati (@data), separati da virgole e una riga per ogni istanza.
Una volta aperto il file in Weka nella versione .arff, nel predire la categoria, ovvero utilizzando la categoria come attributo di classe, si è rimossa la sottocategoria così che non comparissero correlazioni che sarebbero state ovvie, ad esempio: è noto che una certa sottocategoria comparirà sempre con la sua stessa macrocategoria (atti di vandalismo comparirà sempre con allarme sociale perché atti di vandalismo è una sottocategoria appartenente alla categoria allarme sociale). Stessa cosa per predire la sottocategoria, ovvero utilizzando la sottocategoria come attributo di classe, si è eliminata la categoria.
Dopo questo pre-processing dei dati e si è ulteriormente controllato che tutti gli attributi siano di tipo nominale altrimenti con il classificatore associativo si dovrebbe utilizzare la funzione "Discretize", e non sarebbe correttissimo usarla in questo caso perché si vuole fare un re-mapping e non una discretizzazione di un attributo.
Nella classificazione si è potuto scegliere quale supporto minimo utilizzare, quale confidenza minima e, inoltre, è stato possibile scegliere una cartella sul quale salvare i file temporanei, dopo aver premuto "Start" e fatto partire il classificatore.
Si otterrà un output nel quale verranno indicati gli attributi che sono stati presi in considerazione, il numero delle istanze correttamente classificate e con una percentuale sulle istanze totali che rappresenta l'accuratezza del classificatore, e quelle non correttamente classificate. Inoltre, sono predetti altri due parametri: precisione e richiamo riferiti al relativo attributo di classe. Nell'output è presente anche l'informazione di quante regole sono generate dall'algoritmo L3: prima vengono generate le regole di primo livello, considerate le regole di alta qualità, e vi è riportato il numero di quante sono, dopo vengono generate quelle di secondo livello ed è riportato il conteggio anche per queste.
Le regole vengono salvate in una cartella temporanea, se scelgo la modalità di debug "=true" nelle impostazioni dell'algoritmo. Questa cartella temporanea è composta da una serie di file, tra i quali è presente un file chiamato "class_labels-k0.cls" in cui è presente il numero associato all'attributo di classe; vengono associati numeri poiché usando le stringhe il processo impiegherebbe moltissimo tempo in più, infatti si è dovuto creare un programmino in Java per decodificare questi numeri e trasformarli nella corrispondente stringa, leggibile da una persona umana. Gli attributi di classe prendono il valore da 200 in poi, in particolare 200 è quando il classificatore non riesce a classificare un'istanza correttamente ovvero dove non è possibile predire un attributo di classe per quella regola; 201 rappresenterà il primo attributo di classe che il classificatore troverà nei dati di train e così via per tutti gli attributi di classe.
Nel file binario "bin_convertion-k0.bin" sono salvati dei numeri corrispondenti ai valori degli attributi (contenenti nel file di train ovvero "train-k0.arff") che andranno a costituire la creazione del "dizionario" sul quale ci si è basati per la "traduzione" delle regole (da numero intero a stringa). Infatti se l'attributo è già stato trovato precedentemente, ad esso non sarà assegnato un nuovo numero ma sarà riutilizzato il numero assegnato

precedentemente e così via per tutti gli attributi; mentre se non gli è ancora assegnato nessun numero, gli verrà assegnato un numero pari al numero di attributi già trovati (senza doppioni) incrementato di uno. Quindi il dizionario creato sarà:

- numero -> valore attributo 1
- numero -> valore attributo 2
- …

Successivamente si ha il file testuale "levelI-k0.txt" dove sono rappresentate tutte le regole con i valori numerici al posto delle stringhe, per motivi di efficienza e di spazio occupato. Le regole che si trovano sono rappresentate come segue:

- {1,2,3,5} -> 201 5 80.0 0

Nell'esempio citato sopra, i numeri 1,2,3,5 rappresentano quattro item diversi, quindi quattro valori di attributi diversi, che sono correlati tra loro. Dopo la freccia è presente il valore maggiore di 200 che rappresenta l'attributo di classe assegnato a quella regola, in questo caso 201; 5 rappresenta il supporto della regola e 80.0 rappresenta la confidenza di quella regola. Infine, lo zero rappresenta quante volte la regola è stata trovata falsa, ovvero il numero di dati classificati scorrettamente da questa regola: nel caso in esempio zero significa che la regola non è stata trovata sbagliata in nessun caso. Di solito, nelle regole di primo livello, quest'ultima cifra è sempre zero oppure è molto bassa poiché sono regole di alto livello.

Dopo la creazione del dizionario attraverso il file binario, e la traduzione delle regole attraverso il file delle regole, è stato scritto un file testuale (sempre tramite Java) che contenesse le "regole tradotte" e si sono successivamente analizzate.

Quindi, si è dovuto implementare un programmino scritto in Java che fa le seguenti operazioni:

- legge il dataset di input ("train-k0.arff") e il file binario ("bin_convertion-k0.bin") e crea il "dizionario" (salvato in una struttura dati di tipo chiave-valore)
- legge il file numerico delle regole ("levelI-k0.txt") e lo salva in una struttura dati (una mappa) in grado di memorizzare elementi nella forma di coppie chiave-valore; ogni elemento delle mappa è identificato da una determinata chiave, in questo caso la chiave è il numero intero che rappresenta l'attributo e il valore è l'attributo stesso associato a quel numero
- legge il file degli attributi di classe ("class_labels-k0.cls") e salva l'ordine degli attributi
- utilizzando le informazioni salvate precedentemente nelle strutture dati, scrive il file testuale "EstrazioneRegole.txt" contenente le regole leggibili.

Quest'operazione è stata svolta per ogni anno separatamente per categoria e per sottocategoria. Questo programmino è stato di fondamentale importanza, perché con la mente umana non sarebbe stato possibile leggere e decifrare le regole codificate in binario

ottenute dall'algoritmo. Successivamente, queste regole sono state analizzate, come si vedrà nel capitolo 6.

4. Pre-processing del dato

In questo capitolo verranno evidenziati tutti i passi effettuati per la preparazione dei dati alla successiva fase di analisi.

4.1 Arricchimento dati e formule excel

Questa analisi è stata svolta in particolare sul dataset delle segnalazioni al Contact Center della Polizia Municipale.
I dati appartenenti a questo dataset precedentemente erano strutturati secondo 7 attributi principali:

- Categoria
- Sottocategoria
- Circoscrizione
- Località
- Area verde
- Data
- Ora.

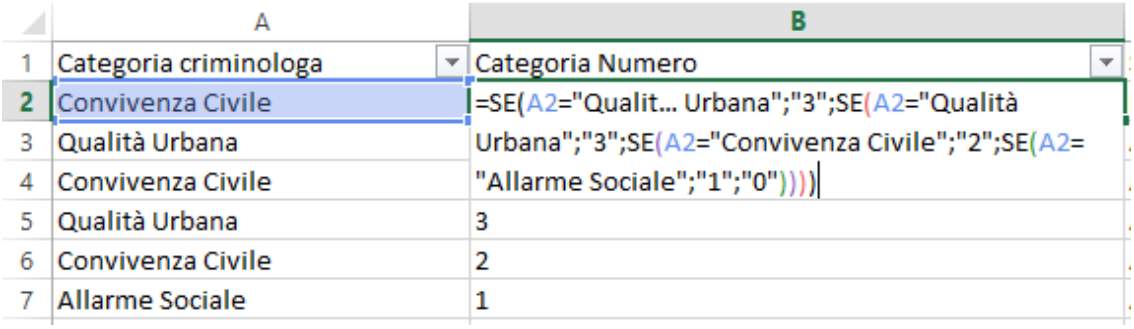
Partendo dal primo attributo, si sono effettuate delle modifiche al dataset di partenza in modo che esso potesse essere meglio leggibile dal software per eseguire l'algoritmo.

Si è notato che le categorie di segnalazioni erano tre. Ogni categoria è stata assegnata ad un numero come lo schema qui sotto riportato:

Allarme Sociale	1
Convivenza Civile	2
Qualità Urbana	3

Tabella 4 – Assegnazione numero alla Categoria

Questi numeri sono stati assegnati attraverso il comando IF, ELSE di Microsoft Excel, come si vede in figura 4.1.



	A	B
1	Categoria criminologa	Categoria Numero
2	Convivenza Civile	=SE(A2="Qualit... Urbana";"3";SE(A2="Qualità
3	Qualità Urbana	Urbana";"3";SE(A2="Convivenza Civile";"2";SE(A2=
4	Convivenza Civile	"Allarme Sociale";"1";"0"))))
5	Qualità Urbana	3
6	Convivenza Civile	2
7	Allarme Sociale	1

Figura 4.1 – Formula che assegna il numero alla categoria

La stessa cosa è stata fatta per le sottocategorie, evidenziando per quale macrocategoria appartenessero:

Atti di vandalismo	11
Altro (A)	12
Aggregazioni giovanili	21
Comportamenti molesti	22
Disturbi altri animali	23
Disturbi cani	24
Disturbi da locali	25
Rumori molesti	26
Uso improprio di parti comuni	27
Altro (C)	28
Decoro e degrado urbano	31
Veicoli abbandonati	32
Altro (Q)	33

Tabella 5 – Assegnazione numero alla Sottocategoria

Come si può vedere la prima cifra sta ad indicare una delle tre macrocategorie, mentre la seconda cifra identifica la sottocategoria.
Anche esse sono state formulate con Microsoft Excel, come rappresentato in figura 4.2, 4.3 e 4.4:

	C	D
1	Sottocategoria Criminologica	Sottocategoria
2	Atti di vandalismo	=SE(C2="Atti di vandalismo";11;12)
3	Altro	12
4	Aggregazioni giovanili	21

Figura 4.2 – Formula che assegna il numero alla sottocategoria

	C	D
1	Sottocategoria Criminologica	Sottocategoria
2	Atti di vandalismo	11
3	Altro	12
4	Aggregazioni giovanili	=SE(C4="Aggregazioni giovanili";21;SE(C4=
5	Comportamenti molesti	"Comportamenti molesti";22;SE(C4="Disturbi
6	Disturbi altri animali	altri animali";23;SE(C4="Disturbi Cani";24;SE(C4=
7	Disturbi Cani	"Disturbi da locali";25;SE(C4="Rumori molesti";
8	Disturbi da locali	26;SE(C4="Uso improprio parti comuni";27;28))))
9	Rumori molesti	)))
10	Uso improprio parti comuni	27
11	Altro	28

Figura 4.3 – Formula che assegna il numero alla sottocategoria

	C	D
1	Sottocategoria Criminologica	Sottocategoria
2	Atti di vandalismo	11
3	Altro	12
4	Aggregazioni giovanili	21
5	Comportamenti molesti	22
6	Disturbi altri animali	23
7	Disturbi Cani	24
8	Disturbi da locali	25
9	Rumori molesti	26
10	Uso improprio parti comuni	27
11	Altro	28
12	Decoro e degrado urbano	=SE(C12="Decoro e degrado urbano";31;SE(C12=
13	Veicoli abbandonati	"Veicoli abbandonati";32;33))
14	Altro	33

Figura 4.4 – Formula che assegna il numero alla sottocategoria

In particolar modo, si sono create tre formule diverse per evidenziare la voce "Altro". Infatti inizialmente leggendo solo la colonna riferita alla sottocategoria, quando si andava incontro alla voce "Altro" non si era in grado di distinguere a quale macrocategoria appartenesse. Con le tre formule separate, invece, si sono distinte le tre voci e sono state chiamate:

- Altro (A): con la "A" che si riferisce ad Allarme Sociale
- Altro (C): con la "C" che si riferisce ad Convivenza Civile
- Altro (Q): con la "Q" che si riferisce ad Qualità Urbana.

Gli attributi circoscrizione e località non sono stati modificati.
L'attributo "Area Verde" inizialmente era un campo vuoto; se la segnalazione era avvenuta in un'area verde allora venivano scritte le parole "Area Verde" nella cella della riga corrispondente. Questo non va bene, perché ogni cella vuota, che significa che non è un'area verde, viene interpretata dal software come informazione mancante. Perciò attraverso una formula, si è aggiunto "Si" dove precedentemente compariva la scritta "Area Verde", "No" altrimenti, come rappresentato in figura 4.5.

	G	H
1	Area Verde	Area Verde
2		No
3		No
4		No
5	Area Verde	=SE(G5="Area Verde";"Si";"No")
6		No
7		No
8		No
9		No
10	Area Verde	Si
11		No
12		No

Figura 4.5 – Formula che assegna il valore all'area verde

Successivamente i "Si" e "No" sono diventati rispettivamente 2 e 1, come rappresentato in figura 4.6. La trasformazione in numero è stata fatta esclusivamente per dare i dati in input al software.

	G	H	I
1	Area Verde	Area Verde	AreaVerdeNume
2		No	1
3		No	1
4		No	1
5	Area Verde	Si	=SE(H5="Si";2;"1")
6		No	1
7		No	1
8		No	1
9		No	1
10	Area Verde	Si	2
11		No	1
12		No	1

Figura 4.6 – Formula che assegna il numero all'area verde

L'attributo "Data" inizialmente compariva come una data nel formato gg/mm/aaaa, ad esempio 24/01/2018. È stato necessario splittarla con Excel in tre celle diverse: giorno, mese, anno, come si vede in figura 4.7.

	J	K	L	M
1	Data	Giorno	Mese	Anno
2	28/05/2012	28	5	2012
3	04/07/2014	4	7	2014
4	09/08/2012	9	8	2012
5	26/03/2016	26	3	2016
6	28/05/2012	28	5	2012
7	18/01/2016	18	1	2016
8	19/06/2013	19	6	2013
9	25/07/2013	25	7	2013
10	26/03/2012	26	3	2012
11	01/07/2013	1	7	2013
12	01/08/2014	1	8	2014
13	06/08/2014	6	8	2014
14	01/09/2014	1	9	2014

Figura 4.7 – Splittamento della data in giorno, mese e anno

La stessa cosa è stata fatta per l'ora, che si trovava nel formato hh.mm, ad esempio 23.30, rappresentato in figura 4.8.

	R	S	T
1	Ora	Ore	Minuti
2	08.00	8	0
3	07.04	7	4
4	07.10	7	10
5	07.17	7	17
6	08.00	8	0
7	07.33	7	33
8	03.57	3	57
9	07.21	7	21
10	08.00	8	0
11	03.37	3	37
12	07.31	7	31
13	08.05	8	5
14	08.00	8	0

Figura 4.8 – Splittamento dell'ora in ore e minuti

Si è pensato successivamente di arricchire il dataset iniziale aggiungendo, in particolare, questi campi:

- Stagione, rappresentato in figura 4.9

	K	L	M	N
1	Giorno	Mese	Anno	Stagione
2	31	05	2017	=SE(O(L2="01";L2="02");"Inverno";SE(O(L2="04";L2="05");"Primavera";SE(O(L2=
3	5	02	2016	"07";L2="08");"Estate";SE(O(L2="10";L2="11");"Autunno";SE(E(L2="03";K2<21);
4	7	07	2014	"Inverno";SE(E(L2="03";K2>20);"Primavera";SE(E(L2="06";K2<21);"Primavera";
5	1	10	2014	SE(E(L2="06";K2>20);"Estate";SE(E(L2="09";K2<21);"Estate";SE(E(L2="09";K2>20)
6	24	11	2014	;"Autunno";SE(E(L2="12";K2<21);"Autunno";"Inverno")))))))))))
7	2	12	2014	Autunno
8	1	09	2014	Estate
9	13	07	2015	Estate
10	28	12	2015	Inverno
11	13	07	2015	Estate
12	23	02	2012	Inverno

Figura 4.9 – Formula che assegna il valore alla stagione

- Giorno della settimana, rappresentato in figura 4.10

	J	K	L	M	N	O
1	Data	Giorno	Mese	Anno	Stagione	Giorno Settimana
2	31/05/2017	31	05	2017	Primavera	=SE(GIORNO.SETTIMANA(J2)=1;"Domenica";SE(
3	05/02/2016	5	02	2016	Inverno	GIORNO.SETTIMANA(J2)=2;"Lunedì";SE(GIORNO.SETTIMANA(J2)=3;
4	07/07/2014	7	07	2014	Estate	"Martedì";SE(GIORNO.SETTIMANA(J2)=4;"Mercoledì";SE(
5	01/10/2014	1	10	2014	Autunno	GIORNO.SETTIMANA(J2)=5;"Giovedì";SE(GIORNO.SETTIMANA(J2)=6;
6	24/11/2014	24	11	2014	Autunno	"Venerdì";SE(E(GIORNO.SETTIMANA(J2)=7;J2>0);"Sabato";"")))))))
7	02/12/2014	2	12	2014	Autunno	Martedì
8	01/09/2014	1	09	2014	Estate	Lunedì
9	13/07/2015	13	07	2015	Estate	Lunedì
10	28/12/2015	28	12	2015	Inverno	Lunedì
11	13/07/2015	13	07	2015	Estate	Lunedì
12	23/02/2012	23	02	2012	Inverno	Giovedì
13	27/02/2012	27	02	2012	Inverno	Lunedì
14	13/03/2012	13	03	2012	Inverno	Martedì

Figura 4.10 – Formula che assegna il valore alla giorno della settimana

- Feriale/Festivo, rappresentato in figura 4.11

	O	P
1	Giorno Settimana	Feriale/Festivo
2	Mercoledì	=SE(O2="Domenica";"Festivo";"Feriale")
3	Venerdì	Feriale
4	Lunedì	Feriale
5	Mercoledì	Feriale
6	Lunedì	Feriale
7	Martedì	Feriale
8	Lunedì	Feriale
9	Lunedì	Feriale

Figura 4.11 – Formula che assegna il valore alla giorno festivo o feriale

- Fascia Oraria, rappresentato in figura 4.12

	Q	R	S	T
1	Ora	Ore	Minuti	Fascia Oraria
2	17.01	17	1	=SE(E(R2=0;S2>0);"Notte";SE(E(R2>4;R2<13);"Mattina";SE(
3	11.02	11	2	E(R2>12;R2<18);"Pomeriggio";SE(E(R2>17;R2<23);"Sera";
4	14.02	14	2	SE(O(R2>22;E(R2<5;R2>0));"Notte";""))))
5	17.11	17	11	Pomeriggio
6	12.24	12	24	Mattina
7	01.25	1	25	Notte
8	15.45	15	45	Pomeriggio
9	11.51	11	51	Mattina
10	16.58	16	58	Pomeriggio
11	20.31	20	31	Sera

Figura 4.12– Formula che assegna il valore alla fascia oraria

4.1.1 Arricchimento ulteriore con aggiunta festività

Per quanto riguarda l'approfondimento dell'arricchimento dati, si è pensato di aggiungere manualmente le festività per la determinazione più esaminata del giorno festivo/feriale. Prima il giorno festivo era stato calcolato su tutti i dati solo come "domenica", mentre tutti gli altri giorni risultavano feriali.
Le festività che sono state aggiunte sono elencate di seguito in ordine cronologico:

- 1 gennaio (Capodanno)
- 6 gennaio (Epifania)
- Pasqua
- Pasquetta
- 25 aprile (Anniversario della Liberazione)
- 1 maggio (Festa dei Lavoratori)
- 2 giugno (Festa della Repubblica)
- 24 giugno (San Giovanni, patrono della città di Torino)
- 15 agosto (Ferragosto)
- 1 novembre (Tutti i Santi)
- 8 dicembre (immacolata Concezione)
- 25 dicembre (Natale)
- 26 dicembre (Santo Stefano)

In particolare, sono state aggiunte quelle festività che non capitano di domenica e quindi nel dataset sarebbero risultate come giorno feriale. Inoltre non è stata aggiunta nessuna etichetta che identifica la festività, ma solamente cambiato da giorno feriale a giorno festivo (qualora non fosse già una domenica).
Per quanto riguarda Pasqua e Pasquetta si sono ricercate negli anni le date a cui sono corrisposte, e queste sono elencate nelle tabelle sottostanti:

- Pasqua, rappresentato in tabella 6:

2012	2013	2014	2015	2016
8 aprile	31 marzo	20 aprile	5 aprile	27 marzo

Tabella 6 – Giorno di Pasqua negli anni

- Pasquetta, rappresentato in tabella 7:

2012	2013	2014	2015	2016
9 aprile	1 aprile	21 aprile	6 aprile	28 marzo

Tabella 7 – Giorno di Pasquetta negli anni

Per quanto riguarda il giorno di Pasqua, sono stati solamente controllati i dati poiché capita sempre di domenica e quindi risultava sempre festivo; mentre per Pasquetta si è dovuto modificare da giorno feriale a giorno festivo seguendo la tabella sopra per ogni anno, come fatto con tutte le altre festività.
Oltre alla classiche feste come Natale, Pasqua e così via, si è deciso di aggiungere anche il 24 giugno, siccome, essendo tutte le segnalazioni della città di Torino, è San Giovanni, la festa patronale della città di Torino; quindi si è pensato potesse essere utile come festività nell'analisi dei dati svolta.

4.2 Tassonomia

Dopo aver compiuto questi passi, si sono quindi classificati gli attributi secondo la tassonomia di figura 4.13.

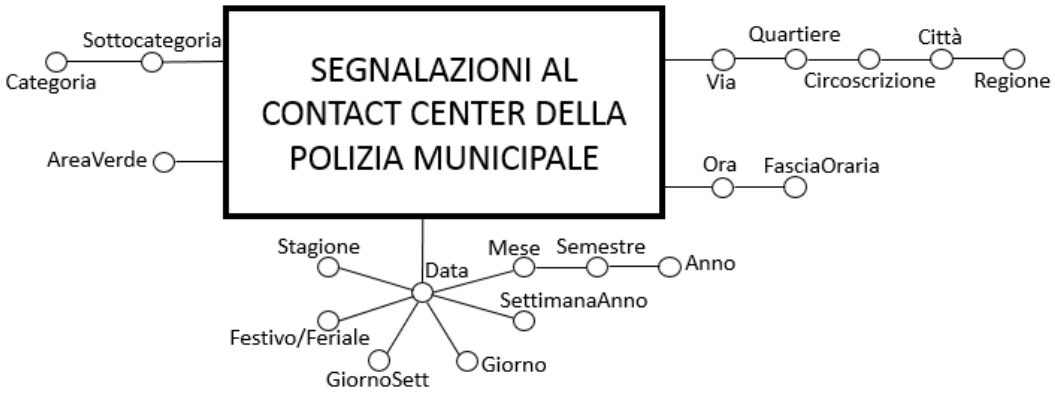


Figura 4.13– Tassonomia delle segnalazioni al contact center della polizia municipale

Si può notare che ci sono cinque rami principali e quindi cinque contenuti informativi diversi.
Il primo riguarda la categoria e la sottocategoria quindi il tipo della segnalazione dell'evento accaduto.
Il secondo ramo riguarda informazioni geografiche, ovvero la via, il quartiere, la circoscrizione, la città e infine la regione di dove è avvenuta la segnalazione; il dataset delle segnalazioni al Contact Center della Polizia Municipale è relativo alla città di Torino, quindi la città è Torino per tutti i record del dataset e la regione sarà Piemonte. Quindi il dato più generale è la circoscrizione. Per quanto riguarda il quartiere invece, in futuro se sarà possibile si potrà svolgere un'analisi nel quale per ogni via, viene associato un quartiere della città con lo scopo di svolgere un'analisi anche per zona e non solo per circoscrizione e per via (località) come è stato fatto in questo elaborato.
Il terzo grande ramo riguarda l'informazione temporale quindi principalmente la data. Dalla data si è risaliti al giorno, mese e anno, al giorno della settimana (lunedì, martedì, mercoledì, ecc), al giorno festivo o feriale (festivo inteso come domenica oppure festività

nazionali), alla stagione. In un eventuale sviluppo futuro potrebbe essere interessante anche considerare i semestri e le settimane dell'anno (dalla prima alla cinquantaduesima). Inoltre sono stati aggiunti, come se ne parlerà nel prossimo paragrafo, i giorni festivi come Natale, Pasqua, ecc.
Il quarto ramo riguarda anch'esso l'informazione temporale ovvero è caratterizzato dall'ora della segnalazione. Con l'ora si è risaliti alla fascia oraria, caratterizzata da quattro periodi del giorno: mattino, pomeriggio, sera, notte.
Infine, nell'ultimo ramo, si può trovare l'area verde, che è una caratterizzazione aggiuntiva della localizzazione geografica del luogo della segnalazione. Questo è semplicemente un dato addizionale che indica semplicemente se il posto dove è accaduto l'evento (quindi la segnalazione) è un'area verde oppure no.

4.3 Divisione del dataset

È stata effettuata una divisione del dataset originale arricchito con i dati menzionati nei precedenti paragrafi, per poi darlo in input al software utilizzato, Rapidminer.
La suddivisione è stata fatta per anni, in particolare si è deciso che il dataset intero ricopriva una fascia annuale dal 2012 al 2017, specificatamente da gennaio 2012 a giugno 2017. Si è deciso di svolgere l'analisi su interi anni, quindi è stato eliminato il semestre da gennaio a giugno 2017, per la mancanza di dati riguardanti l'ultimo semestre del 2017 (luglio-dicembre). Quindi l'analisi dei dati è stata svolta considerando gli anni (interi) dal 2012 al 2016 compresi. Addizionalmente, è stato dato in input al software Rapidminer, sia il dataset intero (anni dal 2012 al 2016), sia i vari spezzi del dataset intero. Infatti il dataset intero è stato suddiviso in altri cinque dataset più piccoli, ognuno rispettivamente per gli anni 2012, 2013, 2014, 2015 e 2016.
La suddivisione del dataset è stata svolta per confrontare cosa venisse fuori con un'estrazione delle regole di associazione sull'intero dataset e paragonarla con un'estrazione delle regole di associazione su ogni piccolo pezzo del dataset (diviso per anno), per vedere se escono gli stessi risultati. Questo, tuttavia, verrà affrontato successivamente nel capitolo sull'estrazione delle regole di associazione tramite Rapidminer con l'algoritmo FP-Growth.

4.4 Costruzione del processo con Rapidminer

In questo paragrafo è descritto come il dataset è stato dato in input al software Rapidminer e come si è svolto il processo passo per passo per arrivare ad ottenere dei risultati.
Innanzitutto, come spiegato precedentemente, il dataset è stato suddiviso in cinque piccoli dataset però mantenendo anche il dataset intero, quindi sono stati dati in input al software sei dataset totali.
In questo paragrafo gli esempi fatti si avvarranno al dataset intero poiché funzionerà allo stesso modo con i dataset più piccoli (ad eccezione dell'attributo anno). Infatti le operazioni fatte con il dataset intero sono state ripetute per tutti i dataset minori.

In principio, il dataset è stato fatto leggere dal software Rapidminer attraverso l'operatore "ReadExcel" in formato foglio di calcolo Excel (il processo intero è rappresentato in figura 4.14).

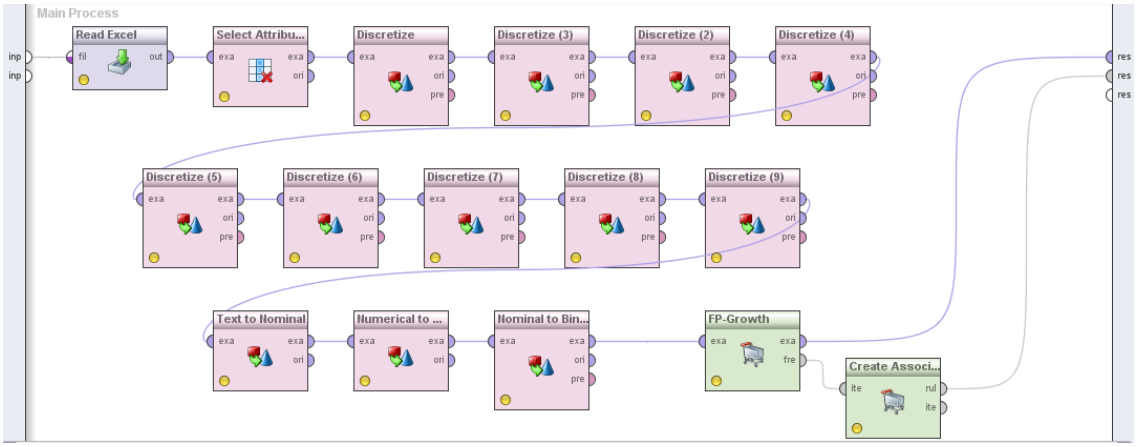


Figura 4.14 – Processo creato con il software Rapidminer

Dopo quest'operatore è stato applicato l'operatore "Select Attributes" dove sono stati selezionati gli attributi di interesse. In questo elaborato si sono considerati quasi tutti gli attributi come attributi di interesse poiché si sono volute vedere le maggiori correlazioni tra essi. Successivamente sono stati applicati gli operatori "Discretize by user specification" non con lo scopo di discretizzare ma per fare un mapping, siccome gli attributi erano quasi tutti di tipo categorico e si aveva bisogno di attributi numerici semplicemente per ottenere una migliore visualizzazione dei risultati.
La maggior parte degli attributi sono stati rimappati, seguendo lo schema spiegato precedentemente riguardo le formule di Excel, e specificatamente sono:

- Categoria, rappresentato in figura 4.15

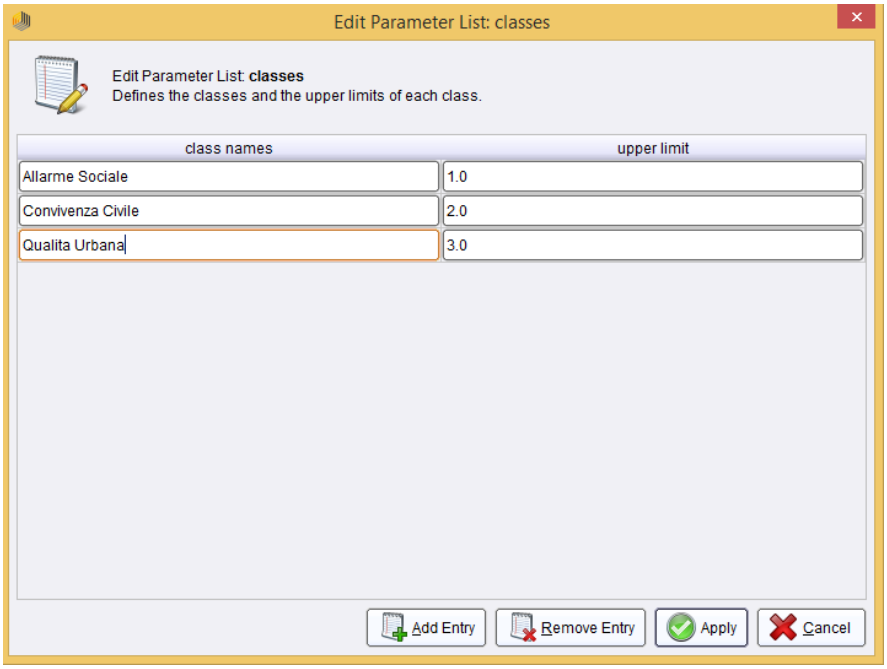


Figura 4.15- Ri-mapping della categoria

- Sottocategoria, rappresentato in figura 4.16

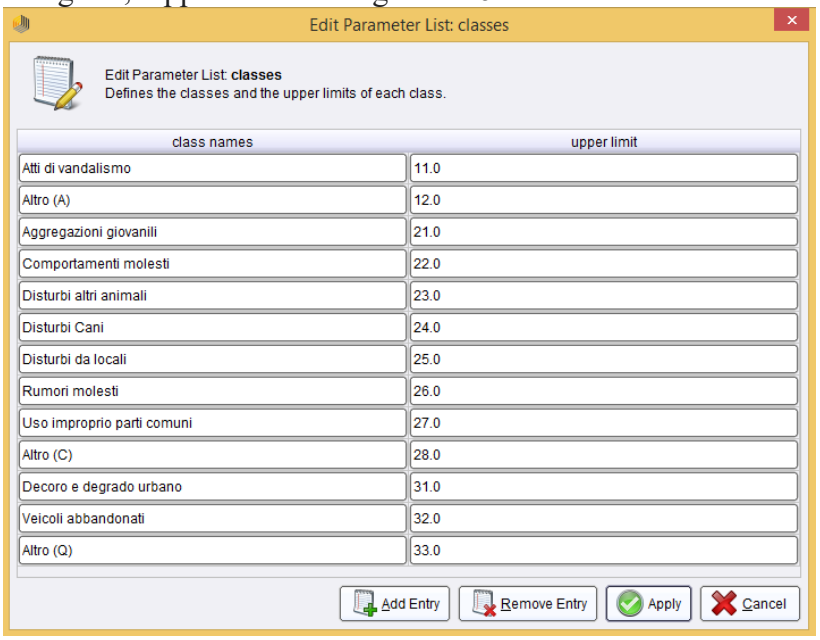


Figura 4.16 – Ri-mapping della sottocategoria

- Circoscrizione, rappresentato in figura 4.17

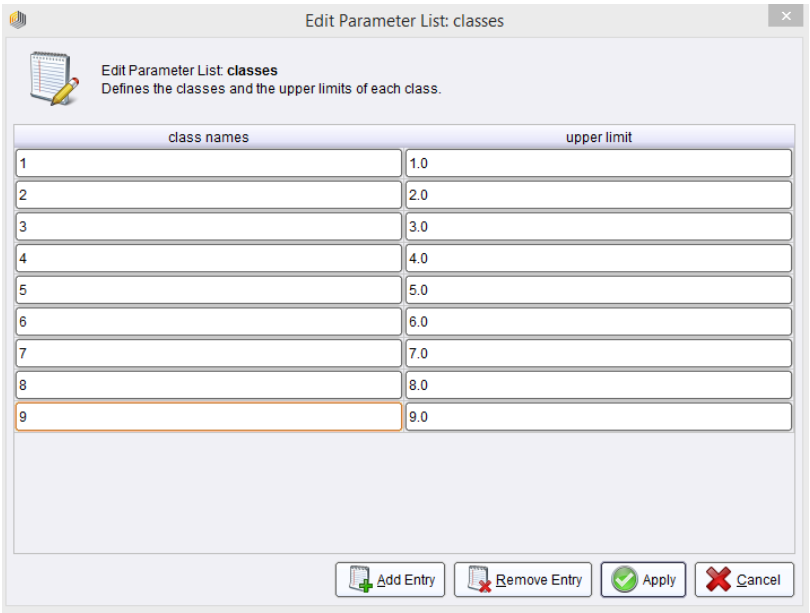


Figura 4.17 – Ri-mapping della circoscrizione

- Giorno, rappresentato in figura 4.18

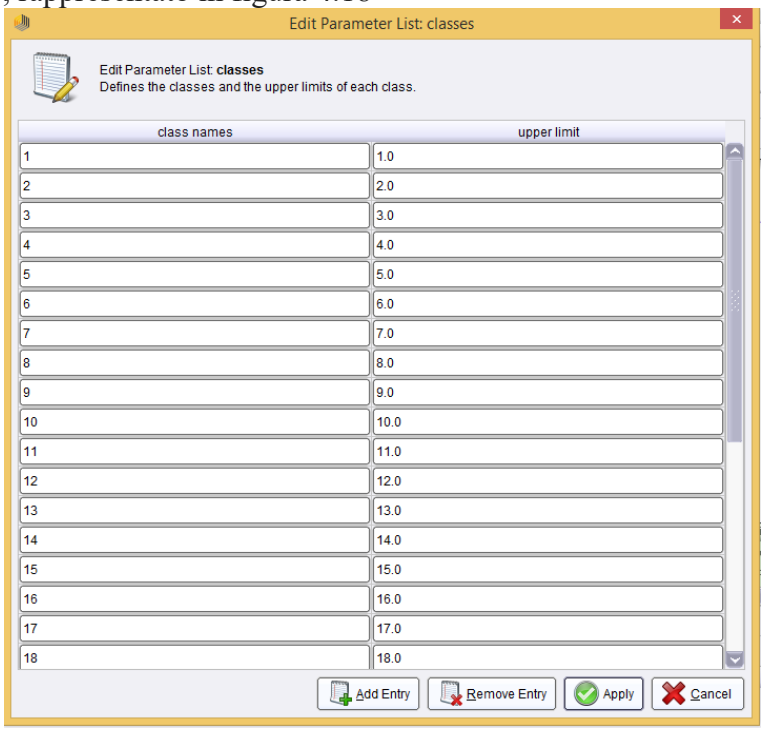


Figura 4.18 – Ri-mapping del giorno

- Mese, rappresentato in figura 4.19

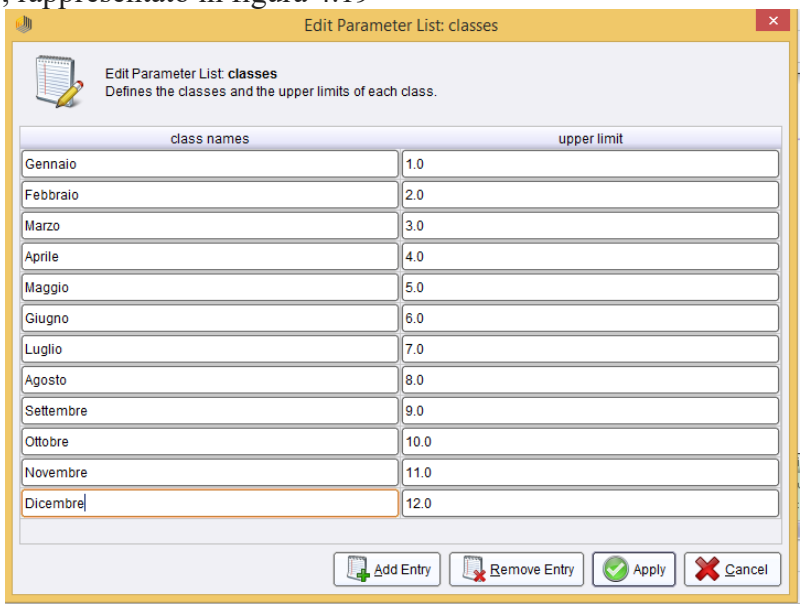


Figura 4.19 – Ri-mapping del mese

- Anno (solo per il dataset intero, perché le piccole porzioni sono già suddivise per anno), rappresentato in figura 4.20

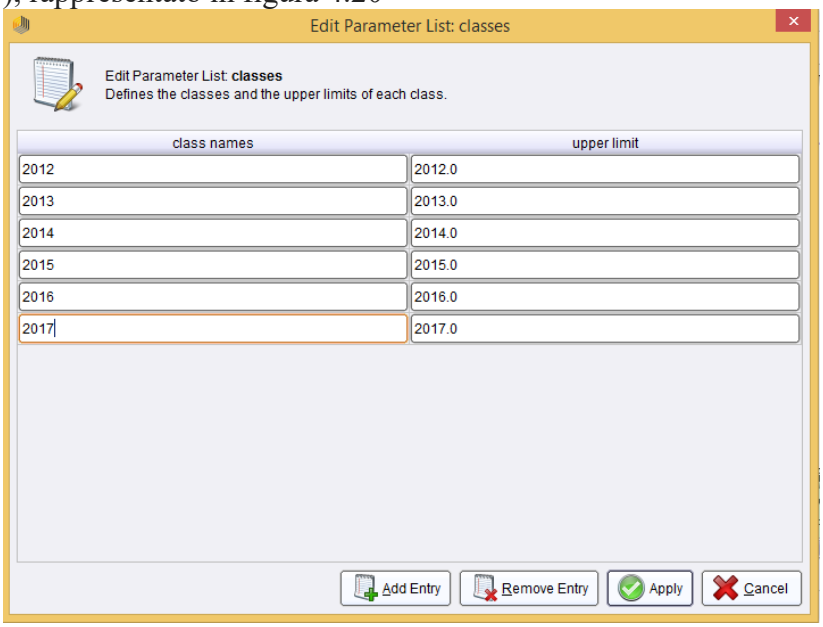


Figura 4.20 – Ri-mapping dell'anno

- Area Verde, rappresentato in figura 4.21

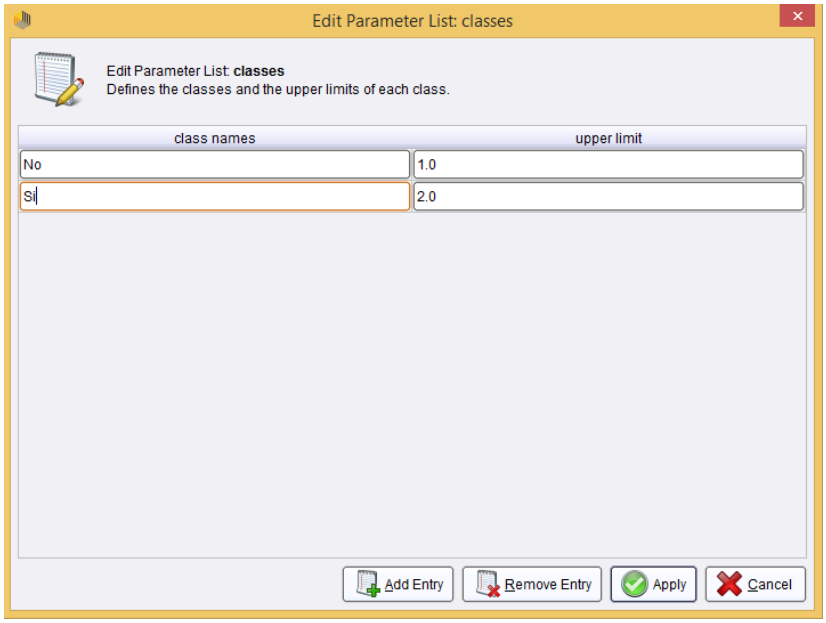


Figura 4.21 – Ri-mapping dell'area verde

- Giorno Feriale o Festivo (comprese le festività), rappresentato in figura 4.22

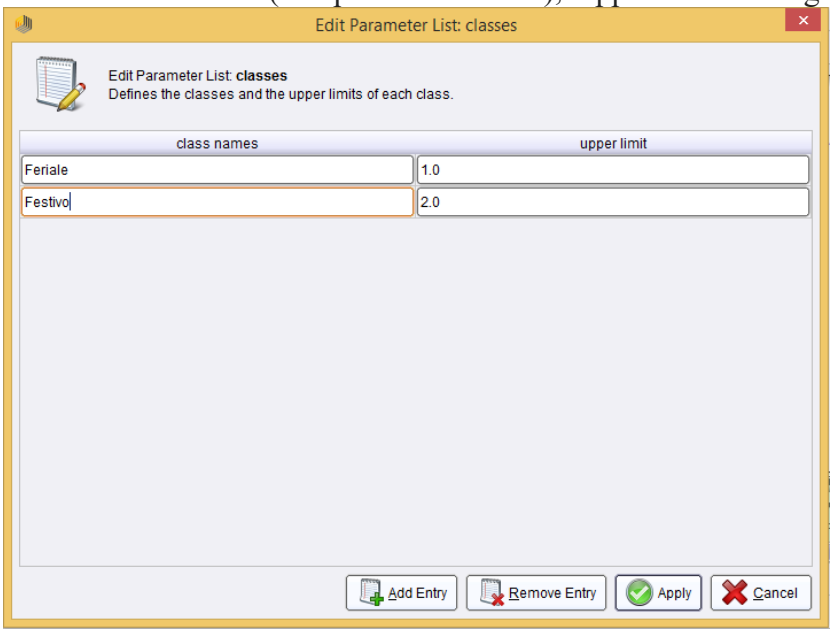


Figura 4.22 – Ri-mapping del giorno feriale o festivo

- Fascia oraria, rappresentato in figura 4.23

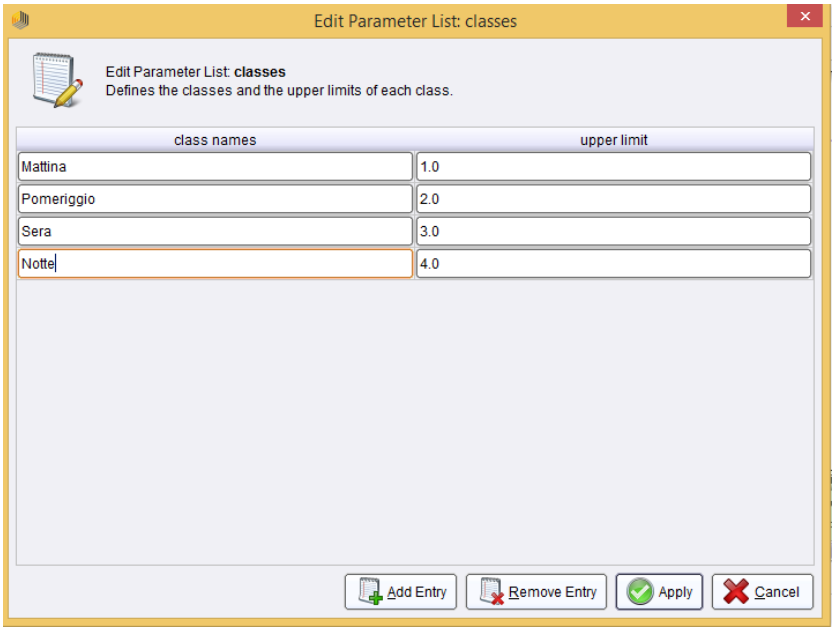


Figura 4.23 – Ri-mapping della fascia oraria

Si è utilizzato l'operatore "Discretize by user specification" tante volte quanti gli attributi elencati nel precedente elenco puntato. Ogni operatore è stato aggiunto o tolto a seconda se l'analisi lo riguardava ovvero se quell'attributo veniva selezionato nell'operatore precedente "Select Attributes".
Successivamente sono stati aggiunti tre operatori: "Text to Nominal", "Numerical to Binomial", "Nominal to Binominal", blocchi necessari per effettuare il run con l'algoritmo FP-Growth, come menzionato nel capitolo riguardante il processo di Rapidminer. Infatti il successivo passo è stato aggiungere l'operatore FP-Growth, che è stato "collegato" all'output del processo e "collegato" a sua volta all'operatore "Create Association Rule". Per questi due operatori sono stati assegnati due parametri fondamentali per l'analisi: per FP-Growth (figura 4.24) è stato necessario assegnare un supporto minimo da utilizzare per l'estrazione delle regole di associazione (figura 4.25), e il minimo supporto utilizzato è stato l'1%.

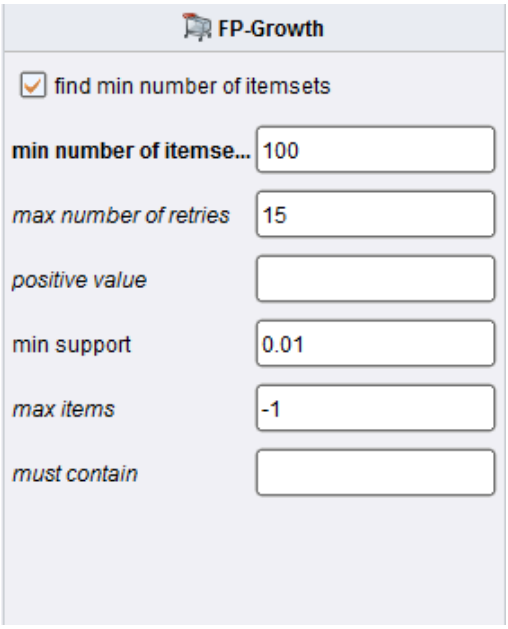


Figura 4.24 – Parametri dell'operatore FP-Growth

La scelta dell'1% è stata fatta perché così l'algoritmo sarebbe stato in grado di trovare più regole e poi si sarebbero selezionate solo successivamente quelle di maggiore importanza. Invece, se si avesse scelto un minimo supporto più alto, l'algoritmo avrebbe potuto eliminare regole di qualità alta che magari sarebbero risultate utili per l'analisi.
Per quanto riguarda invece l'ultimo operatore, si doveva assegnare un livello di minima confidenza; questo livello, in questa analisi, è stato posto al 10%, come in figura 4.25.

Figura 4.25 – Parametri dell'operatore Create association rules

La spiegazione è simile alla precedente riguardante il minimo supporto: filtrando con un livello molto basso, poiché il 10% è da considerarsi molto basso, ci permette di ottenere molte più regole rispetto ad un valore più alto, e quindi anche di non eliminare regole di associazione che potrebbero essere fondamentali per la lettura dei risultati.
Dopo aver impostato questi parametri generali, si è proceduto con il run dell'algoritmo e si è aspettato che uscissero i risultati ovvero le regole di associazione. Queste regole sono poi state catalogate in diversi fogli di calcolo (Excel), suddivisi per anni e per tipi di regole.
Nel capitolo successivo, verrà spiegata più approfonditamente questa suddivisione dei risultati.

5. Analisi delle regole di associazione

Questo capitolo è dedicato alla lettura e all'interpretazione dei risultati ottenuti con il processo citato precedentemente e si sono citate le regole ritenute di fondamentale importanza. Sono state poste diverse domande, la principale tra tutte è stata: "dato un certo tipo di segnalazione (categoria o sottocategoria), a quale tipo di contesto è associato?" oppure "data una certa informazione di contesto (spaziale, temporale o entrambi), a quale tipo di segnalazione è associata?". Si è notato che nell'analisi, l'informazione di contesto spaziale è risultata maggiormente dalla circoscrizione più che dalla località e dall'area verde; mentre per il contesto temporale dall'anno, dal mese, dal giorno festivo o feriale, dalla stagione e dalla fascia oraria. Prima si è cercato di leggere le regole di associazione relative alla categoria per i diversi dataset in input, e successivamente si è cercato di interpretare anche i dati relativi alle sottocategorie (molti di più rispetto ai primi). Quindi questo capitolo sarà utile per leggere e filtrare le regole di associazione, e vedere se ce n'è qualcuna interessante dal punto di vista degli indici e interessante anche dal punto di vista del contenuto. Per fare un esempio con il contesto spaziale, l'obiettivo potrebbe essere: "dato un certo luogo, andare a capire cosa succede (cosa è successo) in quel luogo" oppure esplorare la regola al contrario ovvero: "data una certa segnalazione, dove potrebbe succedere". Pertanto, si è cercato di predire questi dati da dati che già si avevano. Infatti questa analisi potrà essere utile anche in futuro per la predizione di zone pericolose. Ad esempio per il quartiere di San Salvario a Torino, magari possono risultare delle regole di associazione più rilevanti e potrebbe essere un risultato da tenere in considerazione.

5.1 Tipologia di estrazione delle regole di associazione generate

Dopo aver fatto girare il software con dataset diversi in input, si sono estratte tutte le regole di associazione generate dall'algoritmo FP-Growth. Queste sono state catalogate in file Excel, ogni file comprensivo di sei fogli ciascuno. Per ogni dataset sono stati creati due file Excel: uno relativo all'analisi delle regole riguardanti la categoria (sei fogli), e l'altro riguardanti la sottocategoria (altri sei fogli).
Siccome nelle regole di associazione abbiamo una premessa che implica una conclusione (*Premessa* → *Conclusione*), si è notato che si potevano dividere gli attributi con informazione spaziale e temporale.
Secondo la tassonomia riportata precedentemente, si è visto che gli attributi risultanti nelle regole fossero soprattutto di due tipi: spaziale e temporale. Si è assunta, quindi, questa ipotesi, pensando di mantenere solo le regole che avessero come premessa o come conclusione il tipo di segnalazione (quindi la categoria o la sottocategoria).
Come detto prima, per ogni dataset si avranno:
Regole filtrate per categoria (come premessa e separatamente come conclusione)
Regole filtrate per sottocategoria (come premessa e separatamente come conclusione)

Per ogni file con i risultati, si avranno sei fogli che sono strutturati come segue.
Per il file riguardante i risultati filtrati per categoria:
Categoria → Luogo
Categoria → Tempo
Categoria → Luogo, Tempo
Luogo → Categoria
Tempo → Categoria
Luogo, Tempo → Categoria
Per il file riguardante i risultati filtrati per sottocategoria:
Sottocategoria → Luogo
Sottocategoria → Tempo
Sottocategoria → Luogo, Tempo
Luogo → Sottocategoria
Tempo → Sottocategoria
Luogo, Tempo → Sottocategoria
Quindi, ad esempio per la categoria, si sono ricercate tutte le regole di associazione dove la categoria era una premessa e si sono catalogate le conclusioni secondo le informazioni temporali e spaziali. Poi, si sono ricercate le regole al contrario, ovvero dove la categoria era la conclusione e si sono cercate le premesse. Tutto ciò è stato svolto parallelamente con la sottocategoria.
Questa suddivisione è stata fatta soprattutto per riuscire a trovare una correlazione tra il tipo di categoria (ed eventualmente la sottocategoria) e l'informazione spaziale, temporale ed entrambi.
In questa analisi, però, non evince come risultato quasi mai la località (la via della segnalazione) per quanto riguarda l'analisi spaziale. Molte regole trovate sono considerate superflue poiché, soprattutto in quelle temporali; ogni volta che nella regola compare un certo mese, esso sarà appartenente ad una ben chiara stagione, allora comparirà anche la regola relativa alla stagione. Per fare un esempio, il mese di luglio comparirà (nelle premesse o nelle conclusioni) con la stagione estate (rispettivamente nelle premesse o nelle conclusioni) e questo perché è un dato di fatto che il mese di luglio appartiene alla stagione estate, quindi ci sarà una correlazione forte. Si è cercato di non dare troppo peso a queste regole che risultano scontate e si è voluti arrivare a risultati più sensati, studiati e approfonditi.
Queste regole di associazione generate dal software Rapidminer, sono state poi ulteriormente filtrate in base al lift. Si sono cercate le regole con un lift maggiore di uno (uno escluso) e sono state eliminate tutte le regole con lift minore o uguale a uno. Questo perché in questo elaborato si è assunto che le correlazioni dovessero essere tutte positive per avere un senso, e quindi non sono state valutate interessanti le correlazioni negative.
Successivamente si sono tenuti in considerazione il supporto e la confidenza e i risultati sono stati ordinati prima in ordine di supporto (dal maggiore al minore) e successivamente in ordine di confidenza (dal maggiore al minore) poiché più grande è il supporto e più grande è la confidenza e migliore è la regola di associazione.

Questo procedimento è stato svolto ugualmente sia per i risultati generati dal dataset intero (in input), sia per quelli generati dai dataset più piccoli suddivisi per anno ed è stata svolta la catalogazione in categoria e sottocategoria e per il tipo di regola (come spiegato in questo paragrafo).
Successivamente, sarà svolto un lavoro di approfondimento nella lettura e interpretazione dei risultati ottenuti con questo processo segnalando le regole di associazione ritenute di fondamentale importanza.

5.2 Analisi del dataset utilizzando la categoria come conclusione o premessa

Da qui in poi, si inizia la lettura e l'interpretazione dei risultati. Il primo punto di partenza potrebbero essere i risultati relativi alla categoria, poiché se ne hanno molti meno rispetto ai risultati relativi a ciascuna sottocategoria. Inoltre, essendo pochi, potrebbero essere interpretabili in maniera migliore da parte di una mente umana.
Tutti i grafici contengono sull'asse delle ascisse la premessa oppure la conclusione della regola di associazione (dipende dalla tipologia di regola che si sta analizzando), mentre sull'asse delle ordinate il supporto. I blocchi colorati sono ordinati in base a confidenza decrescente, ovvero il primo da sinistra è quello con il maggiore livello di confidenza e l'ultimo a destra è quello con il minore livello di confidenza.
Analizzando per esempio le regole di associazione generate dal processo con in input il dataset intero, quindi parlando di anni dal 2012 al 2016 compresi, e seguendo la tipologia di regola *Categoria → Contesto Spaziale*, si può notare un supporto che va da 0,0465 a 0,4988 delle regole estratte e un livello di confidenza dal 10,43% fino al 96,67%. Pertanto con la tipologia di regola *Categoria → Contesto Spaziale* per il dataset generale (tutti gli anni), la categoria che spicca maggiormente è la Convivenza Civile, figura 5.1.

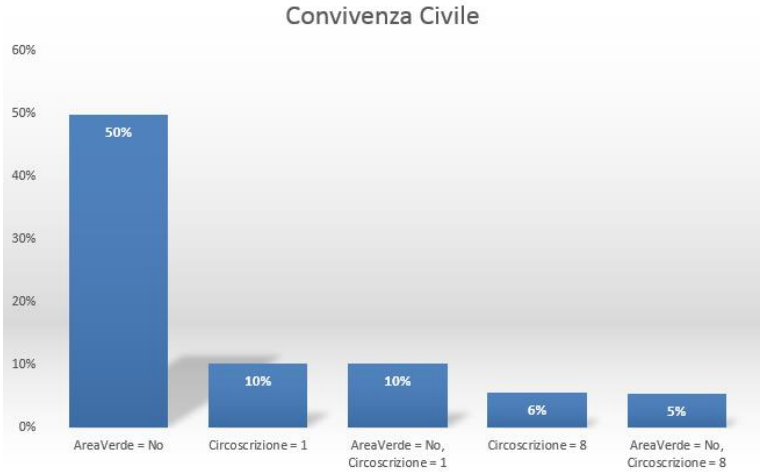


Figura 5.1 – Grafico relativo alla tipologia di regola Categoria →Contesto Spaziale per la categoria convivenza civile

Infatti le tre regole che hanno sia il maggior supporto che una confidenza alquanto alta sono:
Convivenza civile → Non area verde

Convivenza civile → Circoscrizione 1
Convivenza civile → Non area verde, Circoscrizione 1.
Nell'interpretazione si può affermare che data la categoria convivenza civile, i posti dove potrebbe aver luogo maggiormente questo tipo di segnalazione sono la circoscrizione 1 e dove non vi siano aree verdi. Andando ad esplorare la regola al contrario, Contesto spaziale che implica la Categoria, si hanno gli stessi risultati per il supporto, mentre per quanto riguarda il livello di confidenza si ha un livello minore rispetto a quello precedente.

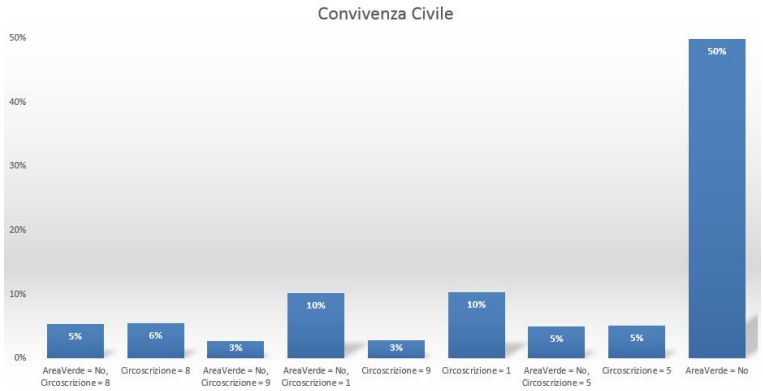


Figura 5.2 – Grafico relativo alla tipologia di regola Contesto Spaziale →Categoria per la categoria convivenza civile

Infatti, per la regola *Convivenza civile → Non area verde* si aveva una confidenza di ben 96,67%, mentre per la regola di associazione al contrario *Non area verde → Convivenza civile* la confidenza è del 52,39%, ovvero molto più bassa, figura 5.2.
In figura 5.3 il grafico relativo alla categoria qualità urbana.

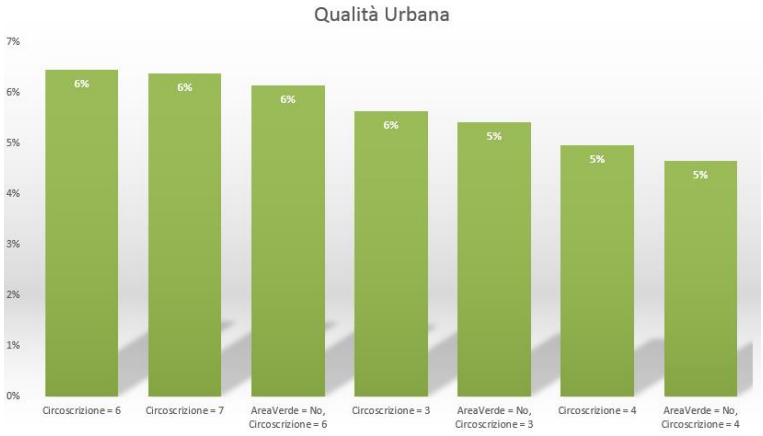


Figura 5.3 – Grafico relativo alla tipologia di regola Categoria →Contesto Spaziale per la categoria qualità urbana

Considerando la regola nel verso opposto (*Spazio → Categoria*), figura 5.4, infatti la regola descritta precedentemente non è tra le migliori trovabili; quella che spicca per quanto riguarda la confidenza è la *Non area verde, Circoscrizione 8 → Convivenza Civile* con un livello di confidenza del 62,96%. La seconda, sempre ordinando per livello di

confidenza, con 62,43% è data da *Circoscrizione 8 → Convivenza Civile*. La terza regola, invece, sempre ordinata per confidenza, *Area verde → Qualità Urbana.*

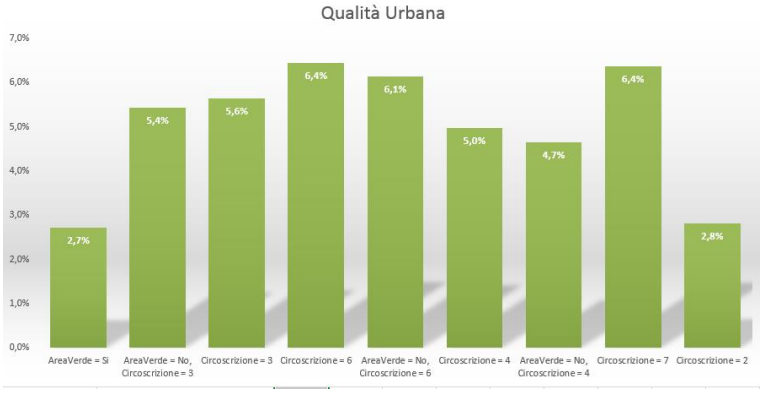


Figura 5.4 – Grafico relativo alla tipologia di regola Contesto Spaziale → Categoria per la categoria qualità urbana

Si può notare che sia nella regola di tipo *Categoria → Luogo* che nella regola al contrario, non compare quasi mai una delle tre categorie principali, l'Allarme Sociale; per lo meno non compare facendo girare il processo con i parametri decisi per questo progetto, ovvero settando un supporto minimo dell'1% e una minima confidenza del 10%.
Per quanto riguarda il tipo di regola *Categoria → Contesto Temporale*, suddividendo e analizzando per categoria si ottiene che, per la categoria convivenza civile, gli anni che spiccano nell'analisi sono l'anno 2013 e l'anno 2015. Ciò significa che, data una segnalazione di tipo convivenza civile, gli anni in cui questa categoria si è maggiormente verificata sono il 2013 e il 2015.

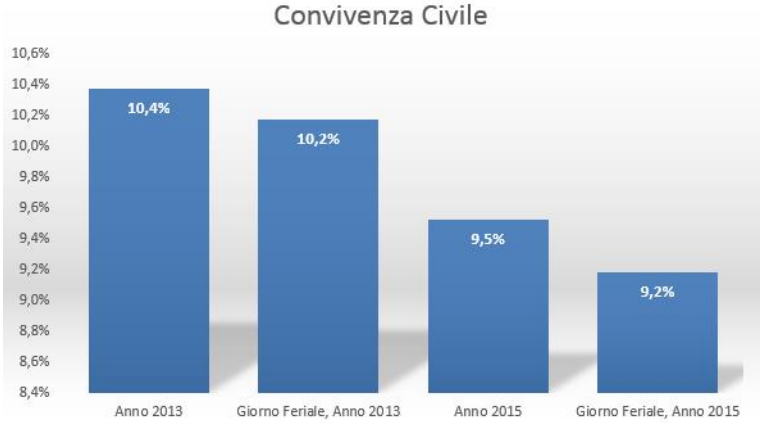


Figura 5.5 – Grafico relativo alla tipologia di regola Categoria →Contesto Temporale per la categoria convivenza civile

Come detto precedentemente, tutti i grafici contengono sull'asse delle ascisse la premessa oppure la conclusione della regola di associazione (dipende dalla tipologia di regola che si sta analizzando), mentre sull'asse delle ordinate il supporto. I blocchi colorati sono ordinati in base a confidenza decrescente, ovvero il primo da sinistra è quello con il maggiore livello di confidenza e l'ultimo a destra è quello con il minore livello di

confidenza. Quindi per fare un esempio nel grafico 5.5, la prima barra avrà la confidenza maggiore rispetto a tutte le altre, ed un supporto anch'esso maggiore rispetto a tutte le altre (poiché è la più alta). Per quanto riguarda la convivenza civile, il range delle confidenze riscontrate va dal 17% (per l'ultimo blocco) al 21% (per il primo blocco). Successivamente verrà fatta un'analisi più approfondita per quanto riguarda la domanda "dato un contesto spaziale, quale segnalazione è associata?" dove i livelli di confidenza saranno specificati nel grafico.
D'ora in avanti la categoria convivenza civile sarà contrassegnata dal colore blu, come è evidente dal grafico in figura 5.5.
Passando alla categoria qualità urbana, in figura 5.6, che d'ora in avanti sarà contrassegnata dal colore verde, si può notare che, data una segnalazione di tipo qualità urbana, gli anni in cui maggiormente questa segnalazione si è verificata sono il 2012 e il 2014. Il range di confidenza va dal 20%al 24%, leggermente più alto rispetto alla categoria analizzata precedentemente.

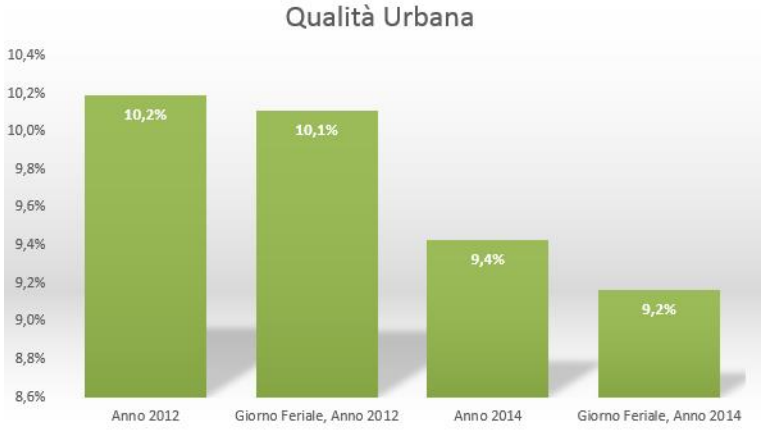


Figura 5.6 – Grafico relativo alla tipologia di regola Categoria →Contesto Temporale per la categoria qualità urbana

Per quanto riguarda invece la terza categoria, allarme sociale, in figura 5.7, si ha che data una segnalazione di tipo allarme sociale, il periodo in cui questa segnalazione si è maggiormente verificata è nell'anno 2014. Questa categoria d'ora in avanti sarà identificata dal colore rosso. Questa categoria, come si può notare dal grafico in figura 5.7, ha un supporto molto più basso rispetto alle altre due categorie, ma ha un range di confidenza che va dal 25% al 26%.

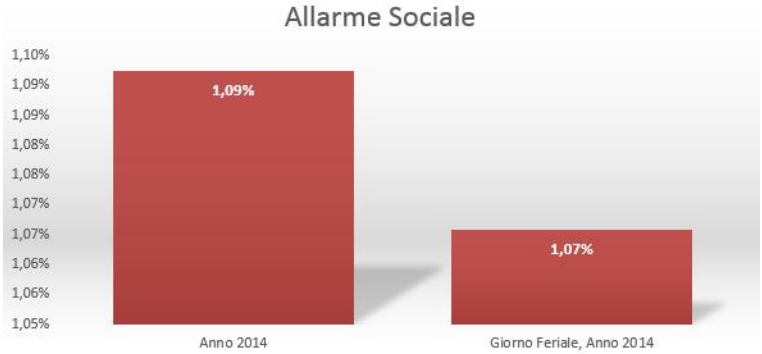


Figura 5.7 – Grafico relativo alla tipologia di regola Categoria →Contesto Temporale per la categoria allarme sociale

Per quanto riguarda questi primi tre grafici descritti, si può concludere che il tipo di regola di associazione *Categoria → Contesto Temporale*, non ha livelli di confidenza altissimi, ma si vuole sottolineare che come contesto temporale si è filtrato solamente per anno.
Tra tutte le regole identificate in un contesto temporale, si può notare che sono due quelle che risaltano maggiormente per supporto, e sono:
Convivenza Civile → Estate con un supporto di 0,1461 e una confidenza del 28,31%
Convivenza Civile → Estate, Giorno Feriale con un supporto di 0,1421 e una confidenza del 27,54%
Per quanto riguarda la confidenza, risalta in particolare un valore altissimo, 99,17%, per la regola di associazione *Allarme Sociale → Giorno Feriale* che ha un supporto di 0,0423.
Prendendo in considerazione, invece, la tipologia di regola opposta ovvero *Contesto Temporale → Categoria*, si può notare che nella lettura dei risultati, la categoria allarme sociale scompare, probabilmente perché ha un supporto minore dell'1% e una confidenza minore del 10% (parametri stabiliti per la selezione delle regole di associazione in output). Quindi, non sono state possibili rilevazioni dei risultati per quella categoria.
Analizzando la convivenza civile, figura 5.8, è stata fatta un'analisi sia sull'anno che sul mese, poiché i risultati generati per questa tipologia sono moltissimi. Come si può notare dal grafico in figura 5.8, risaltano il mese di Giugno dell'anno 2012 e Luglio 2012 con supporti e confidenze relativamente alti. Nel grafico di figura 5.8, il range di confidenza va dal 52% e 62%, quindi l'indicazione di quanto spesso la regola è stata trovata vera è sopra il 50%, risultato molto buono.

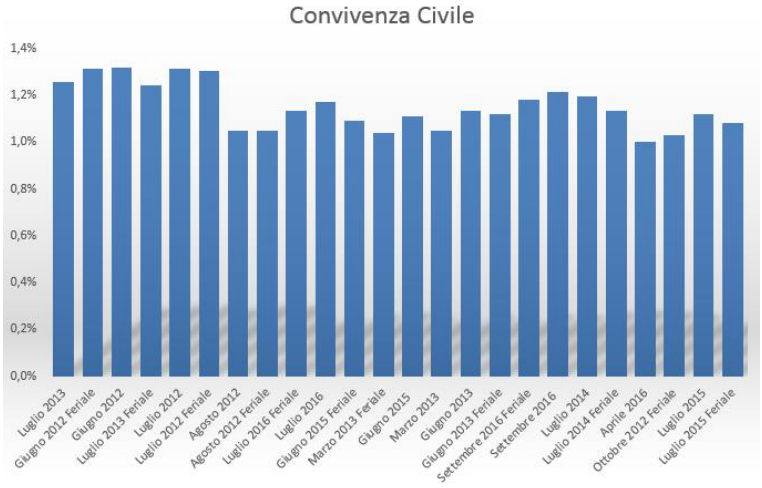


Figura 5.8 – Grafico relativo alla tipologia di regola Contesto Temporale → Categoria per la categoria convivenza civile

Per quanto concerne la qualità urbana, figura 5.9, a livello di supporto risalta il mese di Marzo 2012 con un buon livello di confidenza. Il range di confidenza del grafico in figura 5.9 va dal 44% al 59%, quindi in media un buon 50%.
Come si può notare il mese di Marzo è un'indicazione in più che si ha dell'anno 2012. Infatti questa è la regola di associazione opposta (*Contesto Temporale → Categoria*) della regola che si è analizzata precedentemente (*Categoria → Contesto Temporale*), nel quale si era detto che l'anno che risalta era il 2012. Quindi si può assumere, facendo un controllo incrociato, che l'analisi effettuata torna.

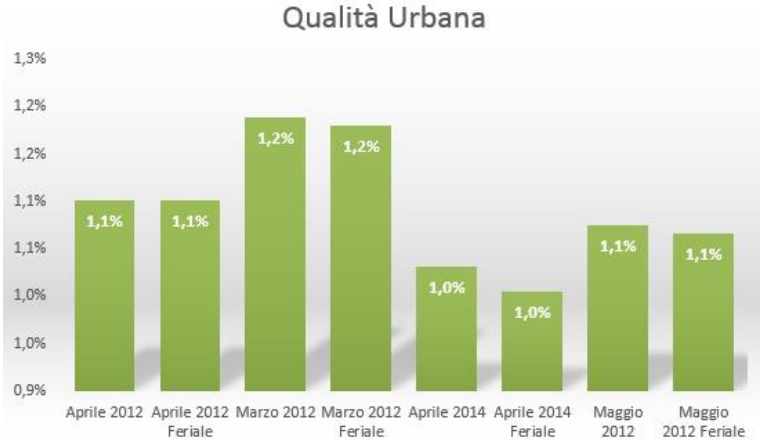


Figura 5.9 – Grafico relativo alla tipologia di regola Contesto Temporale → Categoria per la categoria qualità urbana

A proposito invece della convivenza civile, concludendo si può dire che, secondo la regola al contrario, risalta l'anno 2012, mentre nell'analisi svolta precedentemente risaltavano il 2013 e il 2015.
Considerando la totalità delle regole, quindi non solo quelle relative agli anni e ai mesi, le regole con maggiore supporto sono:

Estate → *Convivenza Civile* con un supporto di 0,1461 e una confidenza del 56,72%
Estate, Giorno Feriale → *Convivenza Civile* con un supporto di 0,1461 e una confidenza del 56,60%.
Invece tra le regole con una più alta confidenza, risaltano:
Anno 2015, Fascia Oraria Mattina → *Convivenza Civile* con una confidenza del 66,29%
Anno 2015, Fascia Oraria Mattina, Giorno Feriale → *Convivenza Civile* con una confidenza del 65,91%.
Successivamente si è esplorata la tipologia di regola di associazione che contiene entrambi contesto temporale e contesto spaziale.
Come prima cosa si è vista la tipologia di regola *Categoria* → *Contesto spaziale e temporale.* Se si suddivide per categoria, si può notare che per la categoria allarme sociale, figura 5.10, risalta nell'anno 2014, nelle non aree verdi e nei giorni feriali con un range di confidenza che va dal 23% al 25%.

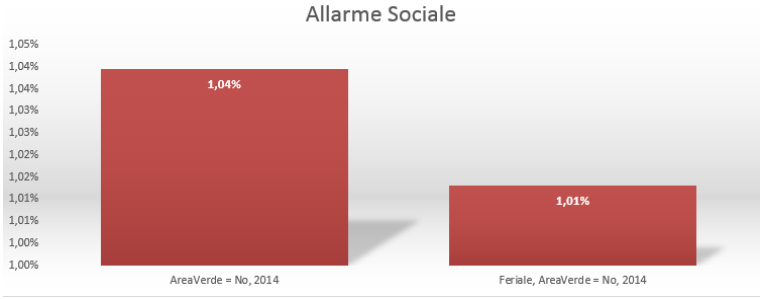


Figura 5.10 – Grafico relativo alla tipologia di regola Categoria → Contesto spaziale e temporale per allarme sociale

A proposito, invece, della convivenza civile, figura 5.11, risaltano gli anni 2013, 2015 e 2016 con una confidenza che va dal 17% al 20%.

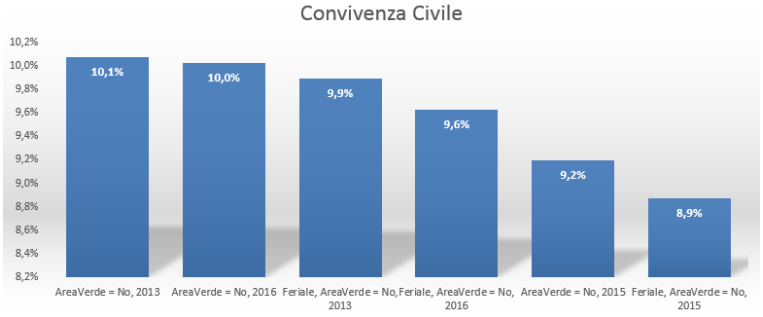


Figura 5.11 – Grafico relativo alla tipologia di regola Categoria → Contesto spaziale e temporale per convivenza civile

Per quanto riguarda la qualità urbana, figura 5.12, risaltano l'anno 2012 e il 2014 con una confidenza che va dal 19% al 23%.

Figura 5.12 – Grafico relativo alla tipologia di regola Categoria → Contesto spaziale e temporale per qualità urbana

Dagli ultimi tre grafici proposti, figure 5.10, 5.11 e 5.12, si può notare che oltre alla suddivisione per anni effettuata in quest'analisi, compare molto spesso l'attributo area verde. Infatti nella maggior parte delle regole analizzate, l'attributo area verde assume valore no.

Per quanto riguarda invece la generalità delle regole, analizzando le regole con una conclusione mista ovvero sia appartenente al contesto spaziale che a quello temporale, risaltano:

Convivenza Civile → Giorno Feriale, Non area verde con supporto 0,4856 e una confidenza del 94,11%.

Questa regola di associazione ha il maggior supporto tra tutte quelle di questa tipologia e la maggior confidenza tra tutte quelle di questa tipologia.

Anche qui, si va poi a vedere la tipologia di regola di associazione opposta, ovvero *Contesto spaziale e temporale → Categoria*, e si sono in particolare prese in considerazione gli attributi anno, per quanto riguarda il contesto temporale, e circoscrizione, per quanto riguarda i contesto spaziale. Tutto ciò per cercare di fare un'analisi più incentrata su alcuni attributi come sarà poi fatta successivamente.

Le due categorie che spiccano maggiormente sono la convivenza civile e la qualità urbana, mentre neanche qui compaiono regole in cui è presente la categoria allarme sociale, probabilmente sempre perché ha un supporto e una confidenza minore di quelli minimi.

Guardando il grafico in figura 5.13, riguardante la convivenza civile, risalta nell'anno 2013 la circoscrizione 1, sempre la circoscrizione 1 nell'anno 2012 e anche nel 2014 e 2016. Il range di confidenza va dal 52% al 69%. Si può concludere che data la circoscrizione 1 nei diversi anni, 2012, 2013, 2014, 2016, la maggior parte delle segnalazioni sono di tipo convivenza civile.

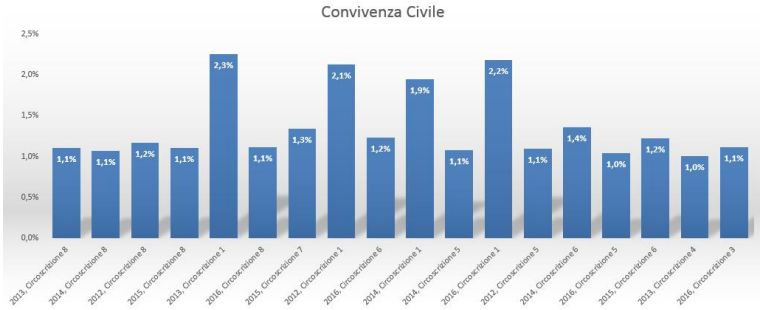


Figura 5.13 – Grafico relativo alla tipologia di regola Contesto spaziale e temporale → Categoria per convivenza civile

Invece, a proposito di qualità urbana, risalta la circoscrizione 6 nell'anno 2012, perciò data la circoscrizione 6 nell'anno 2012, la maggior parte delle segnalazioni sono di tipo qualità urbana. Il grafico in figura 5.14, ha un range di confidenza dal 45% al 60%.

Figura 5.14 – Grafico relativo alla tipologia di regola Contesto spaziale e temporale → Categoria per qualità urbana

Considerando la totalità delle regole, analizzando quelle con una premessa mista ovvero sia appartenente al contesto spaziale che a quello temporale, risaltano:
Giorno Feriale, Non area verde → Convivenza Civile con un supporto di 0,4856 e una confidenza del 52,30%.
Ordinando, invece, per confidenza, si possono notare ben quattro regole con una confidenza del 69% e sono:
Anno 2013, Circoscrizione 8, Giorno Feriale, Non area verde → Convivenza Civile
Anno 2013, Circoscrizione 8, Non area verde → Convivenza Civile
Anno 2013, Circoscrizione 8, Giorno Feriale → Convivenza Civile
Anno 2013, Circoscrizione 8 → Convivenza Civile.
Questa è l'analisi effettuata sulle categorie, successivamente se ne svolgerà una simile sulle sottocategorie per cercare di capire in maniera più approfondita il vero tipo della segnalazione ricevuta al Contact Center della polizia municipale.

5.3 Analisi del dataset utilizzando la sottocategoria come conclusione o premessa

Dopo aver esplorato l'analisi sulle categorie, come spiegato precedentemente si è svolta anche un'analisi sulle sottocategorie e ci si è basati nella catalogazione delle regole di associazione riguardo la categoria di appartenenza delle sottocategorie, così da poter confrontare i risultati.
Si sono nuovamente ricercate le seguenti tipologie di regole:
Sottocategoria → Contesto Spaziale
Sottocategoria → Contesto Temporale
Contesto Spaziale → Sottocategoria
Contesto Spaziale → Sottocategoria.
Andando a esplorare la prima tipologia di regola, si può notare che, in termini di supporto e di confidenza, risalta la circoscrizione 1 a proposito della sottocategoria disturbi da locali (categoria convivenza civile), come si vede dal grafico in figura 5.15. Il range di confidenza va dal 14% al 23%.

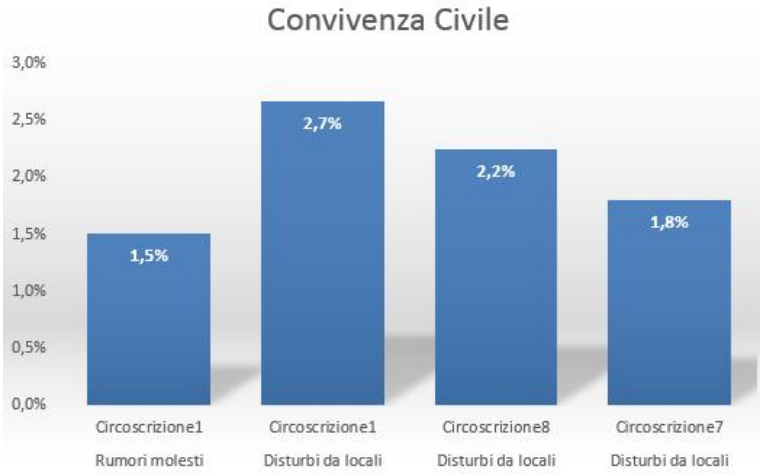


Figura 5.15 – Grafico relativo alla tipologia di regola Sottocategoria → Contesto spaziale per convivenza civile

Per quanto riguarda la qualità urbana, invece, si hanno le seguenti categorie che risaltano maggiormente, come si vede dal grafico in figura 5.16:

- Decoro e degrado urbano
- Veicoli abbandonati

Esse sono all'incirca le due principali sottocategorie appartenenti alla categoria qualità urbana.
Si può notare come nella circoscrizione 1 spicca la sottocategoria decoro e degrado urbano, ovvero data la sottocategoria decoro e degrado urbano, si avrà che le segnalazioni riguardanti quel sottotipo succedono maggiormente nella circoscrizione 1. Il grafico in figura 5.16 ha un range di confidenza che va dal 10% al 26%.

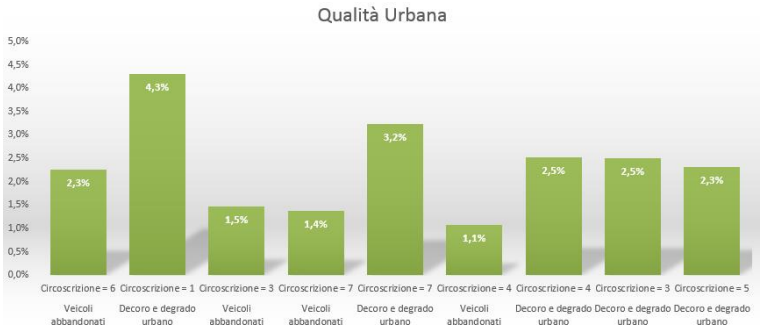


Figura 5.16 – Grafico relativo alla tipologia di regola Sottocategoria → Contesto spaziale per qualità urbana

Andando ad esplorare le regole di associazione al contrario, quindi *Contesto Spaziale → Sottocategoria*, si può notare, analizzando la categoria convivenza civile, che esiste tra le soluzioni la regola trovata precedentemente ovvero data la circoscrizione 1, in quella circoscrizione si avrà una segnalazione di convivenza civile specificatamente come sottocategoria disturbi da locali. Il grafico in figura 5.17 ha un range di confidenza dal 12% al 26%. Si può notare, anche, che risaltano la circoscrizione 8 e la circoscrizione 7 per quanto riguarda i disturbi da locali, che comparivano anche nella regola di verso opposto.

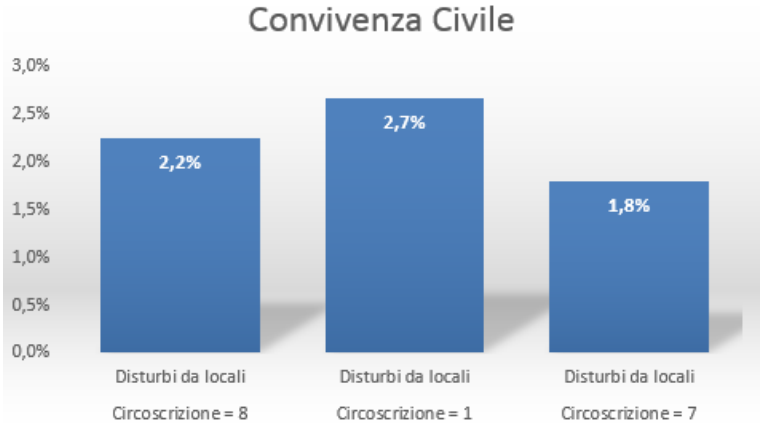


Figura 5.17 – Grafico relativo alla tipologia di regola Contesto spaziale → Sottocategoria per convivenza civile

Per quanto concerne la qualità urbana, invece, si può notare che il picco della sottocategoria decoro e degrado urbano è nella circoscrizione 1, ovvero data la circoscrizione 1, si avrà una segnalazione di sottotipo decoro e degrado urbano appartenente a qualità urbana. Il grafico in questione, figura 5.18, ha un range di confidenza dal 10% al 44%.

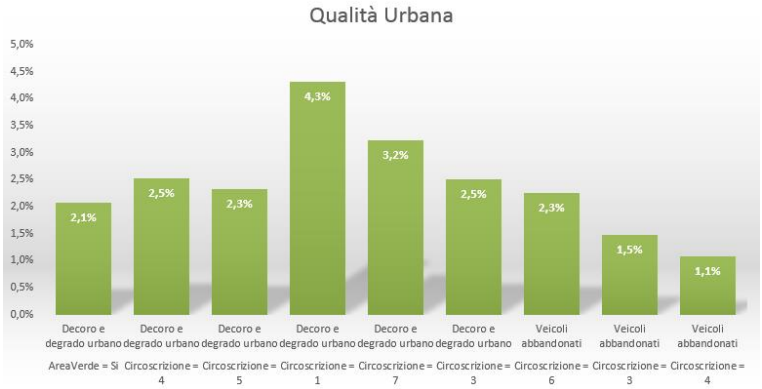


Figura 5.18 – Grafico relativo alla tipologia di regola Contesto spaziale → Sottocategoria per qualità urbana

Inoltre, è possibile notare che la terza categoria principale che è allarme sociale, non compare nell'analisi delle sottocategorie. Evidentemente, nello studiare questa categoria, si devono abbassare ulteriormente il supporto e la confidenza minimi; solo così probabilmente compariranno delle regole di associazione riguardanti questa sottocategoria.
In questa analisi, non si sono tenute in considerazione ulteriori regole, perché ci si è basati su un supporto minimo dell'1% e una minima confidenza del 10% e come ipotesi si è supposto che questi due parametri fossero già tirati a livelli minimi in assoluto, ovvero abbastanza bassi per riuscire ad ottenere risultati importanti.
Per quanto riguarda un'analisi più generale, invece, la regola di associazione che ha maggior supporto e maggior confidenza è *Disturbi da locali → Non area verde*, quindi si può considerare come una tra tutte le regole molto valida in questa analisi.
Un'altra regola abbastanza valida potrebbe essere la sottocategoria *Decoro e degrado urbano → Circoscrizione 1*, con un livello di confidenza leggermente più basso ma comunque ragguardevole.
Un'altra regola simile alla prima è la sottocategoria *Veicoli abbandonati → Non area verde*, con il supporto più alto rispetto a tutte le regole appartenenti a questa tipologia e anche con una confidenza altissima, quindi può essere ritenuta vera a quanto detto dagli indici rilevati.
Per quanto riguarda le regole nel verso opposto si hanno anche *Non area verde → Disturbi da locali*, valida anch'essa. E come validità si può considerare anche la regola nel verso opposto quale *Circoscrizione 1 → Decoro e degrado urbano*.
Tuttavia, la regola di associazione con un supporto minore ma con una maggiore confidenza del 43,20% è data da *Area Verde → Decoro e degrado urbano*.
Spostandoci sul contesto temporale, basandosi sulla tipologia di regola *Sottocategoria → Contesto temporale*, si può notare che anche qui non compare la categoria allarme sociale; l'analisi verrà effettuata sulle altre due categorie convivenza civile e qualità urbana.
Nello specifico, l'analisi è stata svolta non solo per quanto riguarda gli anni, ma anche sulle stagioni e sui mesi in contemporanea.

Infatti a proposito della categoria convivenza civile, figura 5.19, si ha un picco che indica quando la sottocategoria disturbi da locali sono correlati con la stagione primavera, specificatamente data una segnalazione di sottotipo disturbi da locali, essa è stata manifestata specialmente nella stagione primavera. Si hanno anche valori alti per quanto riguarda la stessa sottocategoria con la stagione estate, mentre per quanto riguarda gli anni, si hanno soprattutto l'anno 2012, l'anno 2013, anche se il livello di confidenza è più basso rispetto alla correlazione con le stagioni, e l'anno 2015.
Nell'analisi svolta precedentemente riguardo la categoria (non tenendo conto della sottocategoria) si aveva solamente una correlazione con gli anni 2013 e 2015, ma non il 2012. Infatti il 2012 è associato a disturbi da locali, mentre il 2015 è associato a rumori molesti. Inoltre si può notare anche come compaiono i mesi di Giugno e Luglio, quelli più correlati a questa categoria, avendo il picco per la stagione primavera ed estate; precisamente Giugno è stato considerato nell'analisi come la prima metà mese primaverile e la seconda metà mese estivo, mentre luglio come mese estivo. È pertanto lecito aspettarsi che questi mesi comparissero nell'analisi, riferendosi sempre alla sottocategoria disturbi da locali. Il range di confidenza del grafico in figura 5.19, va dal 13% al 37%.

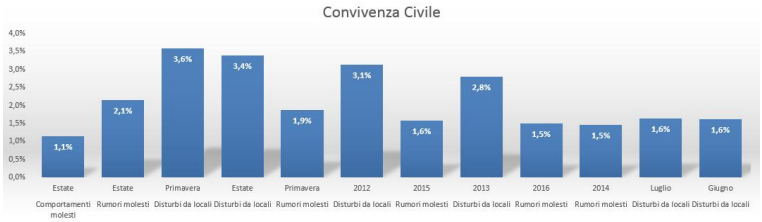


Figura 5.19 – Grafico relativo alla tipologia di regola Sottocategoria → Contesto temporale per convivenza civile

Riferendosi, invece, alla qualità urbana nel contesto temporale, figura 5.20, il picco per quanto riguarda il supporto è dato da decoro e degrado urbano con la stagione autunno, ovvero data una segnalazione con sottocategoria decoro e degrado urbano, la probabilità che si sia verificata è maggiormente nella stagione autunnale. Come seconda regola di associazione, data la stessa segnalazione, la probabilità che maggiormente si è verificata è nell'anno 2016; come terza, la probabilità che si sia verificata è nella stagione inverno; come quarta, sempre a livello di supporto, è per l'anno 2015.
Come detto precedentemente, si può notare tuttavia che la categoria decoro e degrado urbano è correlata al mese di ottobre, infatti risalta la stagione autunno e ottobre in quest'analisi è considerato un mese totalmente autunnale. Mentre considerando la sottocategoria veicoli abbandonati, appartenente alla qualità urbana, si può notare, invece, che la maggior parte delle segnalazioni è avvenuta nelle stagioni primavera ed estate con un picco, anche non indifferente, che mi indica la fascia oraria in cui si sono verificate, ovvero al mattino. Il range di confidenza del grafico in figura 5.20, va dal 10% al 39%. E si ricorda che il grafico è stato ordinato per confidenza, dalla maggiore alla minore.
Nell'analisi delle categorie svolta precedentemente, spiccavano gli anni 2012 e 2014 che si sono collegati con la sottocategoria veicoli abbandonati, mentre analizzando in generale la categoria venivano catalogati comunque come qualità urbana. Quindi, con l'analisi

della sottocategoria, si può concludere che vi è stato un approfondimento specificatamente riguardo al tipo della categoria e quindi a cosa realmente rappresentano questi anni (ad esempio prima qualità urbana in generale e dopo veicoli abbandonati nello specifico).

Figura 5.20 – Grafico relativo alla tipologia di regola Sottocategoria → Contesto temporale per qualità urbana

Guardando la tipologia di regola nel verso opposto, ovvero *Contesto temporale → Sottocategoria*, per quanto riguarda la categoria convivenza civile nell'analisi della categoria in generale si aveva un picco per Giugno e Luglio 2012, infatti qui il picco (considerando la sottocategoria, figura 5.21) è dato solamente dalla sottocategoria disturbi da locali nella stagione primavera. Infatti appare sia Giugno, come analisi del mese, sia 2012, come anno. Inoltre data una stagione, ad esempio estate, si può notare che la segnalazione riguardante la sottocategoria disturbi da locali è accaduta spesso. Ulteriormente si è trovata anche la regola nel quale dato un giorno della settimana, in questo caso venerdì, la maggior parte delle segnalazioni sono state disturbi da locali, e questo può essere un approfondimento aggiuntivo per questa ricerca svolta poiché, essendo Torino una città universitaria, il venerdì sono presenti feste nei locali e quindi tutto ciò può causare delle segnalazioni da parte di persone che si trovano o risiedono nel posto della segnalazione e vengono disturbate da musica alta o schiamazzi relativi ai locali. Vi è anche un picco per la giornata del lunedì, che potrebbe spiegare quanto una persona potrebbe fare una segnalazione nella giornata subito dopo il weekend poiché potrebbe essere che durante il weekend si pensi che nessuno potrebbe rispondere al telefono del contact center della polizia municipale, e quindi lunedì poiché primo giorno lavorativo disponibile.
Il range di confidenza per il grafico in figura 5.21, va dal 13% al 17%, ovvero molto più basso rispetto a quella della regola nel verso opposto.

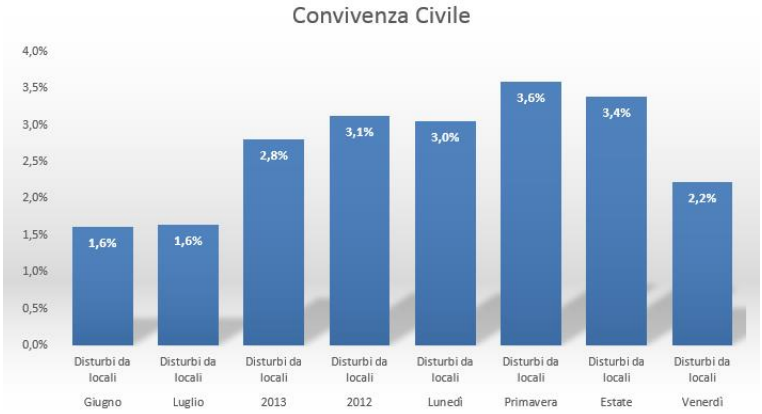


Figura 5.21 – Grafico relativo alla tipologia di regola Contesto temporale → Sottocategoria per convivenza civile

Per quanto concerne, invece, la sottocategoria decoro e degrado urbano, appartenente alla categoria qualità urbana, figura 5.22, data la stagione autunno, si sono verificate diverse segnalazioni di questo sottotipo. Mentre per quanto riguarda la sottocategoria veicoli abbandonati, dato l'anno 2012, si sono verificate diverse segnalazioni di questo sottotipo. Da non sottovalutare il picco del 2016, riferendosi al decoro e al degrado urbano, anche poiché possiede una maggiore confidenza rispetto alla stagione autunno che però ha un supporto leggermente maggiore. Il range di confidenza del grafico in figura 5.22, va dal 10% al 31%, simile a quello della regola nel verso opposto però un pelo più basso.

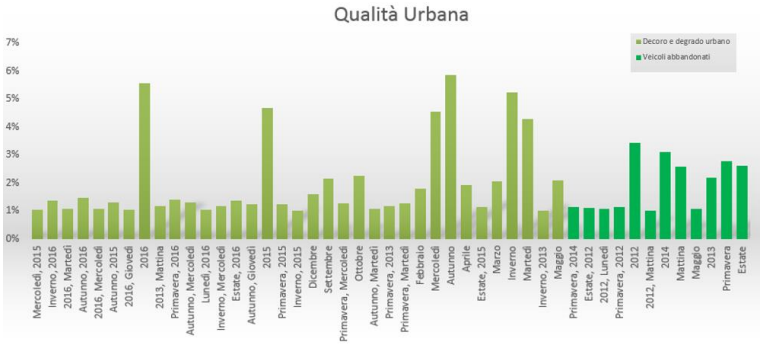


Figura 5.22 – Grafico relativo alla tipologia di regola Contesto temporale → Sottocategoria per qualità urbana

Parlando in generale, tra le regole di associazione sempre legate al contesto temporale, quella che spicca con maggior supporto è data dalla sottocategoria *Decoro e degrado urbano → Autunno.*
Seguono:
Decoro e degrado urbano → 2016
Decoro e degrado urbano → Inverno
Decoro e degrado urbano → 2015.
Per queste quattro regole i livelli di confidenza non sono altissimi, infatti analizzando la regola per quanto riguarda la confidenza, si ha un valore altissimo per la sottocategoria *Veicoli abbandonati → 2012*, oppure per la sottocategoria *Comportamenti molesti → Estate*, oppure *Veicoli abbandonati → 2014*.

Andando a vedere le regole nel verso opposto, si hanno:
Autunno → Decoro e degrado urbano
2016 → Decoro e degrado urbano
Inverno → Decoro e degrado urbano
2015 → Decoro e degrado urbano.
Quindi, in generale, le regole di associazione al contrario, contengono gli stessi valori per gli stessi attributi.
Per quanto riguarda, invece, la regola al contrario, si può dire che i Mercoledì dell'anno 2015 implicano la sottocategoria decoro e degrado urbano.

5.4 Circoscrizioni e quartieri

Prima di introdurre l'analisi che sarà svolta specificatamente sulle circoscrizioni del paragrafo successivo, si è pensato di procedere evidenziando quali quartieri e quali zone appartenessero alle diverse circoscrizioni, figura 5.23.

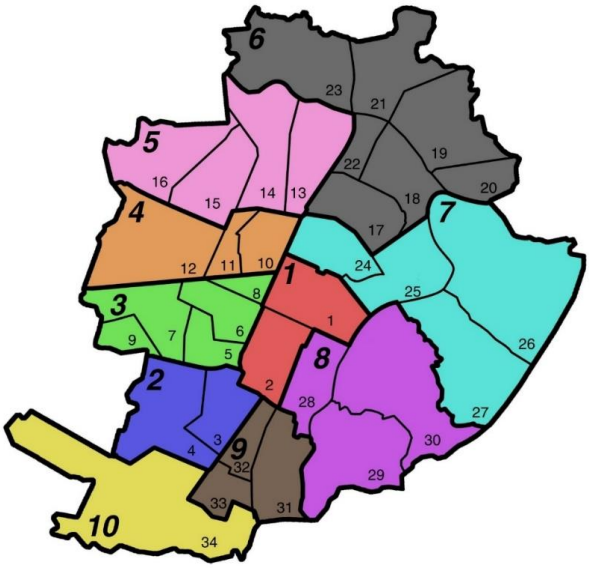


Figura 5.23 – Le circoscrizioni e i quartieri di Torino, tratta da https://i2.wp.com/www.volantinaggio-torino.com/wp-content/uploads/2016/12/Quartieri_torino.jpg?resize=768%2C566

Le circoscrizioni di Torino [42] comprendono i seguenti quartieri/zone, come evidenziato dalla mappa nella figura. La descrizione è rappresentata in tabella 8.

Circoscrizione 1	Centro, Crocetta	In rosso
Circoscrizione 2	Santa Rita, Mirafiori Nord	In blu
Circoscrizione 3	Borgo San Paolo, Cenisia, Pozzo Strada, Cit Turin, Borgata Lesna	In verde
Circoscrizione 4	San Donato, Campidoglio, Parella	In arancione

Circoscrizione 5	Borgo Vittoria, Madonna di Campagna, Lucento, Vallette	In rosa
Circoscrizione 6	Barriera di Milano, Regio Parco, Barca, Bertolla, Falchera, Rebaudengo, Villaretto	In grigio
Circoscrizione 7	Aurora, Vanchiglia, Sassi, Madonna del Pilone	In azzurro
Circoscrizione 8	San Salvario, Cavoretto, Borgo Po	In viola
Circoscrizione 9	Nizza Millefonti, Lingotto, Filadelfia	In marrone
Circoscrizione 10	Mirafiori Sud	In giallo

Tabella 8 – Quartieri appartenenti alle circoscrizioni di Torino

Dopo questa breve spiegazione, riguardante la città di Torino, si può procedere con l'analisi.

5.5 Analisi dei dataset divisi per anno e per sottocategoria

Come descritto nel paragrafo precedente, si sono analizzate le regole di associazione trovate con indici interessanti e si è cercato di riportarle in grafici per capire meglio e cercare di trovare un andamento con tutti i dati a nostra disposizione, ovvero considerando come fascia temporale tutti gli anni dal 2012 al 2016 compresi.
In questo paragrafo, non si sono confrontate tutte le regole ma sono state analizzate le regole di associazione suddivise per anno ed essenzialmente ci si è posti delle domande che potevano essere significative. La domanda più notevole che si è posta è: "dato un determinato luogo (inteso come circoscrizione e in uno sviluppo futuro potrà essere anche il quartiere anziché un insieme di quartieri), cosa succede in quel luogo? E questo come varia nel tempo?". Infatti l'attenzione si è incentrata nell'andare a trovare regole di associazione interessanti sia statisticamente sia per quello che dice la regola; e soprattutto come queste regole variassero nel tempo. Infatti, la suddivisione del dataset per anni è servita per fare un'analisi delle regole di associazione prese separatamente, in particolare nel contesto spaziale (circoscrizione) e per sottocategoria (poiché analisi più approfondita). Successivamente verrà confrontato anno per anno separatamente per vedere come varia e come si sposta di luogo per esempio una segnalazione pericolosa o in uno sviluppo futuro, più nel dettaglio, un quartiere pericoloso.
L'interesse, pertanto, è stato spostato non tanto su una regola di associazione in sé, ma soprattutto su come questa regola varia nel tempo e ciò è stato fatto analizzando l'andamento di supporto e confidenza, come nell'analisi precedente.
Si potrebbe anche notare se compaiono delle regole di associazione che, sull'analisi precedente sul dataset intero, magari, non sarebbero apparse oppure se si modificano gli indici. Questo è il metodo utilizzato per l'esplorazione delle regole.
Gli anni analizzati vanno dal 2012 al 2016 compresi, quindi cinque anni. Si sono ulteriormente creati cinque grafici corrispondenti a questi cinque anni che verranno poi messi a confronto tra di loro per vedere come una segnalazione può propagarsi nel tempo.
Partendo dal 2012, si sono cercate tutte le regole che appartenessero al contesto spaziale che implica una determinata sottocategoria (*Contesto Spaziale* → *Sottocategoria*). Da

notare come il grafico segua un po' la linea dei precedenti: i blocchi sono stati ordinati per confidenza, dal primo a sinistra che ha la maggior confidenza all'ultimo a destra che ha la minor confidenza; l'altezza dei blocchi, invece, varia in base al supporto. Rispetto ai grafici precedenti, la confidenza questa volta è riportata in nero in percentuale sopra il blocco a cui si riferisce. In aggiunta, la categoria convivenza civile (con tutte le sue sottocategorie) sono contrassegnate dal colore blu, la qualità urbana (con tutte le sue sottocategorie) sono contrassegnate dal colore verde e l'allarme sociale da quello rosso (anche se non compare mai in quest'analisi). Non sono risultate regole al di sopra del minimo supporto (1%) e della minima confidenza (10%) per quanto riguarda la categoria allarme sociale, poiché queste regole non sono state ritenute fondamentali per quest'analisi.
Nell'introdurre questa analisi supplementare effettuata, si può notare che non tutte le sottocategorie sono risaltate, ma quelle messe in evidenza sono quattro:

- Disturbi da locali, convivenza civile
- Rumori molesti, convivenza civile
- Veicoli abbandonati, qualità urbana
- Decoro e degrado urbano, qualità urbana.

Il primo grafico in figura 5.24 corrisponde all'anno 2012. Per quest'anno, per la categoria qualità urbana e la sottocategoria decoro e degrado urbano, si può notare che il maggior supporto è associato alla circoscrizione 1, però con un livello di confidenza di circa il 17%, pertanto non altissimo. Mentre per quanto riguarda la maggior confidenza, sempre per la regola riguardante la sottocategoria decoro e degrado urbano, è associata alla circoscrizione 5. Sempre restando nella macrocategoria qualità urbana, si può notare come la sottocategoria veicoli abbandonati abbia un maggior supporto e una maggiore confidenza nella circoscrizione 6.
Passando alla categoria convivenza civile per l'anno 2012, si può notare che compare solamente la sottocategoria disturbi da locali, in cui il maggior supporto si trova nella circoscrizione 1, mentre il maggior livello di confidenza lo si può trovare nella regola di associazione corrispondente alla circoscrizione 8.

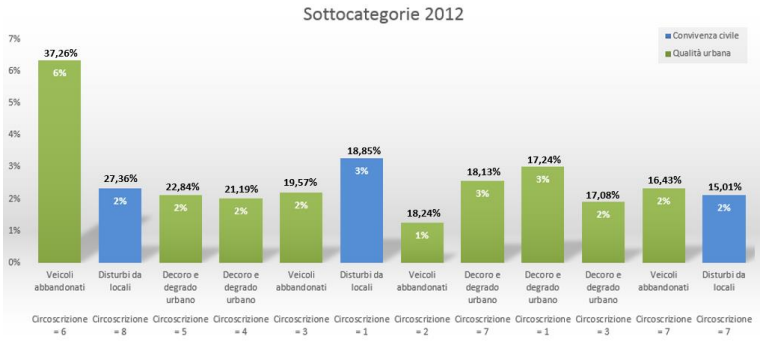


Figura 5.24 – Grafico relativo alla tipologia di regola Contesto spaziale → Sottocategoria per l'anno 2012

Procedendo verso l'anno 2013, grafico in figura 5.25, si può notare per quanto riguarda la convivenza civile, l'unica sottocategoria apparsa è quella riferita ai disturbi da locali e la regola con il maggior supporto è quella regola di associazione in cui è presente la circoscrizione 1, mentre la regola con maggiore confidenza è quella che contiene la circoscrizione 8, esattamente come si è visto per l'anno 2012. Quindi, passando dall'anno 2012 al 2013, per questa categoria ovvero la convivenza civile, non si sono verificati cambiamenti rispetto alle zone, si sono solo verificati a livello di valore di supporto e di confidenza.
Spostando l'attenzione sull'altra macrocategoria, si ha come sottocategoria decoro e degrado urbano con un maggior supporto nella circoscrizione 1, e una maggiore confidenza nella circoscrizione 7. Per l'altra sottocategoria, veicoli abbandonati, si ha il picco della circoscrizione 6 con entrambi maggior supporto e maggior confidenza rispetto ad altri risultati.

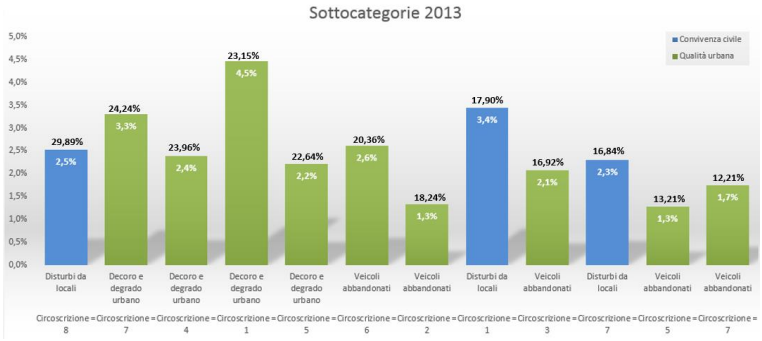


Figura 5.25 – Grafico relativo alla tipologia di regola Contesto spaziale → Sottocategoria per l'anno 2013

Passando al 2014, come descritto dal grafico in figura 5.26, per la categoria convivenza civile, compare sempre e solo la sottocategoria disturbi da locali, nella quale si ha un maggior supporto nella circoscrizione 1 e una maggiore confidenza per la circoscrizione 8. Da notare che le circoscrizioni relative a queste sottocategorie non cambiano nel tempo, infatti, per il terzo anno consecutivo, continuano ancora a rimanere costanti secondo questa analisi.
A proposito, invece, della categoria qualità urbana, si ha decoro e degrado urbano con un maggior supporto nella circoscrizione 7 e una maggior confidenza (anche se di poco) nella circoscrizione 3, anche se tra le due circoscrizioni c'è solamente un punto percentuale di delta nelle confidenze. Assumendo quindi confidenza uguale, viene tenuta più in considerazione la circoscrizione 7 poiché presenta un maggior supporto, quindi è ritenuta una regola di associazione di maggior importanza. Per quanto riguarda i veicoli abbandonati, si ha un maggior supporto e una maggiore confidenza nella circoscrizione 3 e 4. Anch'esse hanno più o meno stesso livello di supporto e stessa confidenza e quindi potrebbero essere guardate contemporaneamente entrambe. Tra l'altro le circoscrizioni 3 e 4 riguardano zone di Torino affiancate e quindi, se la ricerca fosse stata svolta per aree, potrebbero identificarsi come un'unica area.
Concludendo, la sottocategoria decoro e degrado urbano, che rimaneva costante nella circoscrizione 1 fino all'anno 2013, già nel 2013 e per tutto il 2014 si sposta nella

circoscrizione 7. Mentre per la sottocategoria dei veicoli abbandonati, essi non sono più stati ritrovati nella circoscrizione 6, ma bensì tra le circoscrizioni 3 e 4 nel 2014.

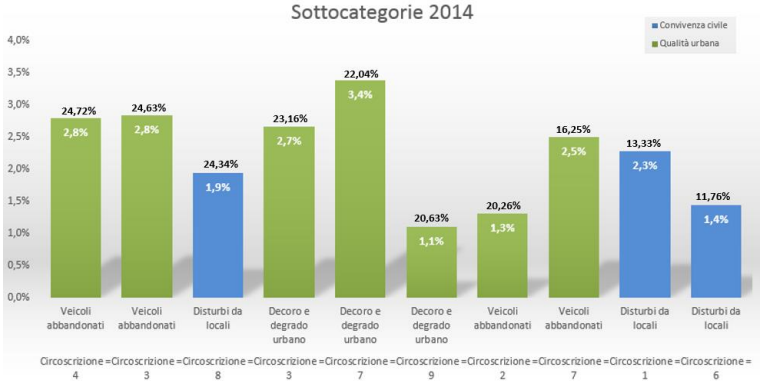


Figura 5.26 – Grafico relativo alla tipologia di regola Contesto spaziale → Sottocategoria per l'anno 2014

Arrivando ad analizzare il 2015, come mostrato dal grafico in figura 5.27, si vede che per quanto riguarda la convivenza civile rientra in gioco la sottocategoria rumori molesti che compare nella circoscrizione 3. Si vedrà poi nell'analisi che per questa sottocategoria non si avranno più apparizioni.
Per quanto riguarda la sottocategoria disturbi da locali, si può notare un maggior supporto e una maggiore confidenza nella circoscrizione 8. Infatti, con l'anno 2015, scompare quasi del tutto per questa sottocategoria, la circoscrizione 1. È come se la "vita dei locali notturni" si fosse spostata e fosse diminuita sempre gradualmente per arrivare nel 2015 a non apparire quasi più, mentre rimane costante nella circoscrizione 8.
Analizzando invece la categoria qualità urbana, si può innanzitutto notare che i veicoli abbandonati sono spariti in quest'anno, mentre per il decoro e degrado urbano si ha un maggior supporto nella circoscrizione 1 e una maggiore confidenza, anche se di poco, nella circoscrizione 4. Si può quindi tenere in considerazione la circoscrizione 1 poiché il livello di confidenza cambia di pochi decimi, perciò si terrà la regola di associazione con maggior supporto che corrisponde alla circoscrizione 1.
Infatti, come si può constatare, si era notato che nel 2012 e 2013, la circoscrizione 1 era contrassegnata dalla sottocategoria decoro e degrado urbano, poi tra il 2013 e il 2014 si è spostata verso la circoscrizione 7 e con il ritornare nella circoscrizione 1 nell'anno 2015.

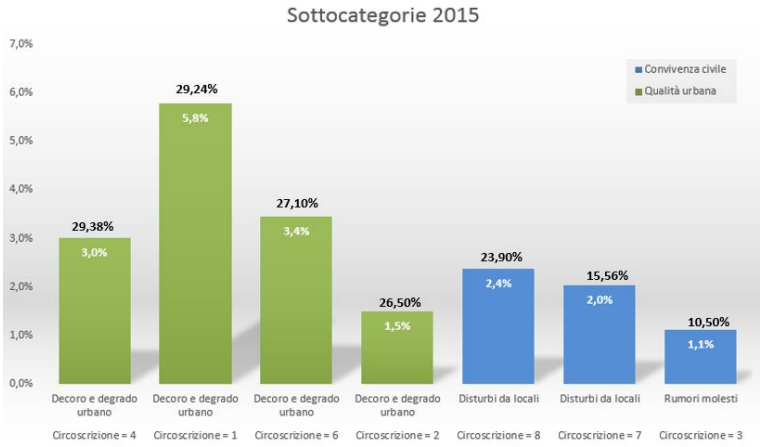


Figura 5.27 – Grafico relativo alla tipologia di regola Contesto spaziale → Sottocategoria per l'anno 2015

Andando ad analizzare l'ultimo anno preso in considerazione, il 2016, figura 5.28, per la categoria convivenza civile, si può trovare un maggior supporto e una maggiore confidenza per la circoscrizione 8, mentre per la categoria qualità urbana, più nello specifico per la sottocategoria decoro e degrado urbano, si nota un maggior supporto nella circoscrizione 5 e una maggiore confidenza nella circoscrizione 9.
Nell'anno 2016, non compare né la sottocategoria rumori molesti per la convivenza civile, né veicoli abbandonati per la qualità urbana.

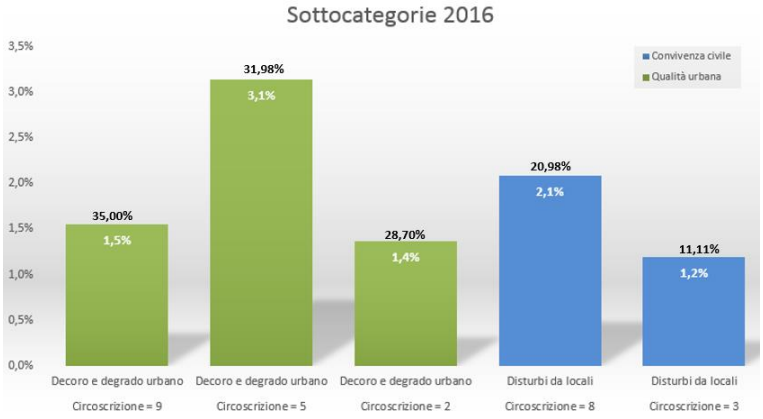


Figura 5.28 – Grafico relativo alla tipologia di regola Contesto spaziale → Sottocategoria per l'anno 2016

In conclusione, si può arrivare ad un determinato risultato che implica che, per quanto riguarda la convivenza civile, a parte la sottocategoria rumori molesti che compare solamente nella circoscrizione 3 nell'anno 2015, per quanto riguarda l'altra sottocategoria ovvero i disturbi da locali, si nota un andamento costante che parte dalla circoscrizione 1 e dalla circoscrizione 8 fino all'anno 2014; nell'anno 2015 scompare la circoscrizione 1 ma rimane costante la circoscrizione 8, e così anche nell'anno 2016. Riassumendo, un andamento costante per la circoscrizione 8, mentre un andamento costante fino all'anno 2014 per la circoscrizione 1.

Per la categoria qualità urbana, considerando la sottocategoria decoro e degrado urbano, si ha la circoscrizione 1 tra gli anni 2012 e 2013, poi, dalla circoscrizione 1, si passa alla circoscrizione 7, tra il 2013 e il 2014 per poi tornare nel 2015 alla circoscrizione 1. Da tenere in considerazione la circoscrizione 5 che compare nel 2012, sparisce negli altri anni, per poi riapparire nel 2016.
Per la sottocategoria veicoli abbandonati, invece, si ha la circoscrizione 6 negli anni 2012 e 2013, per l'anno 2014 la zona si sposta dalla circoscrizione 6 alla circoscrizione 3 e 4, che essendo diverse sono affiancate quindi si potrebbe identificare come una grandissima unica zona, poi negli anni 2015 e 2016 questa sottocategoria sparisce; il che ci porta a dire che nei primi anni c'era un maggior abbandonamento di macchine mentre dopo non vi è più una probabilità così grande di trovare in giro veicoli abbandonati. Ciò è positivo per l'analisi, poiché non si verificano più segnalazioni di questo tipo e gli standard vengono alzati, ed è positivo anche per la sicurezza urbana di Torino.

6. Analisi dei risultati con il classificatore associativo

Questo capitolo è dedicato all'interpretazione dei risultati ottenuti con il classificatore associativo con il software Weka.
Nel primo paragrafo di questa sezione sperimentale sono stati analizzati i parametri dell'algoritmo di classificazione per vedere qual è la configurazione più opportuna e qual è l'impatto sulle performance, in termini di accuratezza e di numero di regole associative trovate. Nel secondo paragrafo, invece, si sono descritti i risultati ottenuti, ovvero le regole associative estratte, e si è cercato di dare un'interpretazione ad esse. Si è cercato, come nel capitolo precedente di separare il contesto spaziale da quello temporale per rendere i risultati in maniera più ordinata. Le regole associative sono state filtrate dal punto di vista degli indici supporto e confidenza, e analizzate dal punto di vista del contenuto informativo.

6.1 Analisi quantitativa delle prestazioni del classificatore associativo

Il punto di partenza di questa analisi è stato l'estrazione di informazione dagli output del software Weka. Esso è stato fatto girare circa 300 volte, 150 per le categorie e 150 per le sottocategorie. I dataset di input sono stati i seguenti:

- Anno 2012 per la Categoria
- Anno 2013 per la Categoria
- Anno 2014 per la Categoria
- Anno 2015 per la Categoria
- Anno 2016 per la Categoria
- Anni tra il 2012 e il 2016 per la Categoria
- Anno 2012 per la Sottocategoria
- Anno 2013 per la Sottocategoria
- Anno 2014 per la Sottocategoria
- Anno 2015 per la Sottocategoria
- Anno 2016 per la Sottocategoria
- Anni tra il 2012 e il 2016 per la Sottocategoria.

Per ognuno di questi dataset, è stata fatta l'analisi dei parametri utilizzando la tecnica di separazione delle variabili: è stata fissata una variabile in una configurazione standard (ad esempio in questo caso il supporto a 1 e la confidenza al 50%), ed è stata fatta variare l'altra variabile.

Per ogni dataset (di quelli nell'elenco puntato precedente), è stato fatto variare il supporto, tenendo la confidenza costante al 50%, per le categorie a: 0.1, 0.5, 1, 5, 10, 15, 20, 25, 30.
Per le sottocategorie a: 0.1, 0.5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10.
È stata fatta la stessa cosa, tenendo il supporto fisso a 1, e fatto variare la confidenza, sia per le categorie che per le sottocategorie a: 30%, 50%, 55%, 60%, 65%, 70%, 75%, 80%
Le performance sono state misurate in termini di accuratezza e in base al numero di regole di primo e secondo livello. È stata misurata, in primo luogo, il cambiamento che subisce l'accuratezza al variare del supporto, tenendo la confidenza al 50%, per ogni anno e per tutti gli anni per la classe categoria (figure dalla 6.1 alla 6.6) e per la classe sottocategoria (figure dalla 6.7 alla 6.12); successivamente è stato misurato sempre l'andamento dell'accuratezza al variare della confidenza, tenendo il supporto a 1, per ogni anno e per tutti gli anni per la classe categoria (figure dalla 6.13 alla 6.18) e per la classe sottocategoria (figure dalla 6.19 alla 6.24).
Successivamente, sono state analizzate il numero di regole associative estratte per ogni modello sempre al variare dei due parametri. La variazione del numero di regole di primo livello (linea azzurra) e la variazione del numero di regole di secondo livello (in blu) sono state riportate nei grafici al variare del supporto per ogni anno e per tutti gli anni, tenendo la confidenza al 50%, per la categoria (figure dalle 6.25 alla 6.30) e per la sottocategoria (figure dalla 6.31 alla 6.36); e al variare della confidenza, tenendo il supporto a 1, per ogni anno e per tutti gli anni per la categoria (figure dalla 6.37 alla 6.42) e per la sottocategoria (figure dalla 6.43 alla 6.48).
Analizzando i grafici nelle figure dalla 6.1 alla 6.11, si può notare che per supporti più bassi si ha un modello un po' più accurato e al crescere del supporto, il modello è sempre meno accurato. Quindi si va da una situazione di underfitting, ovvero che il modello è troppo poco accurato per descrivere la situazione, i dati vengono descritti in maniera troppo grezza e rozza. All'aumentare del supporto, la linea decresce, generando l'overfitting cioè quando il supporto è troppo basso, e si vanno a prendere delle correlazioni che sono a volte troppo specifiche e non opportune). Bisognerebbe trovare un compromesso ottimale tra generalità del modello e specializzazione di questo: questo si può trovare nella fascia centrale grafico, dove il modello non è troppo generale, ma neanche troppo specializzato. Infatti per l'estrazione delle regole associative, descritte nel paragrafo successivo, si è scelta la configurazione con supporto uguale a 1 (configurazione di default), poiché il classificatore performa meglio.
Per quanto riguarda la categoria (figure dalla 6.1 alla 6.6), nell'anno 2012 l'accuratezza massima, del 56,87%, si ha con supporto uguale a 5, per l'anno 2013 l'accuratezza massima del 55,51%, con supporto uguale a 1, per il 2014 del 53,79% con supporto uguale a 10, per il 2015 del 54,28% con supporto uguale a 5 e per il 2016 del 54,25% con supporto uguale a 1. Tenendo in considerazione tutti gli anni, invece, il picco lo si ha per supporti più bassi di 1, ovvero del 54,92% con supporti dello 0,1 e 0,5.

Per la sottocategoria in tutti i grafici (quindi in tutti gli anni, figure dalle 6.7 alla 6.12) il picco lo si trova fino a supporto uguale a 1, e l'accuratezza decresce fino a toccare lo zero, come si nota negli anni 2012 e 2013. L'accuratezza per tutti i grafici riguardanti la sottocategoria, assume un massimo tra 44% e 48%.

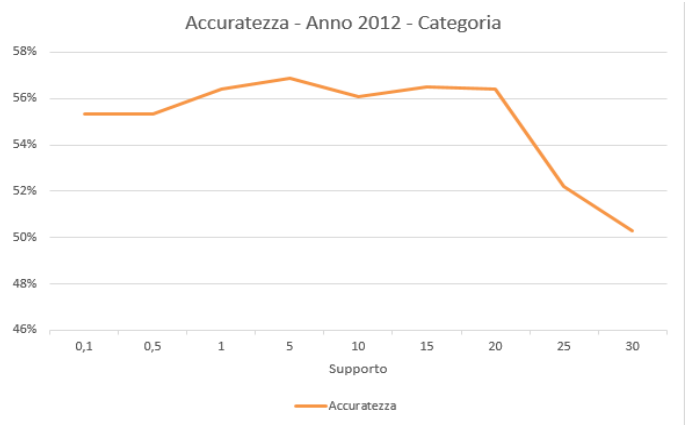


Figura 6.1 – Variazione dell'accuratezza quando varia il supporto, con una confidenza del 50%, per la categoria nell'anno 2012.

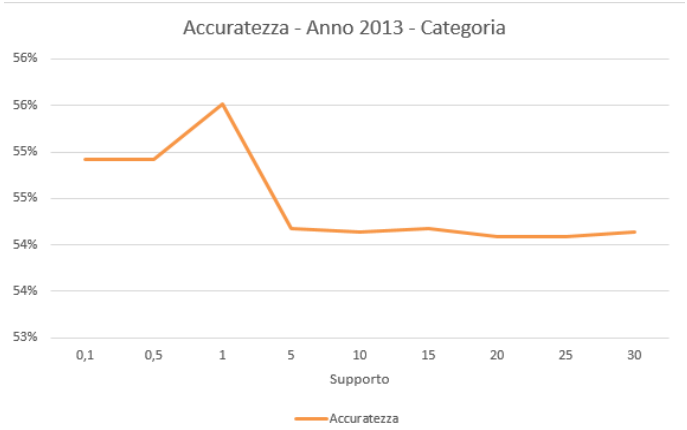


Figura 6.2 – Variazione dell'accuratezza quando varia il supporto, con una confidenza del 50%, per la categoria nell'anno 2013.

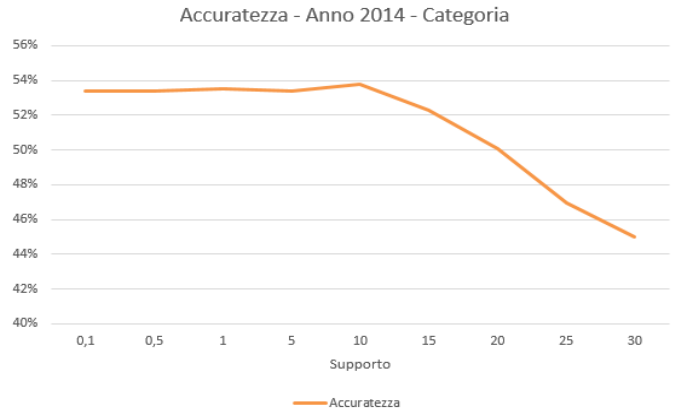


Figura 6.3 – Variazione dell'accuratezza quando varia il supporto, con una confidenza del 50%, per la categoria nell'anno 2014.

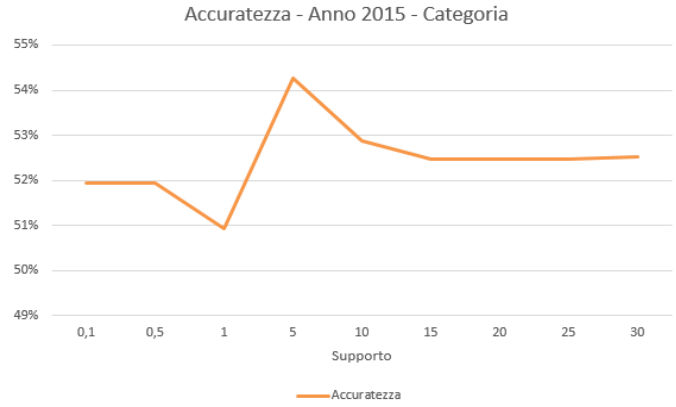


Figura 6.4 – Variazione dell'accuratezza quando varia il supporto, con una confidenza del 50%, per la categoria nell'anno 2015.

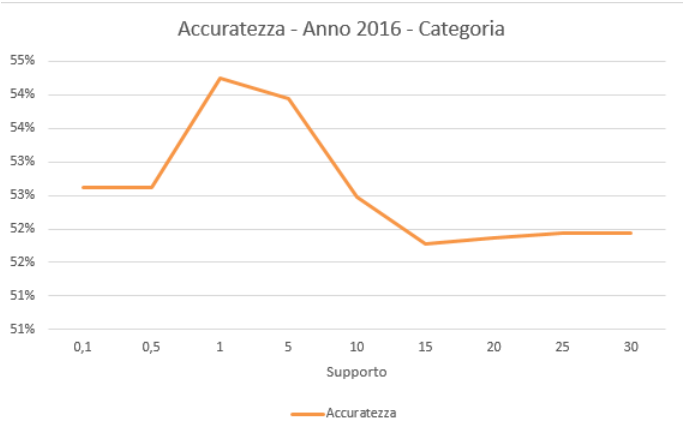


Figura 6.5 – Variazione dell'accuratezza quando varia il supporto, con una confidenza del 50%, per la categoria nell'anno 2016.

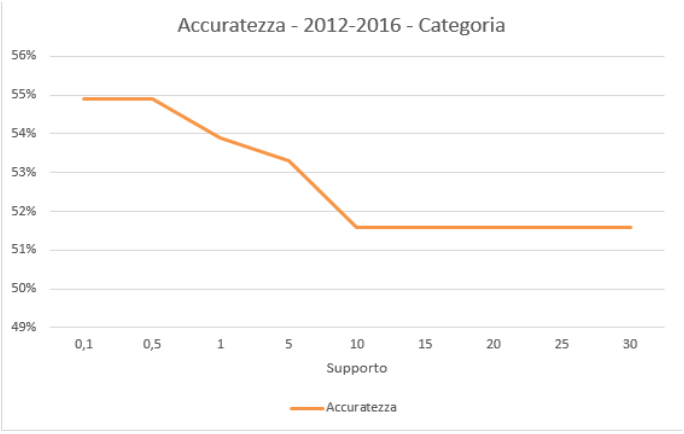


Figura 6.6 – Variazione dell'accuratezza quando varia il supporto, con una confidenza del 50%, per la categoria dal 2012 al 2016.

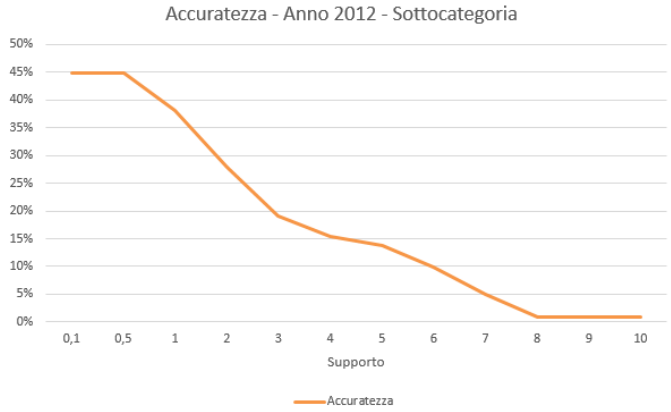


Figura 6.7 – Variazione dell'accuratezza quando varia il supporto, con una confidenza del 50%, per la sottocategoria nell'anno 2012.

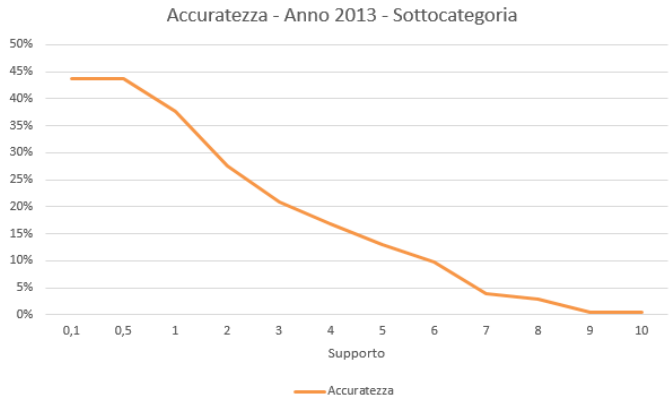


Figura 6.8 – Variazione dell'accuratezza quando varia il supporto, con una confidenza del 50%, per la sottocategoria nell'anno 2013.

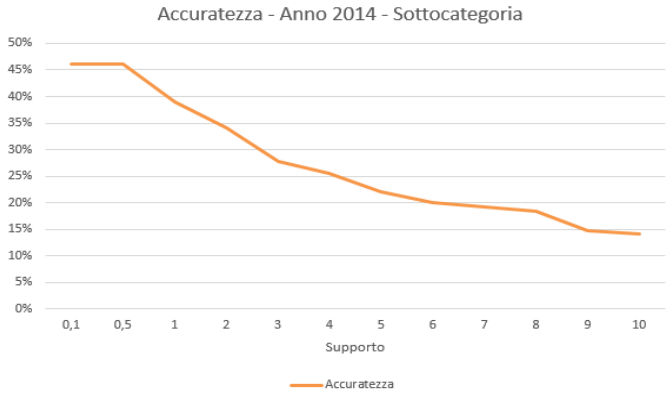


Figura 6.9 – Variazione dell'accuratezza quando varia il supporto, con una confidenza del 50%, per la sottocategoria nell'anno 2014.

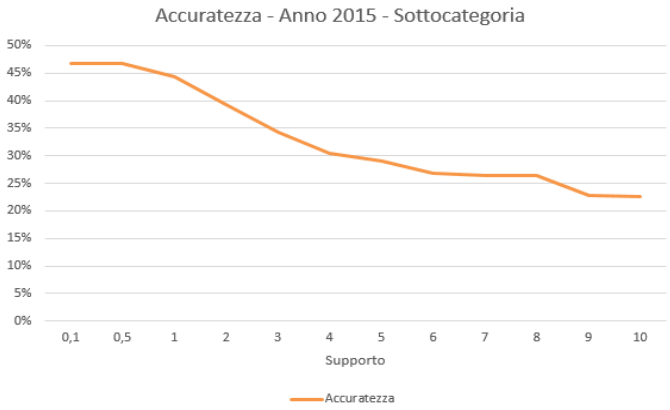


Figura 6.10 – Variazione dell'accuratezza quando varia il supporto, con una confidenza del 50%, per la sottocategoria nell'anno 2015.

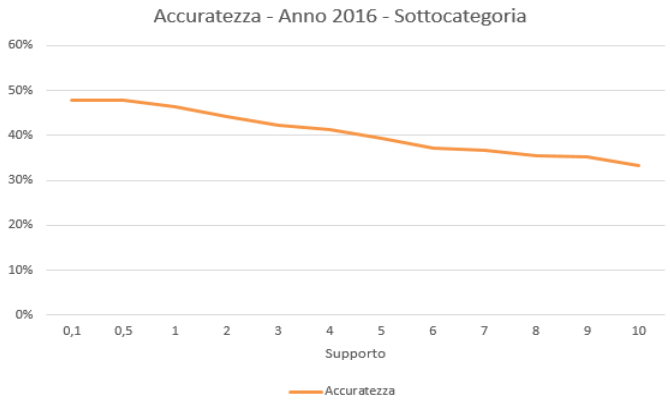


Figura 6.11 – Variazione dell'accuratezza quando varia il supporto, con una confidenza del 50%, per la sottocategoria nell'anno 2016.

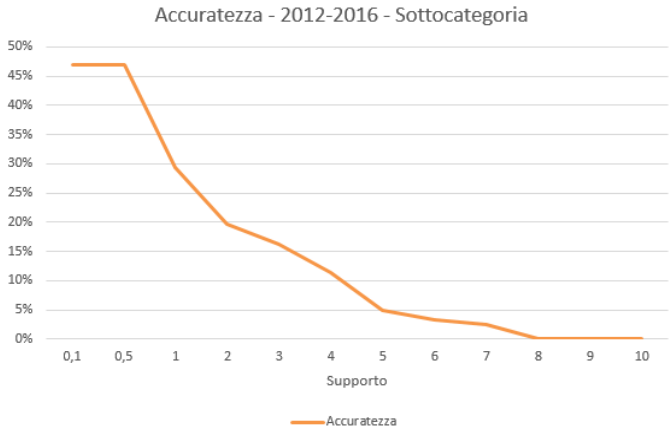


Figura 6.12 – Variazione dell'accuratezza quando varia il supporto, con una confidenza del 50%, per la sottocategoria dal 2012 al 2016.

Un ragionamento analogo è da fare con la confidenza.
Infatti, per quanto riguarda la categoria (figure dalla 6.13 alla 6.18), l'andamento è sempre decrescente con un massimo di accuratezza per confidenze più basse per ogni singolo anno, mentre per quanto riguarda il dataset comprendente tutti gli anni insieme, l'andamento rimane sempre decrescente ma con un picco di minor importanza (17,62%) attorno a una confidenza del 75%. L'accuratezza massima per la categoria va da un 51% riscontrato nel 2015 ad un 56% riscontrato nel 2012 e 2013.
L'andamento è lo stesso anche per quanto riguarda la sottocategoria (figure dalla 6.19 alla 6.24), e l'accuratezza massima va da un 42% riscontrato nel 2013 ad un 48% riscontrato nel 2016.
Si può dire che con il trascorrere degli anni, l'accuratezza cresce.

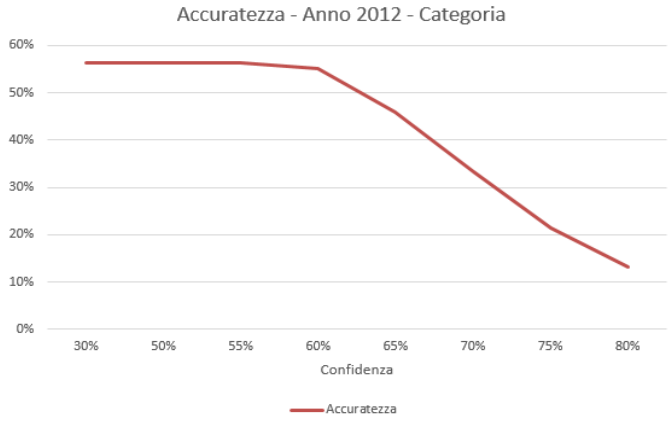


Figura 6.13 – Variazione dell'accuratezza quando varia la confidenza, con un supporto di 1, per la categoria nell'anno 2012.

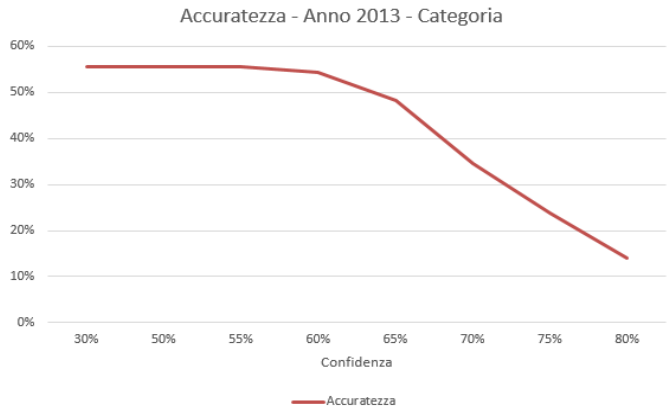


Figura 6.14 – Variazione dell'accuratezza quando varia la confidenza, con un supporto di 1, per la categoria nell'anno 2013.

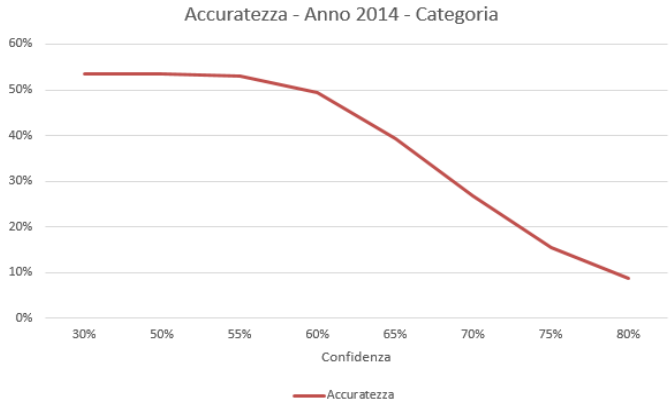


Figura 6.15 – Variazione dell'accuratezza quando varia la confidenza, con un supporto di 1, per la categoria nell'anno 2014.

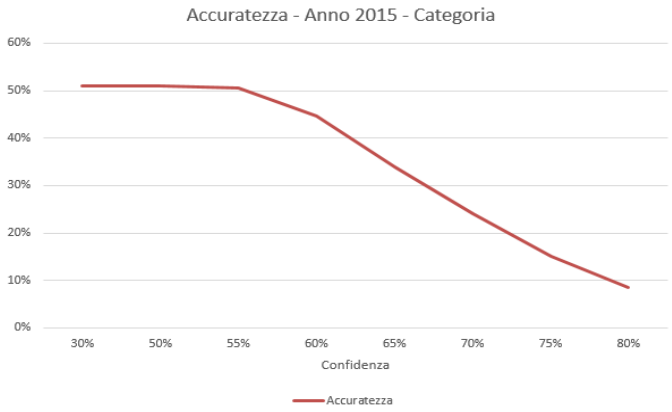


Figura 6.16 – Variazione dell'accuratezza quando varia la confidenza, con un supporto di 1, per la categoria nell'anno 2015.

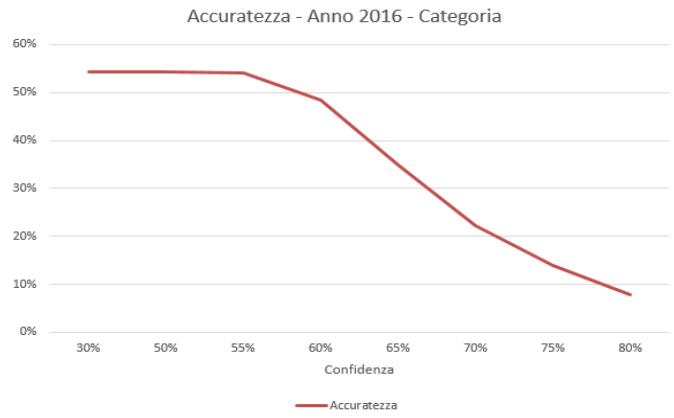


Figura 6.17 – Variazione dell'accuratezza quando varia la confidenza, con un supporto di 1, per la categoria nell'anno 2016.

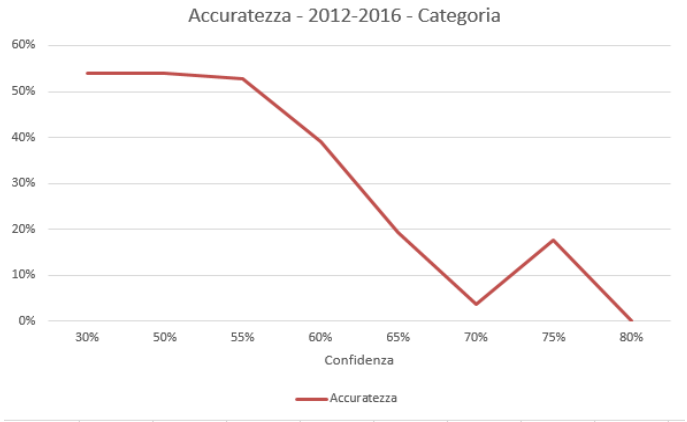


Figura 6.18 – Variazione dell'accuratezza quando varia la confidenza, con un supporto di 1, per la categoria dal 2012 al 2016..

Figura 6.19 – Variazione dell'accuratezza quando varia la confidenza, con un supporto di 1, per la sottocategoria nell'anno 2012.

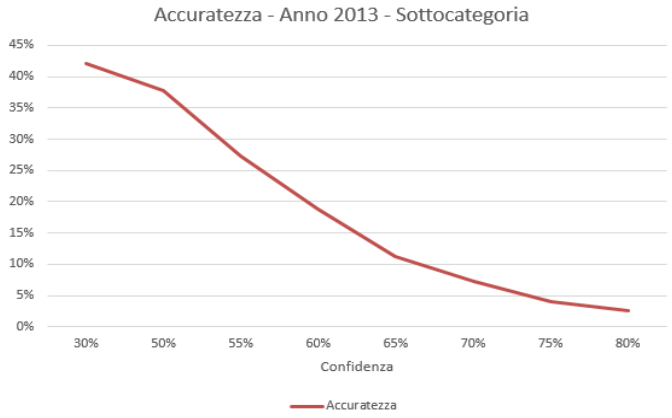


Figura 6.20 – Variazione dell'accuratezza quando varia la confidenza, con un supporto di 1, per la sottocategoria nell'anno 2013.

Figura 6.21 – Variazione dell'accuratezza quando varia la confidenza, con un supporto di 1, per la sottocategoria nell'anno 2014.

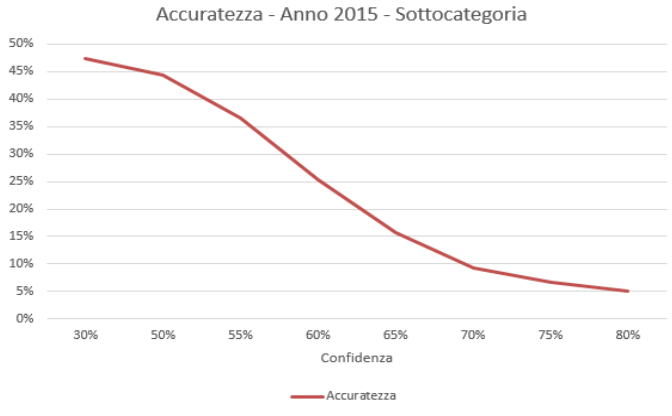


Figura 6.22 – Variazione dell'accuratezza quando varia la confidenza, con un supporto di 1, per la sottocategoria nell'anno 2015.

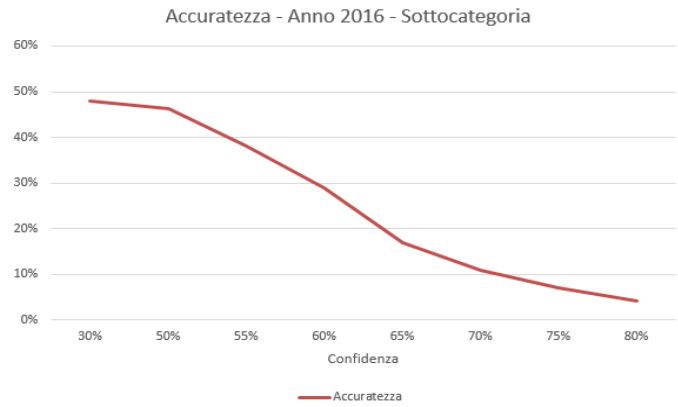


Figura 6.23 – Variazione dell'accuratezza quando varia la confidenza, con un supporto di 1, per la sottocategoria nell'anno 2016.

Figura 6.24 – Variazione dell'accuratezza quando varia la confidenza, con un supporto di 1, per la sottocategoria dal 2012 al 2016.

Successivamente, sono state analizzate il numero di regole associative estratte al variare del supporto. Si può notare che, mantenendo una confidenza costante del 50%, l'andamento è decrescente per quanto riguarda la categoria e anche per la sottocategoria. Con supporti più bassi si ha un maggior numero di regole associative, infatti, era da aspettarsi questo risultato, perché, man mano che il supporto cresce. Le regole vengono filtrate e quindi sono mantenute soltanto le più significative, e quindi si riducono anche di numero. Per la categoria (figure dalla 6.25 alla 6.30), come regole di primo livello si va da un minimo di 3 per l'anno 2014 e 2015 (range da 3 a 6), fino ad un massimo di 1578 per l'anno 2014 (range da 1391 a 1578); per le regole di secondo livello si va da nessuna nell'anno 2014 (range da 0 a 4), fino ad un massimo di 16652 per l'anno 2012 (range da 12875 a 16652). Considerando il dataset intero si ha per le regole di primo livello un minimo di 4 e un massimo di 6783, mentre per il secondo un minimo di 3 e un massimo di 131541 regole.

Per la sottocategoria (figure dalla 6.31 alla 6.36), come regole di primo livello si va da un minimo di 32 per l'anno 2014 (range da 32 a 58), fino ad un massimo di 1894 per l'anno 2012 (range da 1546 a 1894); per le regole di secondo livello si va da 133 nell'anno 2014 (range da 133 a 420), fino ad un massimo di 16567 per l'anno 2012 (range da 12231 a 16567). Considerando il dataset intero si ha per le regole di primo livello un minimo di 19 e un massimo di 8119, mentre per il secondo un minimo di 205 e un massimo di 127938 regole.

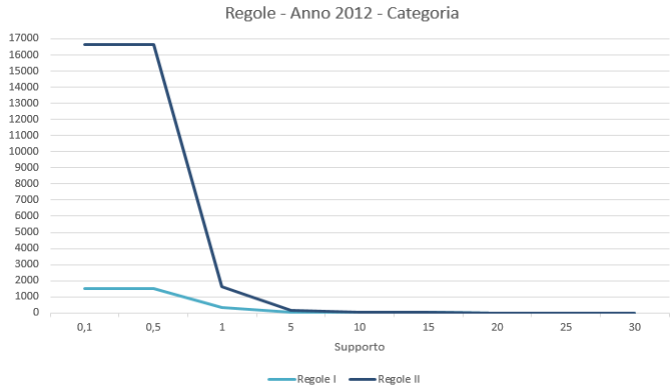


Figura 6.25 – Numero di regole associative al variare del supporto, con una confidenza del 50%, per la categoria nell'anno 2012.

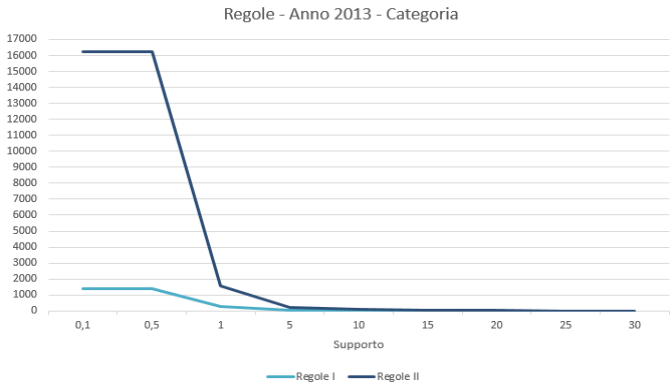


Figura 6.26 – Numero di regole associative al variare del supporto, con una confidenza del 50%, per la categoria nell'anno 2013.

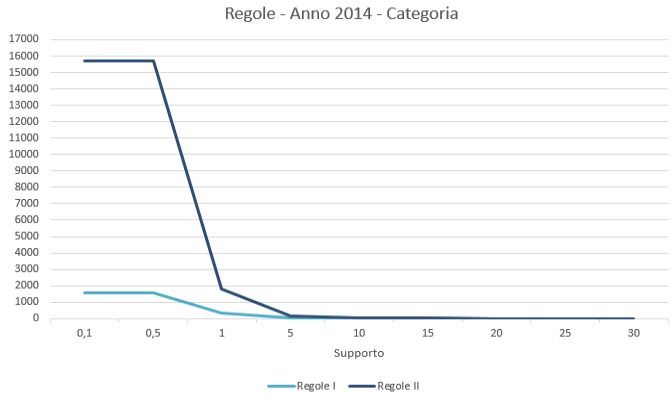


Figura 6.27 – Numero di regole associative al variare del supporto, con una confidenza del 50%, per la categoria nell'anno 2014.

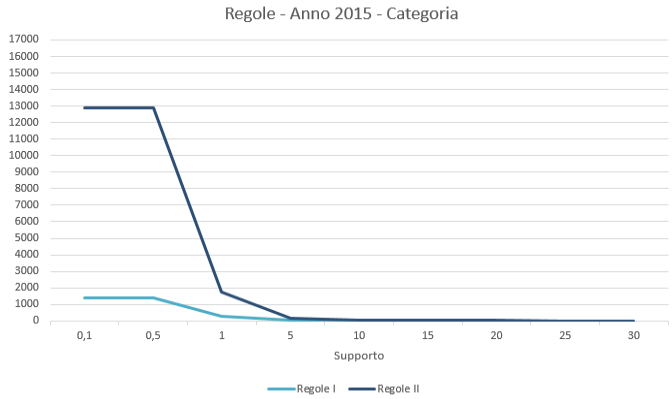


Figura 6.28 – Numero di regole associative al variare del supporto, con una confidenza del 50%, per la categoria nell'anno 2015.

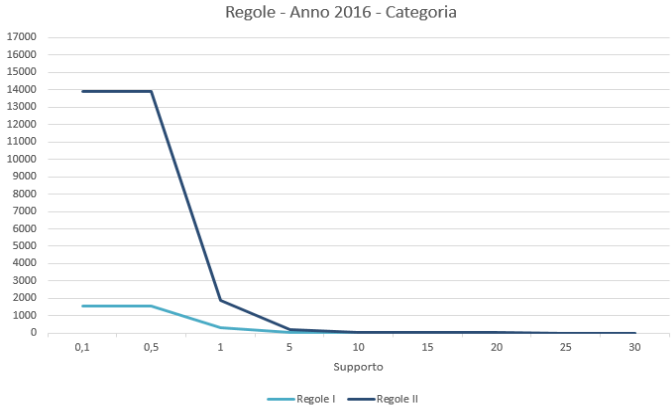


Figura 6.29 – Numero di regole associative al variare del supporto, con una confidenza del 50%, per la categoria nell'anno 2016.

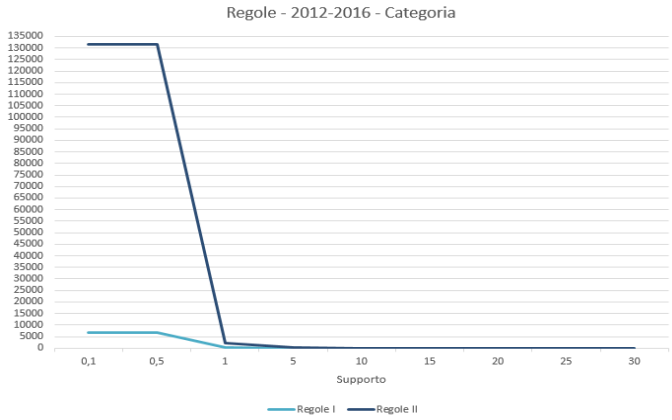


Figura 6.30 – Numero di regole associative al variare del supporto, con una confidenza del 50%, per la categoria dal 2012 al 2016.

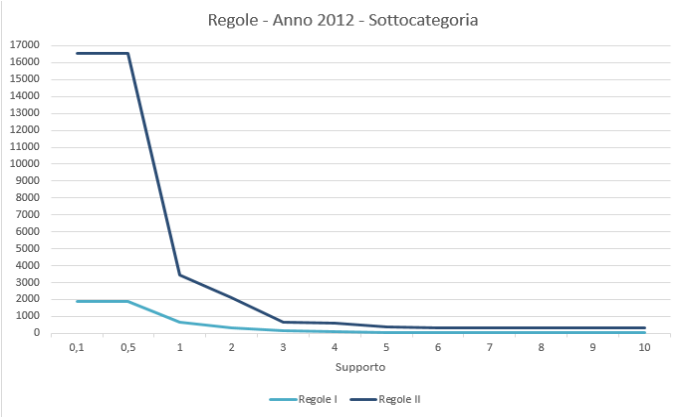


Figura 6.31 – Numero di regole associative al variare del supporto, con una confidenza del 50%, per la sottocategoria nell'anno 2012.

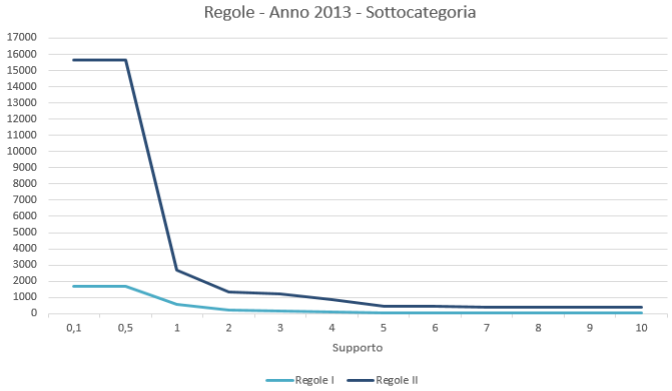


Figura 6.32 – Numero di regole associative al variare del supporto, con una confidenza del 50%, per la sottocategoria nell'anno 2013.

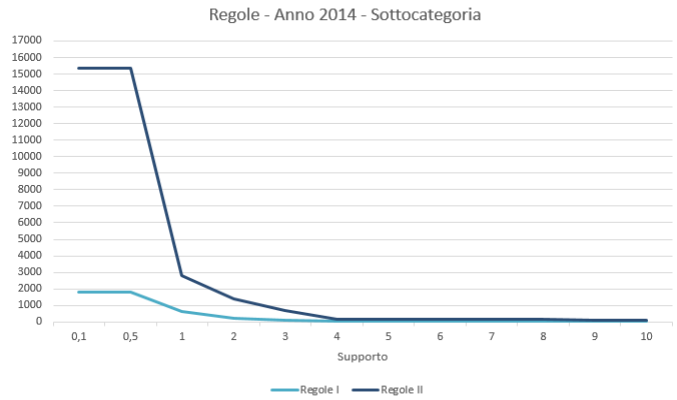


Figura 6.33 – Numero di regole associative al variare del supporto, con una confidenza del 50%, per la sottocategoria nell'anno 2014.

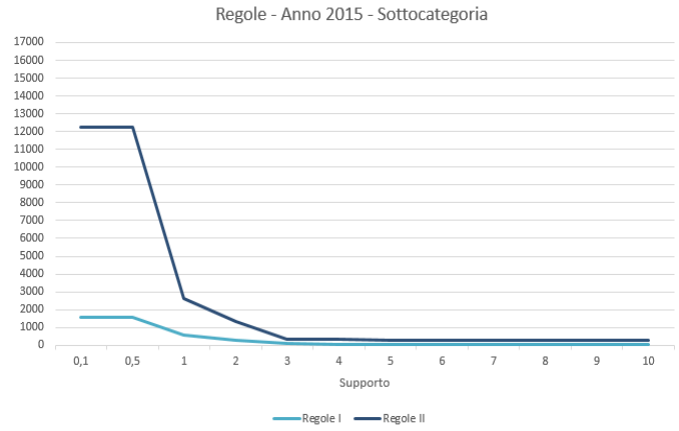


Figura 6.34 – Numero di regole associative al variare del supporto, con una confidenza del 50%, per la sottocategoria nell'anno 2015.

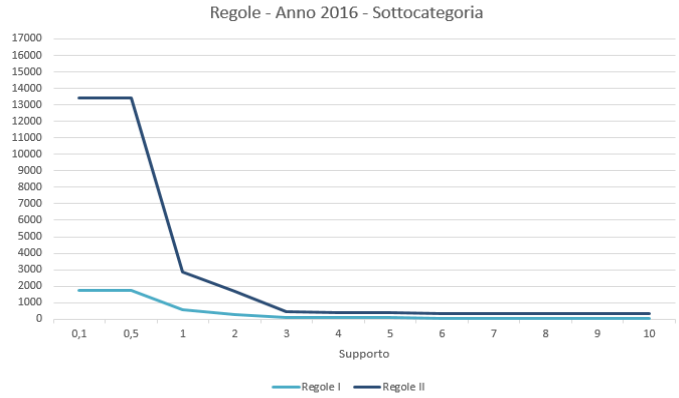


Figura 6.35 – Numero di regole associative al variare del supporto, con una confidenza del 50%, per la sottocategoria nell'anno 2016.

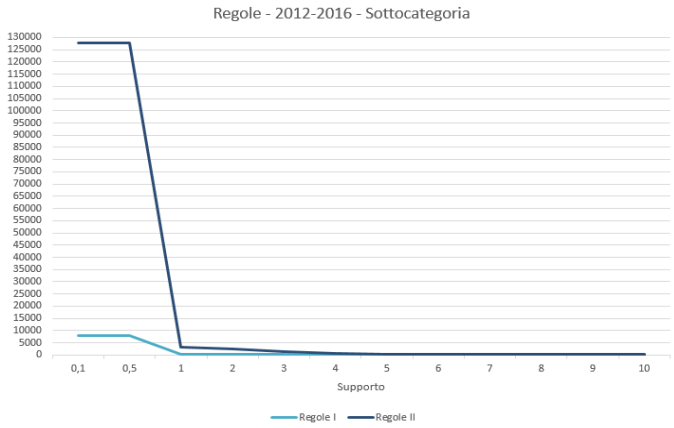


Figura 6.36 – Numero di regole associative al variare del supporto, con una confidenza del 50%, per la sottocategoria dal 2012 al 2016.

Infine, per gli ultimi 12 grafici, sono state analizzate il numero di regole associative estratte al variare della confidenza. Si è notato che l'andamento per le regole di primo e secondo livello all'aumentare della confidenza è decrescente, mentre nel caso precedente però le regole di secondo livello ad un certo supporto diminuivano drasticamente, in questo caso tenendo il supporto costante e aumentando la confidenza, la linea tende a decrescere in maniera costante.
Con confidenze più basse si ha un maggior numero di regole associative: è vero perché aumentando la confidenza vengono filtrate solamente le regole più affidabili quindi il numero di regole decrementa.
Per la categoria (figure dalla 6.37 alla 6.42), come regole di primo livello si va da un minimo di 97 per l'anno 2013 (range da 97 a 139), fino ad un massimo di 365 per l'anno 2014 (range da 295 a 365); per le regole di secondo livello si va da 330 nell'anno 2013 (range da 330 a 573), fino ad un massimo di 3376 per l'anno 2016 (range da 2737 a 3376). Considerando il dataset intero si ha per le regole di primo livello un minimo di 2 e un massimo di 366, mentre per il secondo un minimo di 1 e un massimo di 5366 regole.
Per la sottocategoria (figure dalla 6.43 alla 6.48), come regole di primo livello si va da un minimo di 346 per l'anno 2013 (range da 346 a 417), fino ad un massimo di 744 per l'anno 2012 (range da 580 a 744); per le regole di secondo livello si va da 1578 nell'anno 2015 (range da 1578 a 2381), fino ad un massimo di 5302 per l'anno 2012 (range da 4553 a 5302). Considerando il dataset intero si ha per le regole di primo livello un minimo di 212 e un massimo di 592, mentre per il secondo un minimo di 2139 e un massimo di 6727 regole.

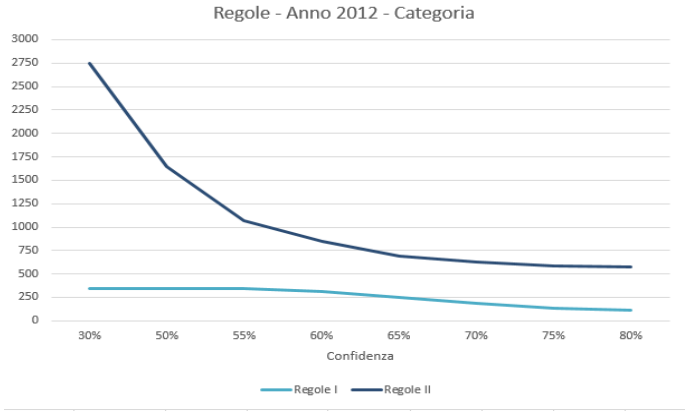


Figura 6.37 – Numero di regole associative al variare della confidenza, con un supporto di 1, per la categoria nell'anno 2012.

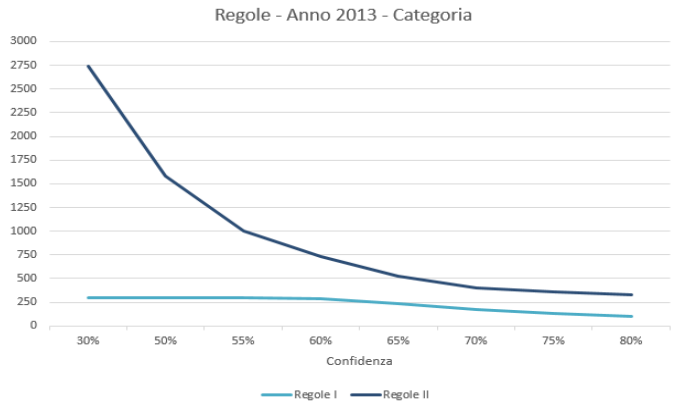


Figura 6.38 – Numero di regole associative al variare della confidenza, con un supporto di 1, per la categoria nell'anno 2013.

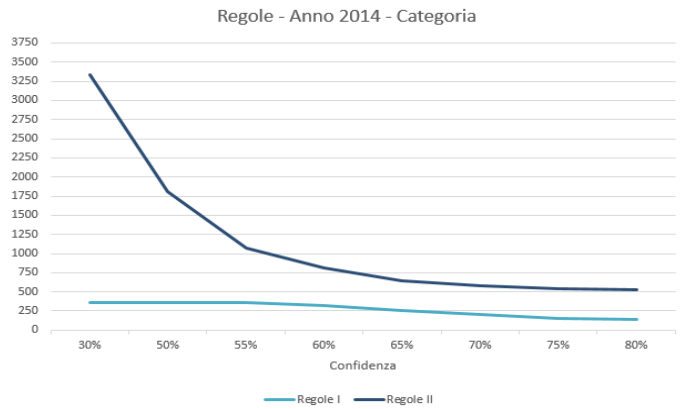


Figura 6.39 – Numero di regole associative al variare della confidenza, con un supporto di 1, per la categoria nell'anno 2014.

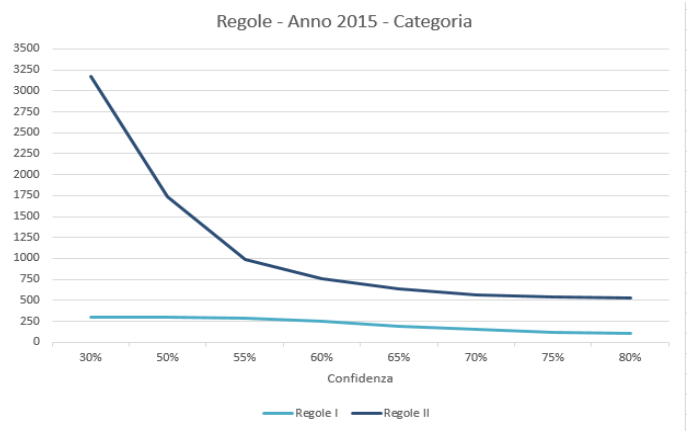


Figura 6.40 – Numero di regole associative al variare della confidenza, con un supporto di 1, per la categoria nell'anno 2015.

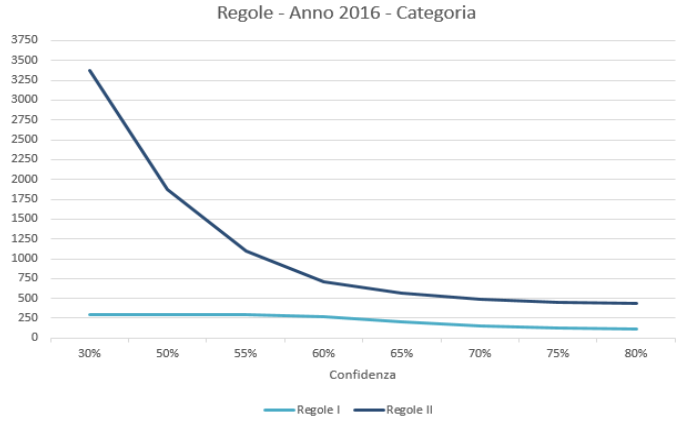


Figura 6.41 – Numero di regole associative al variare della confidenza, con un supporto di 1, per la categoria nell'anno 2016.

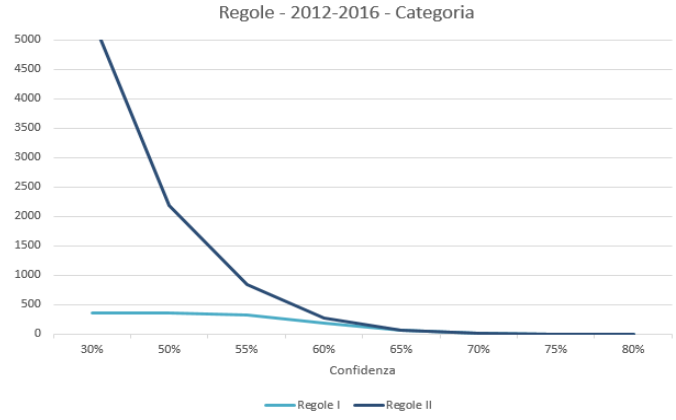


Figura 6.42 – Numero di regole associative al variare della confidenza, con un supporto di 1, per la categoria dal 2012 al 2016.

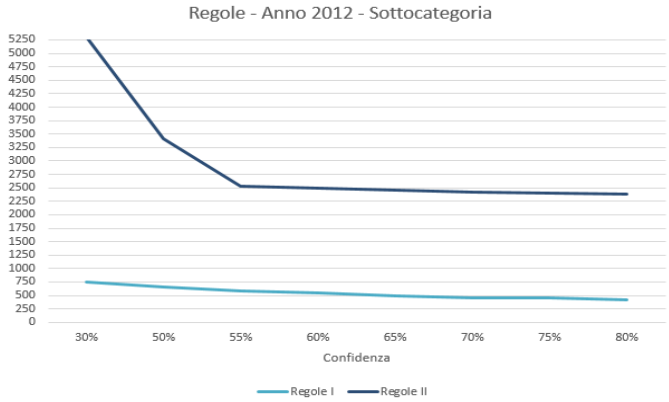


Figura 6.43 – Numero di regole associative al variare della confidenza, con un supporto di 1, per la sottocategoria nell'anno 2012.

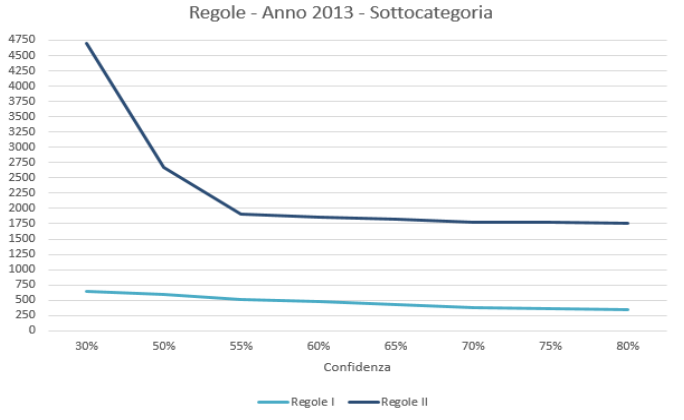


Figura 6.44 – Numero di regole associative al variare della confidenza, con un supporto di 1, per la sottocategoria nell'anno 2013.

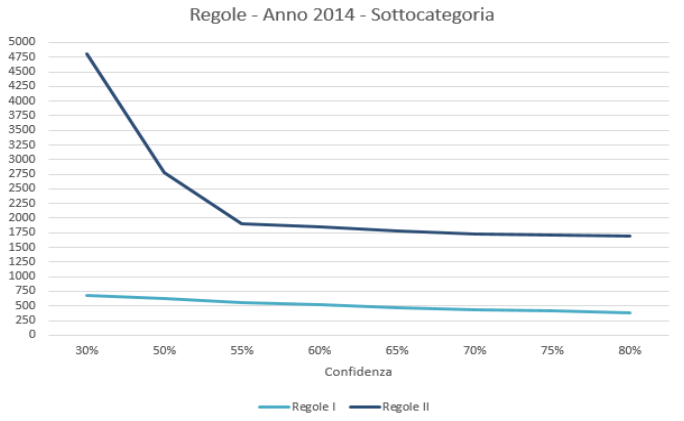


Figura 6.45 – Numero di regole associative al variare della confidenza, con un supporto di 1, per la sottocategoria nell'anno 2014.

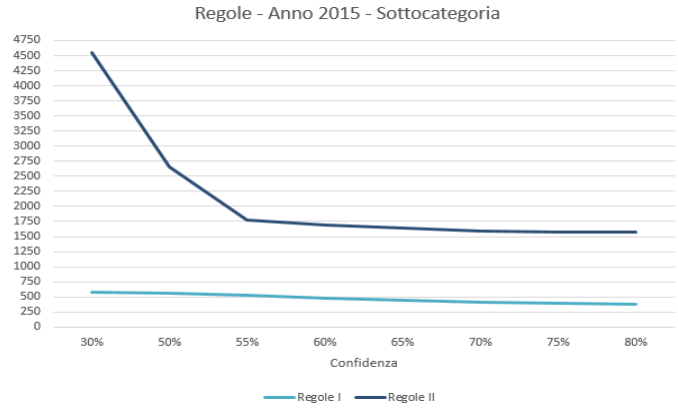


Figura 6.46 – Numero di regole associative al variare della confidenza, con un supporto di 1, per la sottocategoria nell'anno 2015.

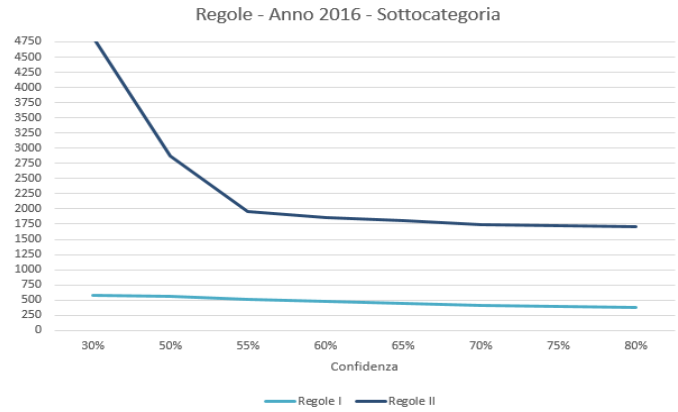


Figura 6.47 – Numero di regole associative al variare della confidenza, con un supporto di 1, per la sottocategoria nell'anno 2016.

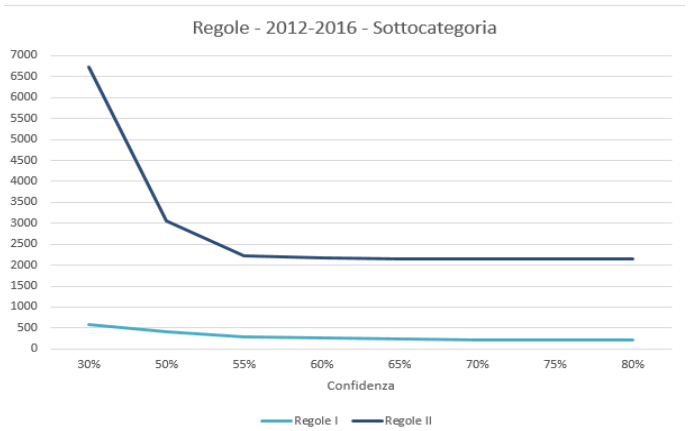
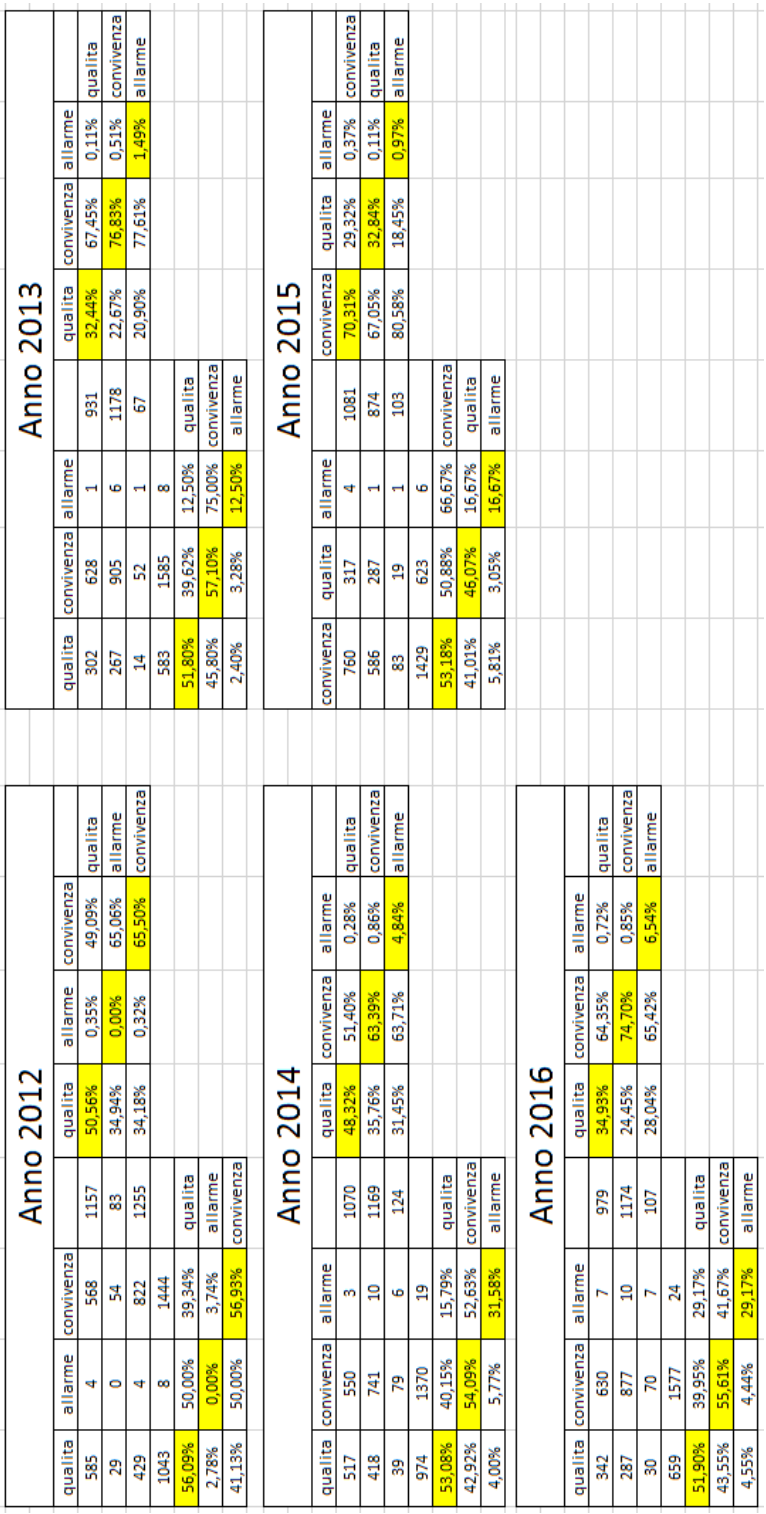


Figura 6.48 – Numero di regole associative al variare della confidenza, con un supporto di 1, per la sottocategoria dal 2012 al 2016.

6.1.1 Matrice di confusione, precisione e richiamo



Anno 2012

qualita	allarme	convivenza	
585	4	568	1157
29	0	54	83
429	4	822	1255
1043	8	1444	
56,09%	50,00%	39,34%	qualita
2,78%	0,00%	3,74%	allarme
41,13%	50,00%	56,93%	convivenza

qualita	allarme	convivenza	
50,56%	0,35%	49,09%	qualita
34,94%	0,00%	65,06%	allarme
34,18%	0,32%	65,50%	convivenza

Anno 2013

qualita	convivenza	allarme	
302	628	1	931
267	905	6	1178
14	52	1	67
583	1585	8	
51,80%	39,62%	12,50%	qualita
45,80%	57,10%	75,00%	convivenza
2,40%	3,28%	12,50%	allarme

qualita	convivenza	allarme	
32,44%	67,45%	0,11%	qualita
22,67%	76,83%	0,51%	convivenza
20,90%	77,61%	1,49%	allarme

Anno 2014

qualita	convivenza	allarme	
517	550	3	1070
418	741	10	1169
39	79	6	124
974	1370	19	
53,08%	40,15%	15,79%	qualita
42,92%	54,09%	52,63%	convivenza
4,00%	5,77%	31,58%	allarme

qualita	convivenza	allarme
48,32%	51,40%	0,28%
35,76%	63,39%	0,86%
31,45%	63,71%	4,84%

Anno 2015

convivenza	qualita	allarme	
760	317	4	1081
586	287	1	874
83	19	1	103
1429	623	6	
53,18%	50,88%	66,67%	convivenza
41,01%	46,07%	16,67%	qualita
5,81%	3,05%	16,67%	allarme

convivenza	qualita	allarme	
70,31%	29,32%	0,37%	convivenza
67,05%	32,84%	0,11%	qualita
80,58%	18,45%	0,97%	allarme

Anno 2016

qualita	convivenza	allarme	
342	630	7	979
287	877	10	1174
30	70	7	107
659	1577	24	
51,90%	39,95%	29,17%	qualita
43,55%	55,61%	41,67%	convivenza
4,55%	4,44%	29,17%	allarme

qualita	convivenza	allarme
34,93%	64,35%	0,72%
24,45%	74,70%	0,85%
28,04%	65,42%	6,54%

Figura 6.49 – Matrice di confusione, precisione e richiamo per tutti gli anni

Nella figura 6.49, possiamo notare la matrice di confusione (prime nove celle), con sotto il calcolo effettuato per ottenere la precisione e a destra il calcolo effettuato per ottenere il richiamo (che sono gli stessi che restituisce l'output del modello).
Verranno analizzati un campione di due anni, ad esempio il 2012 e il 2014.
Nell'esempio riferito all'anno 2012, si può notare che delle 1043 segnalazioni di qualità urbana reali, il sistema ne ha classificate erroneamente 458 (429 come segnalazioni di convivenza civile e 29 come segnalazioni di allarme sociale). Infatti, le segnalazioni che sono state classificate correttamente sono indicate sulla diagonale della matrice, si può velocemente osservare se il classificatore ha commesso errori o no. Stessa cosa per le segnalazioni di tipo convivenza civile: delle 1444 reali, il sistema ne ha classificate 622 erroneamente (568 come qualità urbana e 54 come allarme sociale); e come ultimo delle 8 di allarme sociale reali, il sistema le ha classificate erroneamente tutte (4 come segnalazioni di qualità urbana, 4 come convivenza civile).
Si può notare che il modello ha difficoltà a distinguere tra allarme sociale e le altre categorie e quindi nella predizione della stessa categoria; riesce bene per le altre due categorie.
Ricordando che, la precisione mi indica quanti item selezionati dal modello sono rilevanti. Quindi facendo il calcolo si è calcolata la porzione dei veri positivi sulla somma dei veri positivi più i falsi positivi. Invece, il richiamo mi indica quanti item rilevanti sono selezionati dal modello. Quindi facendo il calcolo si è calcolata la porzione dei veri positivi sulla somma dei veri positivi più i falsi negativi.
Quindi, la precisione è la probabilità che un item riportato (selezionato casualmente) sia rilevante. Il richiamo è la probabilità che un item rilevante (selezionato casualmente) venga riportato.
In questo caso (anno 2012), tra tutte le categorie, la probabilità che la qualità urbana sia rilevante è del 56,09%; che la convivenza civile sia rilevante del 56,93%; che lo sia l'allarme sociale 0%.
Invece, la probabilità che, essendo la qualità urbana rilevante, ne venga tenuto conto nell'analisi è del 50,56%; per la convivenza civile 65,50% e per l'allarme sociale 0%.
Guardando, ad esempio, l'anno 2014, si può notare che delle 974 segnalazioni di qualità urbana reali, il sistema ne ha classificate erroneamente 379 (418 come segnalazioni di convivenza civile e 39 come segnalazioni di allarme sociale). Stessa cosa per le segnalazioni di tipo convivenza civile: delle 1370 reali, il sistema ne ha classificate 629 erroneamente (550 come qualità urbana e 79 come allarme sociale); e come ultimo delle 19 di allarme sociale reali, il sistema ne ha classificate erroneamente ben 13 (3 come segnalazioni di qualità urbana, 10 come convivenza civile).
Possiamo notare dalla matrice che, in entrambi i casi (2012 e 2014), il sistema ha difficoltà a distinguere tra allarme sociale e le altre categorie, ma può fare abbastanza bene la distinzione tra qualità urbana e convivenza civile.

Inoltre, nel 2014, tra tutte le categorie, la probabilità che la qualità urbana sia rilevante è del 53,08%; che la convivenza civile sia rilevante del 54,09%; che lo sia l'allarme sociale 31,58%.
Invece, la probabilità che, essendo la qualità urbana rilevante, ne venga tenuto conto nell'analisi è del 48,32%; per la convivenza civile 63,39% e per l'allarme sociale 4,84%.

6.2 Analisi qualitativa delle regole di classificazione estratte

In questo paragrafo verranno descritte le regole associative estratte e ne verrà data un'interpretazione. Si è cercato nell'analisi di separare le regole con contesto spaziale da quelle con contesto temporale.
Si sono prese alcune regole (lunghezza 2 e 3 item) che sono più facilmente interpretabili e si sono analizzate per capire qual è il contenuto informativo. Si sono scelte le regole associative di livello 1, poiché di alta qualità rispetto a quelle di secondo livello.
Sono stati tenuti in considerazione due parametri nell'analisi: il supporto che mi filtra le regole che sono meno frequenti e la confidenza che mi serve per tenere solo le regole più affidabili; infatti le performance cambiano a seconda di come variano queste soglie.
Come punto di partenza, verrà esaminata la categoria e la sottocategoria per ogni anno.
Per quanto riguarda l'anno 2012, si può notare che come attributo di classe compare la convivenza civile. Per i luoghi, si può constatare che in particolare si ha la circoscrizione 6, nei giorni feriali e nel pomeriggio, con un supporto di 21 e una confidenza dell'80,77%. Sempre per la circoscrizione 6, nella stagione autunno con un supporto di 61 e una confidenza del 67,78% e sempre la circoscrizione 6 con le non aree verdi con un supporto di 222 e una confidenza del 60,16%. Successivamente, si ha la circoscrizione 2, nel mese di marzo e nei giorni feriali, con una confidenza del 70,59% e un supporto di 12; la circoscrizione 7, nel mese di marzo con una confidenza del 65% e un supporto di 13; la circoscrizione 10, in generale nei giorni feriali, con una confidenza del 60% e un supporto di 24; infine la circoscrizione 5, nei giorni feriali e di martedì con una confidenza del 54,55% e un supporto di 18. Tutte queste regole implicano l'attributo di classe "convivenza civile".
Spostando l'attenzione verso il contesto temporale, si ha che la maggior parte delle segnalazioni è avvenuta di giorno mercoledì nelle aree verdi con 66,67% di confidenza e un supporto di 12; inoltre con un 60,53% di confidenza ed un supporto di 23 si ha che, sempre di mercoledì, le segnalazioni sono maggiormente avvenute nel mese di marzo, e nelle regole successive compare anche il mese di marzo con la stagione inverno ma si era precedentemente supposto che marzo è un mese invernale, perciò tutto torna.
Guardando alla sottocategoria, sempre per l'anno 2012, si ha come attributo di classe la sottocategoria rumori molesti, facente parte della macrocategoria convivenza civile, mentre le altre regole riguardano il decoro e il degrado urbano, che appartiene alla qualità urbana. Si può notare che la circoscrizione 6 nelle non aree verdi, soprattutto in via Caprie (quartiere Cit Turin, circoscrizione 3), implicano la sottocategoria rumori molesti; questa

regola ha un supporto di 6 e una confidenza del 75%. Le altre tre regole implicano il decoro e degrado urbano e tutte concentrano attributi i quali non area verde, giorno feriale ma con tre vie della città differenti: via Sempione (tra le zone Barriera di Milano e Regio Parco nella circoscrizione 6) con un supporto di 6 e una confidenza del 75%, corso Palermo (tra le zone Barriera di Milano e Aurora, tra le circoscrizioni 6 e 7) con un supporto di 3 e una confidenza del 75%, via Chiomonte (Borgo San Paolo nella circoscrizione 3) con un supporto di 3 e una confidenza del 60%.
Si può notare che è possibile estrarre un'informazione da quanto appena analizzato.
Come citato precedentemente, la circoscrizione 6 è correlata maggiormente con la sottocategoria rumori molesti e decoro e degrado urbano. Infatti, come afferma TorinoToday [43], un quotidiano online che tratta le cronache e le notizie principali riguardanti la città e i suoi quartieri, nell'articolo intitolato "Barriera di Milano: bivacchi e schiamazzi durante la notte 'Basta! Chiudete questi locali'" dell'11 febbraio 2012 [44], la zona Barriera di Milano appartenente alla circoscrizione 6 è una zona rumorosa e grazie al via vai di gente per via dei locali la pulizia è abbastanza scarsa: precisamente i residenti sono costretti a ripulire i marciapiedi o l'uscio di casa propria per via della troppa sporcizia.
Tra i risultati compare anche via Sempione, che si sdoppia tra le zone Barriera di Milano e Regio Parco. Anche qui, avendo usato la funzionalità Street View di Google Maps [45], si può notare che il degrado non è basso: i marciapiedi si stanno rompendo per via degli alberi che crescono sotto ed è pieno di graffiti sui muri, che non sono mai stati cancellati.
Per quanto riguarda, invece, l'anno 2013, si può trovare maggiormente l'attributo di classe qualità urbana. Approfonditamente, la regola che riguarda le non aree verdi, nella circoscrizione 8 e in corso Ferraris (che si sdoppia tra la circoscrizione 8 e la circoscrizione 1, nello specifico tra i quartieri di San Salvario e Crocetta), concentrata più sulla circoscrizione 8 che sulla 1, ha un supporto di 11 e una confidenza di ben 100%. Si ritrova poi la circoscrizione 8 la maggior parte di volte nel giorno venerdì, con una confidenza dell'88,46% e un supporto di 23, sempre la circoscrizione 8 nel giorno lunedì, con una confidenza del 71,79% e un supporto di 28 e infine nella stagione Autunno con una confidenza del 71,43% e un supporto di 30. Invece, con una confidenza del 93,75% e un supporto di 15, si trova nelle non aree verdi e nei giorni feriali, via Filadelfia (quartiere Santa Rita nella circoscrizione 2). Con 93,33% di confidenza e un supporto di 14, sempre nelle non aree verdi e nei giorni feriali, si trova via Cigna (che si sdoppia tra le zone Aurora e Barriera di Milano, tra le circoscrizioni 6 e 7). Con un 90,91% di confidenza ed un supporto di 10, nei giorni feriali e nel mese di marzo si ha via Pettinati (quartiere Nizza Millefonti, circoscrizione 9). Con un 80% di confidenza e un supporto di 16, nelle non aree verdi, corso Verona (zona Aurora, circoscrizione 7); con un 78,79% di confidenza e un supporto di 26, la circoscrizione 1 nel mese di marzo e nelle non aree verdi; con una confidenza del 71,43% e un supporto di 10, la circoscrizione 2 nei giorni feriali nel mese di settembre.
Per quanto riguarda il contesto temporale spiccano i mesi di giugno, luglio (in particolare a stagione estiva), e, con una confidenza minore, il mese di novembre.
Tutte queste regole implicano la qualità urbana.

Andando invece a vedere più approfonditamente la sottocategoria, l'attributo di classe predominante per quanto riguarda il contesto della qualità urbana è quello dei veicoli abbandonati, mentre per quanto riguarda la convivenza civile si trovano rumori molesti e disturbi da locali.
Infatti la regola che riguarda le non aree verdi, nella circoscrizione 8 e in via IV Marzo (quartiere Centro nella circoscrizione 1), che implica l'attributo di classe veicoli abbandonati, ha la confidenza maggiore rispetto a tutte le altre regole, 81,82%, e un supporto di 9.
Andando ad analizzare prima le regole con confidenza maggiore, per quanto riguarda la convivenza civile, la regola che riguarda la circoscrizione 1, nelle aree verdi e nei giorni feriali che implica rumori molesti, con 76,92% di confidenza e un supporto di 10; nelle non aree verdi e nei giorni feriali, in via San Marino (zona Santa Rita nella circoscrizione 2) che implica disturbi da locali, con una confidenza del 70% ed un supporto di 7; la circoscrizione 3 nei giorni feriali e, nel dettaglio, in via Perosa (zona Cenisia nella circoscrizione 3) che implica disturbi da locali, con una confidenza del 64,29% ed un supporto di 9; con una confidenza più bassa sempre per quanto riguarda i disturbi da locali, si trova la circoscrizione 7 nel mese di dicembre, la circoscrizione 5 di mattina nelle non aree verdi e la circoscrizione 3 in inverno nei giorni feriali. Invece, la regola che implica i rumori molesti è la circoscrizione 5 nelle aree verdi nei giorni feriali, con una confidenza di 55,56% e un supporto di 5.
Per quanto riguarda la fascia temporale si hanno risultati relativi all'attributo di classe rumori molesti soprattutto nel mese di settembre e dicembre, nella fascia oraria pomeridiana.
Anche qui è possibile estrarre un'informazione da questi risultati: ad esempio i famosi Murazzi di Torino, che fanno parte della circoscrizione 1, erano una zona rumorosissima, come si può leggere anche dall'articolo "Murazzi: sequestrati impianti sonori di due locali" del 14 maggio 2013 [46], che tratta di abusi edilizi con lo scopo di fare festa, e quindi disturbare i residenti con rumori fino a tarda notte.
Anche nella circoscrizione 5 i residenti si lamentano del rumore e dei disturbi da locali, principalmente per una discoteca, dall'articolo "I clienti della discoteca escono alle 7 del mattino, residenti esasperati" del 16 luglio 2017 [47]; la data dell'articolo è recente ma è citato che in passato era già stata organizzata una raccolta firme e si vorrebbe replicare poiché i rumori sono spalmati in tutta la notte compresa anche la mattina.
Inoltre, in zona Santa Rita (circoscrizione 2), è presente il Pala Alpitur, dove si tengono frequentemente concerti e può essere oggetti di segnalazioni da magari gente anziana residente nei dintorni, che è costretta a subire disturbi di questo tipo, oppure sempre lì vicino è presente lo Stadio Olimpico dove quasi ogni fine settimana si tengono le partite e potrebbero creare disturbi i cori, le trombette, e così via.
Per quanto riguarda il 2014, spicca la qualità urbana come attributo di classe predominante, e, in questo anno, sono moltissime le regole importanti. Le principali, partendo da una confidenza più alta, sono: con l'87,50% di confidenza ed un supporto di 14, la circoscrizione 1 d'estate e in corso Novara (che si sdoppia tra le zone Aurora e Barriera di Milano, tra le circoscrizioni 6 e 7); con un 75,26% e un supporto di 22, la

circoscrizione 8 nelle non aree verdi di lunedì e in largo Palermo (che si sdoppia tra le zone Aurora e Barriera di Milano, tra le circoscrizioni 6 e 7); via Carmagnola (zona Aurora, tra le circoscrizioni 6); via Mercadante (tra le zone Aurora e Regio Parco, nella circoscrizione 6) nella stagione estiva; la circoscrizione 7 d'estate nella fascia oraria mattiniera; la circoscrizione 6 in autunno, e andando nel dettaglio, nel mese di novembre.
Per quanto riguarda la sottocategoria, la regola con confidenza più alta, riguarda la macrocategoria convivenza civile, nello specifico, rumori molesti. Questa regola dice che la circoscrizione 1 e in via Lagrange (zona Centro, circoscrizione 1) implica rumori molesti con una confidenza dell'80% e un supporto di 4. Successivamente, si ha la circoscrizione 3 nel mese di dicembre e nelle non aree verdi che implica decoro e degrado urbano con una confidenza del 66,67% e un supporto di 8; sempre per quanto riguarda il decoro e degrado urbano, la circoscrizione 5 nel mese di marzo, la circoscrizione 4 nella stagione autunnale, la circoscrizione 7 nella fascia oraria pomeridiana, la circoscrizione 7 nel giorno giovedì. Inoltre, la circoscrizione 8 per quanto riguarda il decoro e degrado urbano, compare spesso, nei giorni giovedì e lunedì, e la circoscrizione 6 nella fascia oraria mattiniera e nel giorno venerdì. Passando alla sottocategoria veicoli abbandonati, si ha con il 66,67% di confidenza ed un supporto di 4, la circoscrizione 4 nella stagione inverno al pomeriggio; mentre con una confidenza più bassa, del 50%, e un supporto di 5, sempre la circoscrizione 4 nelle aree verdi e nei giorni feriali.
Per quanto riguarda il discorso temporale e il decoro e degrado urbano, sono ricorrenti le regole che contengono l'attributo autunno, mentre per i veicoli abbandonati, la maggior parte delle segnalazioni sono state fatte nella stagione invernale.
Quindi, l'informazione che si può estrarre dal 2014, è che in via Lagrange, in pieno centro, i locali sono aperti fino a tarda sera con conseguenti rumori che possono dar fastidio ai residenti; in questa via, infatti, sono presenti dei pub "gettonati" dai giovani come ad esempio il "Jumping Jester", che trasmette anche partite di calcio e chiude tutti i giorni 3 di notte oppure il "One Apple Concept Bar", dove è presente un dj set dove si può andare a fare aperitivo, e chiude alle 2 di notte. Anche la circoscrizione 7, in particolare la zona di Vanchiglia, è molto rumorosa, come citato dall'articolo "In Vanchiglia la movida disturba i residenti" del 17 settembre 2014 [48], che tratta del sorgere di nuovi locali frequentati dagli universitari fino a notte fonda, e ciò rende difficile la convivenza con quella parte di residenti che ormai giovani non sono più.
Ulteriormente, per quanto riguarda il decoro e degrado urbano, si può citare la circoscrizione 8, in particolare la zona di San Salvario, dove, come citato nell'articolo "Degrado a San Salvario, siringhe abbandonate per giorni lungo le strade" del 19 novembre 2014 [49], è stata ritrovata una siringa in mezzo alla strada e nessuno si è operato per rimuoverla, mettendo a rischio passanti o addirittura bambini.
Per quanto riguarda l'anno 2015, compare come attributo di classe predominante la categoria allarme sociale. La regola in cui compare via Quincinetto (zona Madonna di Campagna, circoscrizione 5) nel giorno mercoledì è presente con una confidenza del 70,59% e un supporto di 12; sempre nella circoscrizione 5 e in corso Grosseto (che attraversa diverse zona tra cui Madonna di Campagna e Borgo Vittoria, tutte nella circoscrizione 5), con una confidenza del 69,23% e un supporto di 9; lungo Stura Lazio

(zona Barca nella circoscrizione 6) nei giorni feriali e nelle non aree verdi, che implica sempre allarme sociale, con un 63,64% di confidenza e un supporto di 14; infine, la circoscrizione 5 e la circoscrizione 8, con una confidenza del 61,54% e un supporto di 8. Per quanto riguarda la sottocategoria, si possono trovare ben due regole che implicano gli atti di vandalismo, facenti parte della macrocategoria allarme sociale. In particolare, si può notare la circoscrizione 7 e in via Magenta (zona Crocetta nella circoscrizione 1) nella stagione invernale con un 80% di confidenza e un supporto di 4; mentre la circoscrizione 4 nelle aree verdi con un 64,29% di confidenza e un supporto di 9. Per quanto riguarda invece le regole appartenenti alla macrocategoria convivenza civile, si possono notare ben 4 regole con il 100% di confidenza, tra cui:

- la circoscrizione 10 nei giorni feriali e nelle non aree verdi, che implica disturbi da locali con un supporto di 3
- la circoscrizione 10 nei giorni feriali e nelle non aree verdi, che implica rumori molesti con un supporto di 4
- la circoscrizione 1 nei giorni feriali e nelle non aree verdi, che implica disturbi da locali con un supporto di 2
- la circoscrizione 4 nei giorni feriali e nelle non aree verdi, che implica rumori molesti con un supporto di 2.

In generale, tra le altre regole che implicano i rumori molesti, si ha anche la circoscrizione 3, sempre nei giorni feriali e nelle non aree verdi, con una confidenza del 66,67% e un supporto di 2; mentre con una confidenza minore, compare anche la circoscrizione 6 sempre nei giorni feriali e nelle non aree verdi.

Quindi è possibile estrarre un'informazione anche nel 2015 da questi risultati: ad esempio a proposito di atti di vandalismo, si è trovato un articolo intitolato "La circoscrizione Sette è dovuta intervenire per reprimere una serie di furti e atti vandalici" del 2 aprile 2015 [50], che tratta di bulletti che entrano nella piscina Colletta durante le ore notturne, commettendo reati. Sempre rimanendo il tema atti di vandalismo, spostandosi nella circoscrizione 4, un altro articolo intitolato "Parco Dora, i cittadini vogliono le telecamere di sorveglianza contro la criminalità" del 29 settembre 2015 [51], tratta del volere da parte dei cittadini di mettere telecamere di sorveglianza con lo scopo di contrastare spaccio di droga, scippi e furti. Inoltre, il Parco Dora è oggetto di rumore costante a causa del centro commerciale, infatti nei risultati la circoscrizione 4 è correlata anche con i rumori molesti. Per quanto riguarda quest'ultimi, come già citato precedentemente, continuano i disturbi nella circoscrizione 6, maggiormente nei quartieri di Barriera di Milano e Villaretto. Sono stati trovati due articoli che affermano ciò, il primo "Basta con la movida selvaggia, in Barriera è guerra alle aperture notturne" del 9 febbraio 2015 [52], è un disperato appello da parte dei residenti che chiedono una moratoria contro quei locali che tengono aperto notte e giorno provocando un continuo disturbo alla quiete pubblica; il secondo "Villaretto chiede barriere contro i rumori della tangenziale" del 15 dicembre 2015 [53], che tratta dell'inquinamento rumoroso causato dalla tangenziale nord e i residenti sostengono che non riescono nemmeno più a dormire la notte a causa di questo.

Guardando all'ultimo anno, il 2016, si può notare che l'attributo di classe predominante è quello riguardante le segnalazioni di tipo qualità urbana. Infatti, tra le regole, si ha la

circoscrizione 1 e via Baveno (zona Parella nella circoscrizione 4) e la circoscrizione 2 e la circoscrizione 8 entrambe nelle non aree verdi, ed entrambe con una confidenza circa dell'84% ed un supporto la prima di 11 e la seconda di 16. Inoltre, nella circoscrizione 8, con una confidenza intorno al 76%, le segnalazioni avvengono frequentemente nel giorno sabato, e di venerdì nella stagione estiva, e nel mese di giugno. Nel mese di giugno, con una confidenza minore, del 65%, e un supporto di 13, questo tipo di segnalazioni si possono trovare anche nella circoscrizione 4 e in via Nizza (zona Nizza Millefonti nella circoscrizione 9) e, come ultima regola, nella circoscrizione 3 nel giorno di mercoledì.
Per quanto riguarda la sottocategoria, compaiono moltissime regole con confidenza alta e di buona qualità, infatti si andranno a vedere solamente quelle con una confidenza almeno del 65%. L'attributo di classe predominante in questo anno è veicoli abbandonati, facente parte della macrocategoria qualità urbana. Analizzando la prima regola, si ha nei giorni feriali, nella circoscrizione 9 e in via Guastalla (zona Vanchiglia nella circoscrizione 7) con una confidenza dell'81,82% e un supporto di 9; con più o meno gli stessi valori di confidenza e supporto si ha la circoscrizione 7 e corso De Gasperi (zona Crocetta nella circoscrizione 1); e la circoscrizione 9 e corso Regina Margherita (grandissimo corso di Torino che attraversa diverse zone che fanno parte delle circoscrizioni 4 e 5) con un supporto di 17 e una confidenza del 73,91%.
Con confidenze più basse, tra il 65% e il 68%, sempre per quanto riguarda i veicoli abbandonati, si trova la circoscrizione 7 che è presente in ben 3 regole soprattutto nel giorno mercoledì, nei giorni feriali e nelle non aree verdi. Come ultima regola esaminata, con una confidenza del 70% e un supporto di 7 si ha la circoscrizione 10 nei giorni feriali e in via Casalis (che attraversa la zona Cit Turin nella circoscrizione 3 e la zona San Donato nella circoscrizione 4) che implica l'attributo di classe comportamenti molesti, facente parte della categoria convivenza civile.
Infine, dai risultati ottenuti in questo anno si può capire che, per quanto riguarda i veicoli abbandonati, si interpreta che la circoscrizione 9 sia il cuore di questa sottocategoria. Infatti, da un articolo intitolato "Controlli auto abbandonate di via Dina, vigili sequestrano 22 veicoli" del 18 febbraio 2015 [54], si nota che le zone Lingotto e Mirafiori sono le più gettonate per abbandonare la propria auto, infatti ne sono state trovate 22.
In questo paragrafo, si è quindi cercato di dare un'interpretazione alle regole associative risultanti e si può concludere che dai risultati, ovvero dalle regole estratte e filtrate, si possono estrarre informazioni utili per l'interpretazione dell'analisi svolta.

7. Conclusioni

L'obiettivo di questa tesi è stata l'analisi di dati relativi alla sicurezza urbana con lo scopo di trasformarli in informazione. I dati trattati sono open data, quindi facilmente consultabili ed analizzabili. Dal lavoro di tesi riguardante esplorazione degli open data relativi alla sicurezza urbana è emerso che in Italia si stanno sempre ampliando di anno in anno; si è concluso che le nazioni con più open data resi disponibili sono Canada e Stati Uniti per quanto riguarda il contesto extra-europeo e Regno Unito e Francia per quanto riguarda quello europeo.
I dati analizzati sono quelli relativi alla città di Torino. Si è scelto di analizzare le segnalazioni fatte dai cittadini al Contact Center della Polizia Municipale e si è valutato di utilizzare l'algoritmo FP-Growth attraverso l'uso del software Rapidminer e di creare le regole di associazione.
Dalle analisi effettuate con questo algoritmo è emerso che, per tutte le tipologie di regole analizzate, le categorie che dominano sono la convivenza civile e la qualità urbana. In particolare le sottocategorie disturbi da locali e rumori molesti per la prima, e decoro e degrado urbano e veicoli abbandonati per la seconda. È stata svolta una analisi temporale, suddividendo per anno, e si è trovato che per quanto riguarda i disturbi da locali, i luoghi più soggetti sono la zona centro, la zona crocetta e san Salvario fino al 2014, poi la zona si è spostata totalmente sul quartiere di san Salvario negli ultimi anni dell'analisi. Per quanto riguarda i rumori molesti, si nota solamente un picco nell'anno 2015 precisamente nella zona di Borgo San Paolo, zona Cenisia, Cit Turin e zona Pozzo Strada. Nel contesto qualità urbana, soprattutto la sottocategoria decoro e degrado urbano rimane costante negli anni, ma cambia sempre di luogo. Nel 2012 e 2013 è localizzata nella zona del centro e nella Crocetta, per poi spostarsi nel 2013 e nel 2014 nelle zone Vanchiglia, Aurora, Sassi e Madonna del Pilone. Nel 2015 torna a comparire nella zona del centro e della Crocetta per poi spostarsi nel 2016 in zona borgo Vittoria, Madonna di Campagna, Lucento e Vallette. Per i veicoli abbandonati, invece, si hanno solo segnalazioni nei primi anni: nel 2012-2013 nelle zone Barriera di Milano, Regio Parco, Barca, Bertolla, Falchera, Rebaudengo, Villaretto; nel 2014 nelle zone Borgo San Paolo, Cenisia, Pozzo Strada, Cit Turin, Borgata Lesna e San Donato, Campidoglio, Parella. Negli ultimi anni questo tipo di segnalazione non compare nell'analisi, il che può voler dire che si ha un minor abbandonamento di macchine con il passare del tempo, anche se questo periodo temporale è solo un campione, quindi non si può affermare con certezza.
Successivamente, si è deciso di utilizzare un'altra tecnica basata su un classificatore associativo con lo scopo di rendere l'analisi più accurata possibile; sono stati studiati approfonditamente i parametri dell'algoritmo e si sono scelti quelli che avrebbero portato a migliori performance. Da ciò è emerso che è possibile estrarre un'informazione dai dati estratti, che era la premessa di questo elaborato. Infatti, si può notare dall'analisi che la circoscrizione 6, e maggiormente il quartiere di Barriera di Milano, è abbastanza rumoroso per la presenza di gente e locali; a causa di quest'ultimi, la pulizia è abbastanza

scarsa, il degrado risulta alto, le strade sono tenute sporche e i residenti sono costretti a ripulire il proprio uscio di casa propria. Si è notata la stessa cosa per i disturbi da locali negli anni successivi per la circoscrizione 1, dove sono presenti i Murazzi, in cui nel 2013 sono stati sequestrati due impianti sonori per il troppo disturbo ai residenti e in pieno centro in via Lagrange, dove i locali sono aperti fino a tarda notte e creano disturbo ai residenti; inoltre, anche nella circoscrizione 2, comprendente il PalaAlpitour e lo Stadio Olimpico, nella circoscrizione 5 e nella circoscrizione 7, quartiere principale della movida rumorosa è quello di Vanchiglia. Si è notato ulteriormente che, per quanto riguarda il decoro e degrado urbano, la circoscrizione 8, in particolare la zona di San Salvario, non risulta molto pulita, e si sono trovate siringhe abbandonate accanto ai marciapiedi.
Spiccano gli atti di vandalismo nella circoscrizione 4 e nella circoscrizione 7, quest'ultima comprendente il Parco Dora, dove nel 2015 i cittadini hanno firmato una petizione per far installare delle telecamere di sorveglianza per la troppa criminalità.
Nel 2015, nella circoscrizione 6, viene firmata una petizione per l'installazione di barriere per l'inquinamento acustico in zona Villaretto causato dalla tangenziale nord di Torino.
Infine, non va tralasciata la situazione delle auto abbandonate in circoscrizione 9.
Quindi, da tutto ciò, si è dimostrato che si possono estrarre informazioni utili dai risultati ottenuti.
L'analisi svolta in questo elaborato è stata effettuata prendendo in considerazione le circoscrizioni di Torino e le vie, ma non i quartieri. Un possibile sviluppo futuro sarebbe quello di considerare, come dati di partenza, i singoli quartieri, così da notare qual è il quartiere che spicca maggiormente in una determinata circoscrizione, analizzando per esempio la via della città e associandola ad un determinato quartiere.
Anche nel contesto temporale è possibile guardare al futuro, analizzando e suddividendo i dati per periodi di tempo più brevi dell'anno, ad esempio il semestre o addirittura il mese dell'anno.
L'analisi, in futuro, potrebbe essere svolta utilizzando nuovi algoritmi e magari anche nuovi software per notare se i risultati ottenuti convergono.
Per concludere, i risultati ottenuti e quelli ottenibili tramite altri algoritmi, potrebbero portare in futuro ad una vera e propria raccolta, schematizzazione di questi dati, aggiornati il più frequente possibile, consultabili ad esempio attraverso pagine web, applicazioni mobili per cellulari e così via, rendendo il cittadino partecipe di ciò che lo circonda, in tempo reale.

Bibliografia e sitografia

[1] *Open Data comune di Torino*, http://aperto.comune.torino.it/, consultato il 21/09/2017

[2] *Open Data comune di Torino, dataset "Segnalazioni al Contact Center della Polizia Municipale"*, http://aperto.comune.torino.it/, consultato il 21/09/2017

[3] *Open Data comune di Torino, dataset "Violazioni ai Regolamenti Comunali",* http://aperto.comune.torino.it/, consultato il 21/09/2017

[4] *Open Data comune di Torino, dataset "Violazioni al Codice della Strada",* http://aperto.comune.torino.it/, consultato il 21/09/2017

[5] *Agenzia per l'Italia Digitale, articolo sugli open data in Italia,* https://www.dati.gov.it/content/quanti-sono-open-data-italia, consultato il 14/11/2017

[6] *Open Data comune di Roma,* http://dati.comune.roma.it/, consultato il 15/11/2017

[7] *Open Data comune di Milano*, http://dati.comune.milano.it/, consultato il 15/11/2017

[8] *Open Data comune di Firenze*, http://opendata.comune.fi.it/, consultato il 15/11/2017

[9] *Articolo di "Open Data in the G8",* http://www2.datainnovation.org/2015-open-data-g8.pdf, consultato il 15/11/2017

[10] *Open Data città di Londra,* https://data.london.gov.uk/dataset/london-fire-brigade-incident-records, consultato il 15/11/2017

[11*] Open Data città di Parigi,* https://opendata.paris.fr/explore/?sort=modified&refine.theme=Equipements,+Services,+Social, consultato il 15/11/2017

[12] *Open Data Francia, ferrovie SNCF,* https://ressources.data.sncf.com/explore/dataset/incidents-securite/table/?sort=-esr&dataChart=eyJxdWVyaWVzIjpbeyJjaGFydHMiOlt7InR5cGUiOiJjb2x1bW4iLCJmdW5jIjoiQ09VTlQiLCJjb2xvciI6InJhbmdlLUFjY2VudCIsInNjaWVudGlmaWNEaXNwbGF5Ijp0cnVlfV0sInhBeGlzIjoiZGF0ZSIsIm1heHBvaW50cyI6IiIsInRpbWVzY2FsZSI6InllYXIiLCJzb3J0IjoiIiwic2VyaWVzQnJlYWtkb3duIjoiZXNyIiwic3RhY2tlZCI6Im5vcm1hbCIsImNvbmZpZyI6eyJkYXRhc2V0IjoiaW5jaWRlbnRzLXNlY3VyaXRlIiwib3B0aW9ucyI6eyJzb3J0IjoiLWVzciJ9fX1dLCJ0aW1lc2NhbGUiOiIifQ%3D%3D, consultato il 15/11/2017

[13] *Open Data città di Vancouver, Canada,* http://data.vancouver.ca/datacatalogue/, consultato il 15/11/2017

[14] *Open Data città di Surrey, Canada,* http://data.surrey.ca/group/health-and-safety, consultato il 15/11/2017

[15] *Open Data città di Edmonton, Canada,* https://data.edmonton.ca/browse?anonymous=true&q=EPS%20Neighbourhood%20Criminal%20Incidents&sortBy=relevance, consultato il 15/11/2017

[16*] Open Data città di Montreal, Canada,* http://donnees.ville.montreal.qc.ca/group/loi-justice-securite-publique, consultato il 15/11/2017

[17] *Open Data città di Toronto, Canada,* https://www1.toronto.ca/wps/portal/contentonly?vgnextoid=8bf6e03bb8d1e310VgnVCM10000071d60f89RCRD, consultato il 15/11/2017

[18] *Open Data città di Calgary, Canada,* https://data.calgary.ca/browse?category=Government&q=safety&sortBy=relevance, consultato il 15/11/2017

[19] *Open Data città di Halifax e Nuova Scozia, Canada,* https://data.novascotia.ca/browse?anonymous=true&q=Crime+Statistics+-+Incidents+and+rates+for+selected+offences&sortBy=relevance, consultato il 15/11/2017

[20] *Open Data città di New York, Stati Uniti d'America,* https://data.cityofnewyork.us/browse?category=Public+Safety&provenance=official, consultato il 15/11/2017

[21] *Open Data città di San Francisco, Stati Uniti d'America,* https://data.sfgov.org/browse?category=Public+Safety, consultato il 15/11/2017

[22] *Open Data città di Los Angeles, Stati Uniti d'America,* https://data.lacity.org/browse?category=A+Safe+City, consultato il 15/11/2017

[23] *Open Data città di Las Vegas, Stati Uniti d'America,* https://opendata.lasvegasnevada.gov/browse?category=Public+Safety&limitTo=datasets, consultato il 15/11/2017

[24] *Open Data città di Seattle, Stati Uniti d'America,* https://data.seattle.gov/browse?category=Public+Safety&provenance=official, consultato il 15/11/2017

[25] *Open Data città di Denver, Stati Uniti d'America,* https://www.denvergov.org/opendata/dataset/city-and-county-of-denver-crime, consultato il 15/11/2017

[26] *Open Data città di Chicago, Stati Uniti d'America,* https://data.cityofchicago.org/browse?category=Public%20Safety, consultato il 15/11/2017

[27] *Open Data città di Houston, Stati Uniti d'America,* http://data.houstontx.gov/group/public-health, consultato il 15/11/2017

[28] *Open Data città di Orlando, Stati Uniti d'America,* https://data.cityoforlando.net/browse?category=Public+Safety, consultato il 15/11/2017

[29] *Open Data città di Detroit, Stati Uniti d'America,* https://data.detroitmi.gov/browse?category=Public+Safety, consultato il 15/11/2017

[30] *Open Data città di Philadelphia, Stati Uniti d'America,* https://www.opendataphilly.org/group/public-safety-group, consultato il 15/11/2017

[31] *Open Data città di Washington, Stati Uniti d'America,* https://data.wa.gov/browse, consultato il 15/11/2017

[32] *Open Data città di Boston, Stati Uniti d'America,* https://data.cityofboston.gov/browse?anonymous=true&category=Public+Safety&limitTo=datasets&q=safety&sortBy=relevance, consultato il 15/11/2017

[33] *Open Data città di Tokyo, Giappone,* http://www.data.go.jp/data/en/dataset?groups=gr_1500, consultato il 15/11/2017

[34] *Open Data città di Moscow, Russia,* http://data.gov.ru/taxonomy/term/10/datasets, consultato il 15/11/2017

[35] *Open Data città di Canberra, Australia,* https://www.data.act.gov.au/browse?category=Justice%2C+Safety+and+Emergency, consultato il 15/11/2017

[36] *Materiale didattico del corso Business Intelligence del Politecnico di Torino,* http://dbdmg.polito.it/wordpress/teaching/business-intelligence/, consultato il 20/01/2018

[37] *Rapidminer,* https://rapidminer.com/, consultato il 01/12/2017

[38] *Manuale d'uso per Rapidminer,* https://docs.rapidminer.com/downloads/RapidMiner-v6-user-manual.pdf, consultato il 01/12/2017

[39] Tan P.N., Steinbach M., Kumar V., *Introduction to Data Mining*, Pearson International Edition, 2006.

[40] *Weka,* https://www.cs.waikato.ac.nz/ml/weka/index.html, consultato il 18/01/2018

[41] *Il classificatore associativo,* http://dbdmg.polito.it/wordpress/research/associative-classification/, consultato il 18/01/2018

[42] *Le dieci circoscrizioni amministrative di Torino,* http://www.comune.torino.it/statistica/osservatorio/annuario/2002/pdf/03_Territorio.pdf, consultato il 21/09/2017

[43] *TorinoToday, il quotidiano online di Torino,* http://www.torinotoday.it/, consultato il 12/03/2018

[44] *Articolo di TorinoToday "Barriera di Milano: bivacchi e schiamazzi durante la notte 'Basta! Chiudete questi locali'" dell'11 febbraio 2012,* http://www.torinotoday.it/cronaca/degrado-barriera-milano-furti-schiamazzi-notturni.html, consultato il 12/03/2018

[45] *Street View Google Maps,* https://www.google.it/maps/, consultato il 12/03/2018

[46] *Articolo di TorinoToday "Murazzi: sequestrati impianti sonori di due locali" del 14 maggio 2013,* http://www.torinotoday.it/cronaca/murazzi-sequestrati-impianti-sonori.html, consultato il 12/03/2018

[47] *Articolo di TorinoToday "I clienti della discoteca escono alle 7 del mattino, residenti esasperati" del 16 luglio 2017,* http://www.torinotoday.it/cronaca/schiamazzi-discoteca-stradella-baldissera.html, consultato il 12/03/2018

[48] *Articolo di TorinoToday "In Vanchiglia la movida disturba i residenti" del 17 settembre 2014,* http://www.torinotoday.it/cronaca/vanchiglia-movida-problemi.html, consultato il 12/03/2018

[49] *Articolo di TorinoToday "Degrado a San Salvario, siringhe abbandonate per giorni lungo le strade" del 19 novembre 2014,* http://www.torinotoday.it/cronaca/siringa-abbandonata-via-saluzzo-corso-raffaello.html, consultato il 12/03/2018

[50] *Articolo di TorinoToday "La circoscrizione Sette è dovuta intervenire per reprimere una serie di furti e atti vandalici" del 2 aprile 2015,* http://www.torinotoday.it/cronaca/ladri-piscine-vanchiglia-cartelli.html, consultato il 12/03/2018

[51] *Articolo di TorinoToday "Parco Dora, i cittadini vogliono le telecamere di sorveglianza contro la criminalità" del 29 settembre 2015,* http://www.torinotoday.it/cronaca/parco-dora-telecamere-di-sorveglianza.html, consultato il 12/03/2018

[52] *Articolo di TorinoToday "Basta con la movida selvaggia, in Barriera è guerra alle aperture notturne" del 9 febbraio 2015,* http://www.torinotoday.it/cronaca/la-proposta-dei-residenti-una-moratoria-per-i-locali-di-barriera.html, consultato il 12/03/2018

[53] *Articolo di TorinoToday "Villaretto chiede barriere contro i rumori della tangenziale" del 15 dicembre 2015,* http://www.torinotoday.it/cronaca/inquinamento-rumore-tangenziale-villaretto.html, consultato il 12/03/2018

[54] *Articolo di TorinoToday "Controlli auto abbandonate di via Dina, vigili sequestrano 22 veicoli" del 18 febbraio 2015,* http://www.torinotoday.it/cronaca/controlli-auto-abbandonate-via-dina.html, consultato il 12/03/2018