Politecnico di Torino

Corso di Laurea Magistrale in Ingegneria Informatica

Caratterizzazione dei livelli di rischio di polizze assicurative mediante estrazione di regole di associazione

Relatore
Prof. Luca Cagliero
Correlatore
Prof.ssa Elena Maria Baralis

Candidato
Antonino Scalone
Matricola: 231908

Dicembre 2018

Ringraziamenti

Desidero ringraziare i professori Luca Cagliero ed Elena Maria Baralis per aver reso possibile lo sviluppo di questo lavoro con la loro professionalità e costante disponibilità.

Vorrei ringraziare la mia famiglia, mia madre e mia sorella Martina, che hanno supportato la mia permanenza a Torino sostenendomi giorno dopo giorno affinchè potessi raggiungere questo traguardo. Grazie a mio padre, che penso mi guidi ancora da lassù.

Inoltre un ringraziamento va ai miei amici, nuovi e ritrovati, per aver reso questa esperienza ancora più importante.

Indice

1 Introduzione **1**

2 Descrizione dello scenario **4**
2.1 Concetto di scatola nera 4
2.2 Dataset iniziale 5
2.2.1 Analisi semantica dei dati reperiti dalle scatole nere . . 13
2.2.2 Analisi semantica dei dati relativi agli assicurati 14
2.3 Scenario 16

3 Data Mining **18**
3.1 Introduzione 18
3.2 Data Mining: concetti base 19
3.2.1 Tipi di dati e modelli di archiviazione 20
3.2.2 Strategie di Data Mining 22
3.2.3 Tecniche di classificazione 23
3.2.4 Regressione 25
3.2.5 Tecniche descrittive 26
3.2.6 Clustering 27
3.2.7 Estrazione di regole di associazione 29
3.3 Knowledge Discovery in Databases 31
3.3.1 Data Cleaning e principali problemi 32
3.3.2 Data Integration e Data Transformation: problemi e soluzioni 34
3.4 La classificazione associativa 35
3.4.1 L'algoritmo Live and Let Live 36

4 Analisi dei dati telematici e aziendali **39**
4.1 Struttura del processo applicativo 39
4.2 Il processo di Knowledge Discovery 41
4.2.1 Selezione delle feature 42
4.2.2 Preprocessing dei dati 47

4.2.3 Trasformazione dei dati 50
4.3 Caso di studio 52
4.3.1 Definizione delle categorie di feature 53
4.3.2 Generazione dei profili caratterizzanti i livelli di rischio 56
4.3.3 Interpretazione dei pattern e definizione dei KPI 58

5 Risultati sperimentali **60**
5.1 Approccio 60
5.2 Risultati ricavati 61

6 Conclusioni e sviluppi futuri **69**

Elenco delle figure

2.1 *Data Integration* 5
3.1 *Overview relativa agli step che compongono il processo KDD [7]* 31
4.1 *Struttura del processo di analisi* 40

Elenco delle tabelle

2.1 *Dati aggregati su base periodica relativi alle informazioni reperite dalle black box* 6
2.2 *Possibili dati relativi a contraenti di polizze assicurative* 10
3.1 *Ipotetico insieme di scontrini rilasciati da un supermarket* . . 30
4.1 *Selezione delle feature ritenute importanti per l'analisi* 43
4.2 *Schema del dataset generato mediante aggregazione settimanale* 50
4.3 *Schema del dataset generato mediante aggregazione mensile* . . 51
4.4 *Schema del dataset generato mediante aggregazione annuale* . 51
4.5 *Sottoinsiemi delle feature per ciascuna categoria* 54
5.1 *Selezione di alcuni risultati ottenuti* 61

Capitolo 1

Introduzione

Con il termine *black box* (scatola nera) in ambito assicurativo si indica un dispositivo telematico capace di reperire informazioni utili circa il veicolo in viaggio e il relativo comportamento del conducente. La scatola nera è principalmente fornita di un rilevatore gps ed ha l'obiettivo primario di riproporre in maniera dettagliata la dinamica esatta di un sinistro, nel caso in cui ne avvenga uno, servendosi di dati quali la forza dell'urto e la velocità dei veicoli coinvolti nell'impatto. Per anni, la scatola nera è stata utilizzata esclusivamente da mezzi di trasporto quali navi e aerei, in modo da tenere traccia di alcune informazioni come l'altitudine, la velocità e così via. I servizi che la scatola nera riesce a fornire al giorno d'oggi sono aumentati a dismisura con l'aggiunta di sensori e moduli che permettono di rilevare tutto ciò che è inerente allo stile di guida dell'assicurato, come ad esempio, consentono di registrare eventi di accelerazioni, decelerazioni e i dettagli relativi alle percorrenze, basandosi sui vari tempi di marcia e tempi di sosta rilevati dalla scatola e distinguendo inoltre i molteplici tipi di strada su cui l'assicurato viaggia.

Le compagnie assicurative hanno mostrato forte interesse, soprattutto negli ultimi anni, nell'agevolare l'installazione della scatola nera a bordo degli autoveicoli, con la volontà principale di eliminare il grave problema delle frodi assicurative e delle truffe relative ai sinistri, cercando in questo modo di limitare i costi a carico delle compagnie e quindi fornire sconti agli assicurati, relativamente alle polizze Rc auto. Tra i vantaggi principali nell'utilizzo della scatola nera a bordo, spiccano la possibilità di rintracciare un autoveicolo in caso di furto attraverso il localizzatore e la possibilità di personalizzare le polizze in base alla condotta dell'assicurato, integrando quest'ultima con i dati relativi alle caratteristiche dei clienti e alla loro situazione assicurativa.

L'obiettivo del lavoro di tesi è analizzare i dati relativi a un insieme di polizze assicurative associate a veicoli e caratterizzarle in base al loro livello

di rischio di causare sinistri. L'analisi sarà basata su una collezione integrata di dati storici, relativi alle polizze, che includono:

1. le caratteristiche principali delle polizze e dei veicoli associati;
2. i percorsi effettuati dai veicoli assicurati;
3. gli stili di guida associati ai guidatori dei veicoli assicurati;
4. i sinistri commessi dai veicoli.

I dati ai punti (2) e (3) sono raccolti da scatole nere installate sui veicoli.

Nel corso degli anni, la quantità di dati che popola la rete e che risiede negli archivi delle aziende è aumentata notevolmente. I volumi di dati sono diventati molto consistenti e le sorgenti dati sono aumentate in numero e tipologia, come ad esempio dati derivanti da sensori, dati telefonici, dati finanziari e così via. Nasce così la necessità di acquisire informazioni in maniera rapida e di gestire dataset il cui volume è così grande da richiedere l'introduzione di nuovi strumenti di analisi, dato che la capacità dei database relazionali non soddisfa più le esigenze di immagazzinamento e analisi di queste moli. I nuovi strumenti e le nuove metodologie che vengono introdotti si pongono l'obiettivo di estrapolare e gestire informazioni nel minor tempo possibile, con la capacità di supportare ed assistere gli utenti nell'estrazione di informazioni, probabilmente non immediate, dai dati, trasformandoli in un secondo momento in conoscenza organizzata. Questa necessità ha portato alla nascita del *Data Mining.*

Il data mining [4, 8, 9] è l'insieme delle tecniche che si occupano di estrarre informazioni utili da grandi moli di dati, principalmente non ordinati o grezzi, attraverso strumenti e tecniche automatizzate o semi-automatizzate che permettono di ricavare relazioni e pattern inizialmente sconosciuti all'interno delle basi di dati. Il concetto di data mining si basa fondamentalmente su due tipi di approcci, quali approccio supervisionato e approccio non supervisionato. La prima tipologia di approccio consiste nell'utilizzo di algoritmi che tendono ad analizzare elementi noti in base al contesto applicativo per poi elaborare classificazioni e/o predizioni sul futuro, tenendo in considerazione un insieme considerevole di fattori. I modelli predittivi, in generale, si pongono come obiettivo principale quello dell'apprendimento, in modo tale che la conoscenza acquisita attraverso dati noti, strutturati e già etichettati nel cosiddetto training set, possa essere utile e altamente performante nel contesto di analisi futuri. La seconda tipologia di approccio invece non basa l'apprendimento su dati già etichettati, ma si pone l'obiettivo di ricavare

pattern significativi dal punto di vista concettuale e descrittivo, attraverso l'analisi esclusiva dei dati in esame.

Sulla base di quanto detto, in questo lavoro di tesi sono state identificate le correlazioni più significative tra i dati raccolti e i livelli di rischio nel medio-lungo termine, con lo scopo di arricchire la base di dati aziendale con i dati della telematica e trarre le correlazioni più rilevanti all'interno dei dati per ogni livello di rischio, distinguendo vari contesti in base alla tipologia di polizze e alla durata contrattuale, come ad esempio, polizze stipulate nell'anno in analisi e della durata di un anno (polizze nuove/annuali) o semestrali (polizze nuove/semestrali) e così via.

La tecnica consiste nell'applicazione di un classificatore basato su regole di associazione, denominato L^3 (*Live and Let Live*) [3]. L'utilizzo del classificatore L^3 permette la generazione di un modello interpretabile, in modo da etichettare dati non noti e/o estrarre i pattern più significativi per un'analisi mirata in base al contesto applicativo. Le regole di associazione [1, 14] rappresentano le co-occorrenze più rilevanti presenti all'interno di una collezione di dati potenzialmente molto ampia. Dato che le regole di associazione possono rappresentare correlazioni arbitrarie tra combinazioni di feature, sarà selezionata solo la parte di co-occorrenze presente all'interno dei dati pertinenti al caso in esame, ovvero tra i dati inerenti alle caratteristiche degli assicurati e al loro stile di guida da un lato, e i dati relativi alle classi di rischio dall'altro. I passi principali che hanno definito questo lavoro di tesi sono:

1. definizione dei livelli di rischio ed etichettatura delle polizze;
2. raccolta, analisi e preparazione dei dati telematici e relativa integrazione con i dati aziendali;
3. estrazione di regole di associazione caratterizzanti ciascuna classe di rischio.

I principali risultati ottenuti riguardano l'estrazione di pattern significativi la caratterizzazione delle polizze e la definizione di Key Performance Indicators (KPIs) utili a caratterizzare i servizi telematici e a monitorare la correlazione tra i dati delle scatole nere e quelli dei sinistri sotto vari aspetti.

Capitolo 2

Descrizione dello scenario

2.1 Concetto di scatola nera

Nel corso degli anni, le compagnie assicurative hanno incentivato l'installazione della scatola nera a bordo dei veicoli dei propri assicurati, in modo tale da monitorare i dati dei clienti ed evitare il problema delle frodi assicurative. La black box è un dispositivo satellitare equipaggiato fondamentalmente di un localizzatore gps che permette di rilevare grandi moli di dati circa i dettagli sullo stile di guida dei propri clienti ed eventuali sinistri e/o infrazioni commesse. Al giorno d'oggi esistono varie versioni di scatole nere che spaziano in base a forma, dimensione ed elementi che ne caratterizzano le specifiche tecniche. L'adozione delle black box non è ancora obbligatoria in Italia, ma le compagnie assicurative cercano di invogliare gli automobilisti ad installarla garantendo degli sconti sulle polizze Rc auto. Tra i principali vantaggi che derivano dall'utilizzo delle scatole nere è possibile citare la possibilità di tracciare la dinamica dettagliata di un ipotetico sinistro, il ritrovamento del mezzo in caso di furto e la fondamentale novità della personalizzazione delle polizze auto in base alla propria condotta di guida.

Il lavoro di tesi si prefigge l'obiettivo di fornire un supporto alle compagnie assicurative nel valutare e valorizzare i dati reperiti dalle scatole nere per una migliore gestione e analisi, esplorando i suddetti dati per ricavarne un valore aggiunto attraverso l'applicazione di tecniche di data mining [4, 8, 9] e determinare l'effettiva conoscenza estraibile e i servizi realizzabili. Per poter applicare al meglio il classificatore adottato in questo lavoro, è stato necessario definire un processo di preparazione dei dati in esame, in modo tale da gestire eventuali problematiche legate all'integrazione di più sorgenti o alla natura dei dati, nonchè alla loro struttura per renderla idonea e compatibile con le tecniche di data mining.

2.2 Dataset iniziale

I dati reperiti dalle black box installate sui veicoli degli assicurati forniscono al giorno d'oggi informazioni importanti e dettagliate che permettono di definire la tracciabilità e la rintracciabilità dei veicoli stessi. Le informazioni reperite sono principalmente inerenti alla condotta e alle percorrenze degli assicurati, allo stato del veicolo e ad altre informazioni in base al tipo di scatola nera che la compagnia assicurativa decide di adottare. Per quanto riguarda la condotta dei guidatori è possibile tenere traccia dello stile di guida dei clienti, nonchè delle infrazioni effettuate, differenziate in base al tipo di strada e alla fascia oraria. Le percorrenze invece includono informazioni circa i chilometri che i guidatori percorrono durante la validità della polizza, differenziate anch'esse in base alla tipologia di strada. I suddetti dati sono integrati con le caratteristiche delle polizze, in modo tale da avere per ogni record sia i dettagli sullo stile di guida e sulle percorrenze, sia le caratteristiche relative ai dettagli anagrafici dell'assicurato e al suo stato assicurativo. Il concetto di *Data Integration*, rappresentato in Fig. 2.1, è essenziale per fornire una visione unificata di più database. Lo scopo è quello di ottenere un unico archivio completo di dati, in modo tale da garantire un processo di analisi il più dettagliato possibile.

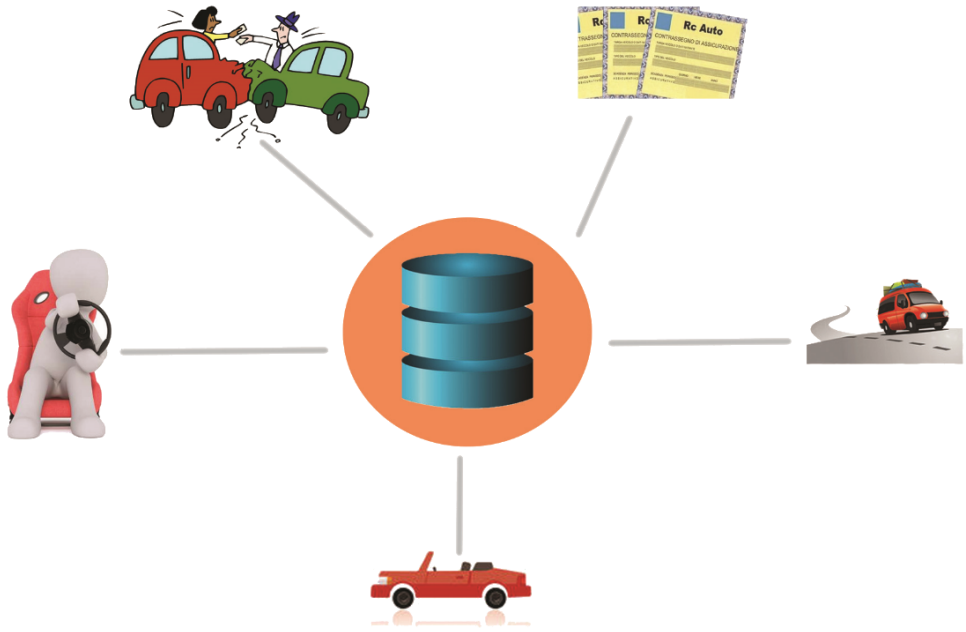

Figura 2.1: *Data Integration*

Nella Tabella 2.1 e nella Tabella 2.2 vengono rappresentate rispettivamente alcune feature reperite potenzialmente da una scatola nera e alcune possibili informazioni relative agli assicurati.

Tabella 2.1: *Dati aggregati su base periodica relativi alle informazioni reperite dalle black box*

Categoria	**Attributo**
Aggregazioni temporali	-Giorno rilevamento dati -Data inizio del periodo di osservazione -Ora inizio del periodo di osservazione -Data fine del periodo di osservazione -Ora fine del periodo di osservazione -Giorno -Settimana -Anno -Data del sinistro -Codice della settimana
Percorrenze in metri	-Numero totale di corse -Numero totale di metri -Metri percorsi in autostrada -Metri percorsi in città -Metri percorsi in periferia -Metri percorsi in altre strade -Metri percorsi nella fascia diurna -Metri percorsi nella fascia notturna -Metri percorsi il lunedì -Metri percorsi il martedì -Metri percorsi il mercoledì -Metri percorsi il giovedì -Metri percorsi il venerdì -Metri percorsi il sabato -Metri percorsi la domenica -Numero totale di metri percorsi

Continua nella prossima pagina

Continua dalla pagina precedente

Categoria	**Attributo**
	durante il giorno -Numero totale di km percorsi durante il giorno -% metri percorsi in autostrada -% metri percorsi in città -% metri percorsi su strade extra-urbane -% metri percorsi altrove -% metri percorsi durante la fascia diurna -% metri percorsi durante la fascia notturna -% metri percorsi nei giorni feriali -% metri percorsi nei giorni festivi
Percorrenze in secondi	-Numero totale di secondi -Secondi trascorsi in autostrada -Secondi trascorsi in città -Secondi trascorsi in periferia -Secondi trascorsi in altre strade -Secondi in viaggio il lunedì -Secondi in viaggio il martedì -Secondi in viaggio il mercoledì -Secondi in viaggio il giovedì -Secondi in viaggio il venerdì -Secondi in viaggio il sabato -Secondi in viaggio la domenica -Secondi nella fascia diurna -Secondi nella fascia notturna
Stile di guida	-Eccessi di velocità in città durante la fascia diurna -Eccessi di velocità in autostrada durante la fascia diurna -Eccessi di velocità su strade extra-urbane durante la fascia diurna -Eccessi di velocità altrove

Continua nella prossima pagina

Continua dalla pagina precedente

Categoria	Attributo
	durante la fascia diurna -Eccessi di velocità in città durante la fascia notturna -Eccessi di velocità in autostrada durante la fascia notturna -Eccessi di velocità su strade extra-urbane durante la fascia notturna -Eccessi di velocità altrove durante la fascia notturna -Accelerazioni brusche in città durante la fascia diurna -Accelerazioni brusche in autostrada durante la fascia diurna -Accelerazioni brusche su strade extra-urbane durante la fascia diurna -Accelerazioni brusche altrove durante la fascia diurna -Accelerazioni brusche in città durante la fascia notturna -Accelerazioni brusche in autostrada durante la fascia notturna -Accelerazioni brusche su strade extra-urbane durante la fascia notturna -Accelerazioni brusche altrove durante la fascia notturna -Decelerazioni brusche in città durante la fascia diurna -Decelerazioni brusche in autostrada durante la fascia diurna -Decelerazioni brusche su strade extra-urbane durante la fascia diurna -Decelerazioni brusche altrove durante la fascia diurna

Continua nella prossima pagina

Continua dalla pagina precedente

Categoria	Attributo
	-Decelerazioni brusche in città durante la fascia notturna -Decelerazioni brusche in autostrada durante la fascia notturna -Decelerazioni brusche su strade extra-urbane durante la fascia notturna -Decelerazioni brusche altrove durante la fascia notturna -Curvature ad alta velocità in città durante la fascia diurna -Curvature ad alta velocità in autostrada durante la fascia diurna -Curvature ad alta velocità su strade extra-urbane durante la fascia diurna -Curvature ad alta velocità altrove durante la fascia diurna -Curvature ad alta velocità in città durante la fascia notturna -Curvature ad alta velocità in autostrada durante la fascia notturna -Curvature ad alta velocità su strade extra-urbane durante la fascia notturna -Curvature ad alta velocità altrove durante la fascia notturna -Cambi repentini di direzione in città durante la fascia diurna -Cambi repentini di direzione in autostrada durante la fascia diurna -Cambi repentini di direzione su strade extra-urbane durante la fascia diurna -Cambi repentini di direzione

Continua nella prossima pagina

Continua dalla pagina precedente

Categoria	Attributo
	altrove durante la fascia diurna -Cambi repentini di direzione in città durante la fascia notturna -Cambi repentini di direzione in autostrada durante la fascia notturna -Cambi repentini di direzione su strade extra-urbane durante la fascia notturna -Cambi repentini di direzione altrove durante la fascia notturna -Numero assoluto circa lo stile di guida -Numero relativo circa lo stile di guida

Si conclude dalla pagina precedente

Tabella 2.2: *Possibili dati relativi a contraenti di polizze assicurative*

Categoria	Attributo
Informazioni sulle polizze	-Numero di polizza -Inizio copertura della polizza -Fine copertura della polizza -Data di scadenza polizza -Nuova polizza -Durata polizza (Semestrale o Annuale) -Tipo di polizza -Mese di scadenza della polizza
Informazioni black box	-ID black box del veicolo -Black box a bordo -Tipo di black box
Informazioni sui veicoli	-Cavalli fiscali -Cavalli potenza -Codice veicolo

Continua nella prossima pagina

Continua dalla pagina precedente

Categoria	**Attributo**
	-Numero massimo passeggeri a bordo -Alimentazione del veicolo -Tipo di veicolo -Utilizzo del veicolo -Anno di prima immatricolazione -Data di prima immatricolazione -Età del veicolo -Data dell'ultima voltura -Anni di possesso del veicolo -Gancio per auto
Informazioni assicurati	-Codice sesso -Provincia -Area territoriale -Regione -Codice geografico -CAP -Sesso -Anni di patente -Professione -Ritardo ottenimento patente -Età dell'assicurato
Stato clienti	-Classe di merito -Età della patente -Massimale -Classe agevolata -Numero di rinnovi -Anni senza assicurazione -Attestato di rischio -Distanza dall'ultimo sinistro -Tipo di danno -Forfait che l'assicurazione deve pagare per causa sinistro -Forfait ottenuto -Cifra pagata ma non recuperata -Cifra recuperata

Continua nella prossima pagina

Continua dalla pagina precedente

Categoria	Attributo
	-Anni di rischio -Costo complessivo -Cifra pagata -Riservato -Rinuncia alla surrogazione -Compagnia -Classe di merito interna -Flag proprietario-contraente -Costo degli incidenti con colpa grave causati -Numero di incidenti con colpa grave causati -Costo degli incidenti senza colpa grave causati -Numero di incidenti senza colpa grave causati -Costo degli incidenti subiti, pagati dalla compagnia -Forfait ottenuto dalla compagnia debitrice -Numero di incidenti subiti dagli assicurati per i quali la compagnia paga la cifra -Numero di incidenti subiti dagli assicurati per i quali la compagnia ottiene il forfait dalla compagnia debitrice -Numero totale di sinistri (causati e subiti) -Costo complessivo prima della gestione dei dati -Numero di incidenti causati -Costo degli incidenti causati -Codice aziendale

Si conclude dalla pagina precedente

2.2.1 Analisi semantica dei dati reperiti dalle scatole nere

Nella Tabella 2.1, i primi attributi indicano il periodo di osservazione dei dati inerenti agli assicurati che hanno adottato le black box, specificando sia la data esatta di inizio e fine periodo con gli attributi "Data inizio del periodo di osservazione" e "Data fine del periodo di osservazione", sia l'intervallo orario esatto con "Ora inizio del periodo di osservazione" e "Ora fine del periodo di osservazione". L'attributo "Giorno rilevamento dati" indica la data in cui sono stati prelevati alcuni dati e determina più record per la stessa polizza, dovuti all'integrazione dei database in cui i dettagli relativi alla condotta di guida hanno granularità giornaliera. Una sezione molto importante riguarda gli attributi che descrivono le percorrenze degli assicurati. La granularità dei dati relativi alle percorrenze, in termini di metri percorsi e secondi trascorsi in percorrenza, consiste nel riportare sia i dettagli giorno per giorno, nell'arco di una settimana alla volta, sia il numero complessivo settimanale. L'attributo "Numero totale di metri" indica il totale complessivo di metri percorsi durante la settimana di osservazione di una determinata polizza.

Le black box in adozione da parte delle compagnie sono capaci di riconoscere la tipologia di strada relativamente ai dati delle percorrenze e la fascia oraria che separa la percorrenza durante il giorno da quella durante la notte, distinguendo principalmente 6 categorie:

1. urbana;
2. suburbana;
3. autostrada;
4. altro;
5. diurna;
6. notturna.

Un altro attributo ritenuto importante per l'analisi è "Settimana", che indica il numero della settimana di osservazione rispetto al totale delle settimane in un anno. I dati inerenti allo stile di guida degli assicurati sono suddivisi allo stesso modo delle percorrenze, cioè tenendo conto della tipologia di strada e della fascia oraria in cui gli eventi sono stati reperiti. Gli attributi relativi a questa categoria comprendono numerose infrazioni, quali:

- eccessi di velocità;

- accelerazioni brusche;
- decelerazioni brusche;
- cambi repentini di direzione;
- curvature ad alta velocità.

2.2.2 Analisi semantica dei dati relativi agli assicurati

I dati inerenti alle proprietà degli assicurati rappresentano dettagli significativi dei clienti che stipulano un contratto di polizza con un'azienda assicurativa. Le quattro principali categorie che meglio rappresentano i numerosi attributi presenti nel dataset, sono le seguenti:

1. dettagli anagrafici del cliente;
2. dati relativi all'auto posseduta dal cliente;
3. informazioni sullo stato assicurativo del cliente;
4. informazioni sui sinistri e attestato di rischio.

I dettagli anagrafici dei clienti includono informazioni, quali età, sesso, provincia di residenza, professione e così via. Le situazioni in cui il contraente risulta diverso da colui che utilizza l'auto, sono regolate dall'attributo "Flag proprietario-contraente", impostato a *true* o a *false* in base alle condizioni. Per quanto riguarda i dati relativi all'autoveicolo ritenuti più significativi, è possibile citare gli attributi:

- età del veicolo;
- anno della prima immatricolazione;
- cavalli fiscali del veicolo;
- cavalli potenza del veicolo;
- numero totale di persone che è possibile trasportare in auto;
- alimentazione del veicolo.

La terza categoria racchiude tutti gli attributi che indicano le caratteristiche delle polizze e la condotta degli assicurati nel corso degli anni. Gli attributi principali, utilizzati nell'analisi, riguardano:

- classe di merito;
- numero della polizza;
- inizio e fine copertura della polizza;
- durata della polizza;
- nuova polizza;
- mese di scadenza della polizza;
- tipologie di contratto, distinte principalmente in *polizza principale nella compagnia*, *polizza principale in altra compagnia* e *classe non agevolata*.

L'ultima categoria include tutti gli attributi inerenti alla situazione sui sinistri e sul relativo stato economico degli assicurati nei confronti dell'azienda. Un'importante sottocategoria relativa ai sinistri comprende principalmente l'attestato di rischio, ovvero il numero dei sinistri denunciati negli ultimi cinque anni, supportato da altri attributi come "Anni senza assicurazione" e "Distanza dall'ultimo sinistro" che indicano rispettivamente il numero di anni in cui un cliente è stato assicurato con la stessa compagnia assicurativa e la distanza temporale dall'ultimo sinistro, nel caso in cui negli ultimi cinque anni ce ne sia stato uno o più. Nel caso in cui il cliente abbia commesso un incidente durante l'anno di riferimento della validità dei dati analizzati, gli attributi "Data del sinistro" e "Numero totale di sinistri" forniscono rispettivamente indicazioni circa la data esatta del giorno in cui è avvenuto il sinistro e il numero totale dei sinistri cumulati fino ad allora. È possibile inoltre conoscere attraverso alcuni attributi fondamentali, dettagli sul tipo di incidente, distinguendo se l'incidente è stato subito o causato e nell'ultimo caso si tiene traccia anche del costo dell'incidente causato, con l'attributo "Costo degli incidenti causati". Due attributi di questo genere inoltre, "Numero di incidenti con colpa grave causati" e "Numero di incidenti senza colpa grave causati", riportano due principali categorie di incidenti, differenziate in base al fatto che si tratti rispettivamente di un incidente che reca una disabilità maggiore del 9% o un incidente con disabilità minore del 9%.

I livelli di rischio che sono stati determinati nell'analisi sono principalmente quattro:

1. numero degli incidenti con colpa grave causati;
2. numero degli incidenti senza colpa grave causati;
3. numero degli incidenti causati;
4. costo degli incidenti causati.

2.3 Scenario

Prima di applicare le tecniche supervisionate e non, di esplorazione dei dati, che permettono di supportare la compagnia assicurativa per una migliore gestione dei dati telematici, è necessaria l'attuazione di una fase di *preprocessing*. Questa fase si occupa principalmente di preparare i dati affinché le tecniche successive possano essere applicate al meglio e possano ricavare risultati non fuorvianti. Infatti, quando si affronta un problema di machine learning, il primo passo consiste nel predisporre un buon training set a partire dai dataset disponibili, in modo da costruire un modello piuttosto accurato. È necessario quindi fare un'analisi preliminare in modo da evidenziare eventuali criticità e nel caso, ristrutturare i dati in modo da eliminarle e rendere gli stessi dati compatibili con gli strumenti utilizzati.

I processi aziendali analizzano la correlazione tra il livello di rischio di una polizza e una o più variabili caratteristiche per identificare profili di clienti tipici o di particolare rilievo. La telematica consente di arricchire la base di dati aziendale con dati eterogenei, spesso non considerati o considerati solo in parte durante i processi decisionali. Questo lavoro di tesi è finalizzato ad arricchire la caratterizzazione delle polizze con i dati della telematica e studiare le correlazioni più significative all'interno dei dati separatamente per ogni livello di rischio. L'analisi si basa su modelli di classificazione interpretabili, quali i classificatori associativi [3, 15]; essi combinano l'accuratezza delle predizioni effettuate con la possibilità di esplorare il contenuto dei modelli predittivi per identificare le correlazioni più significative che pilotano l'assegnazione di un certo profilo di polizza ad un determinato livello di rischio. Nello specifico, è stato applicato il classificatore associativo L^3 (*Live and Let Live*) [3], costruito selezionando un insieme di regole di associazione di alta qualità per quel modello, di cui si discuterà teoricamente nel capitolo 3 e del suo utilizzo nel capitolo 4. Con lo scopo di profilare le polizze aventi livello di rischio simile, questo studio si propone di generare diversi modelli predittivi interpretabili, esplorarne le caratteristiche, selezionare i pattern più significativi e identificare insiemi di feature che, assunti determinati valori, delineano un livello di rischio specifico. A differenza dei modelli statistici tradizionali, i modelli generati attraverso l'uso del classificatore L^3, e in generale grazie ad una buona parte delle tecniche di data mining, consentono di analizzare le co-occorrenze tra combinazioni di feature di varia natura e associarle a specifici livelli di rischio con elevata attendibilità, analizzando moli di dati potenzialmente molto grandi e quindi non approcciabili su base statistica. I pattern selezionati per ciascun livello di rischio saranno poi confrontati per evidenziare le differenze tra polizze di diversa tipologia. Inoltre, è stato studiato anche l'impatto della scelta di alcuni indicatori utilizzati per descrivere

i vari livelli di rischio, denominati *Key Performance Indicators*, definendo strutture correlate usufruibili per scopi futuri.

Capitolo 3

Data Mining

In questo capitolo, verranno introdotte le tecniche principali del concetto di data mining e del processo di cui fa parte. Si discuterà in dettaglio del classificatore L^3 e delle regole di associazione delle quali il classificatore si serve.

3.1 Introduzione

L'evoluzione del concetto di interconnessione e condivisione dei dati e l'introduzione di supporti hardware più performanti hanno portato ad una crescita sostanziale di dati in circolazione in qualsiasi campo applicativo, a partire da sistemi bancari e di mercato, per arrivare al campo della sanità, e così via. Effettuare una ricerca online, compiere transazioni bancarie, rilevare dati metereologici e quant'altro, introducono al giorno d'oggi moli di dati piuttosto significative che hanno la necessità di essere gestite al meglio. L'aumento dei dati ha portato all'introduzione di nuovi modelli di archiviazione in ambito aziendale, tra cui spicca il concetto di *Data Warehouse* [6]. Con il termine data warehouse si intende un archivio di dati di vari tipi, reperiti da fonti altrettanto differenti tra di loro e organizzati in maniera tale da ricavarne informazioni approfondite e dettagliate, per garantire gestione e analisi migliori ed efficaci. I dati ai quali il data warehouse attinge, sono quelli archiviati nel database, che va mantenuto separato. Il motivo principale per cui i due archivi vengono tenuti separati è principalmente quello di non influenzare le continue transazioni sul database con le operazioni complesse di analisi del data warehouse. Indipendentemente dagli ambiti applicativi, l'obiettivo principale di quest'ultimo è quello di gestire la notevole raccolta di dati in modo da trarne strategie aziendali rilevanti.

La crescita esplosiva di dati disponibili ha portato a rendere le tecniche prima conosciute per l'analisi di dati, poco idonee. Il metodo tradizionale di trasformare i dati in conoscenza si basava sull'analisi ed esplorazione manuali. Solitamente, in base al dominio applicativo, gli specialisti avevano il compito di stilare, secondo determinate scadenze, dei report in cui venivano fatte delle analisi sui dati reperiti in quell'arco temporale, relative all'andamento informativo [7]. Questa strategia si è rivelata carente e soprattutto costosa nel corso degli anni, sia a causa della crescita della mole di dati da analizzare che richiedesse l'utilizzo di algoritmi altamente scalabili, sia per la varietà dei dati rilevati, in base al tipo di contesto applicativo. L'obiettivo principale è quindi quello di rendere il processo di analisi, almeno parzialmente, automatizzato per assistere gli utenti nella gestione dei dati e nell'estrazione di informazioni a valore aggiunto. Nasce così la necessità di introdurre nuovi strumenti sofisticati e metodologie adatte per gestire, analizzare ed estrarre informazioni entro tempistiche ritenute accettabili. Viene introdotto così il concetto di *data mining*.

3.2 Data Mining: concetti base

Il termine data mining [9] indica l'insieme di tecniche che permettono l'estrazione di conoscenza e informazioni da quantità di dati consistenti, in modo da individuare correlazioni, probabilmente nascoste, da rendere disponibili per analisi future. Con il termine conoscenza utilizzato in quest'ambito, si intende l'insieme di dati che associati tra di loro riescono a fornire informazioni e connessioni significative non note a priori. Il concetto di data mining nasce dall'esplosione di dati avvenuta in questi anni, con l'obiettivo di introdurre tecniche che potessero al meglio estrapolare informazioni da basi di dati considerevolmente ampie. Il volume dei database è stato soggetto ad un incremento sia in termini di numero di record, sia per il numero di attributi presenti in esso. Le strategie di data mining si prefiggono di ricavare conoscenze non immediatamente disponibili ed evidenti, cercando di definire dei modelli astratti, denominati *pattern*, che indicano prevalentemente una rappresentazione sintetica dei dati. I pattern consistono quindi in modelli altamente comprensibili che hanno lo scopo di fornire degli schemi utili e soprattutto validi con un certo grado di confidenza. Molte tecniche di data mining richiedono un'adeguata formattazione e preparazione dei dati, in modo da evitare che i risultati ottenuti, ossia i pattern sopra esposti, possano indurre a conclusioni fuorvianti.

3.2.1 Tipi di dati e modelli di archiviazione

Le sorgenti che riescono a rilevare quantità considerevoli di dati sono ormai numerose, così come sono altrettanto numerose le tipologie di dati stessi. In base al contesto applicativo e alle situazioni, i dati da dover analizzare includono numerose categorie e tipologie di rappresentazione eterogenee, come ad esempio dati multimediali, dati relativi ai social network, stream di dati e così via. I dati possono essere fondamentalmente classificati in tre categorie in termini di struttura [6]:

1. dati strutturati: le informazioni sono organizzate secondo uno schema chiaro e ben preciso;

2. dati non strutturati: i dati appartenenti a questa categoria non seguono un determinato schema o regole ben definite, spingendo a delineare un trattamento specifico per questa tipologia;

3. dati semi-strutturati: i dati sono parzialmente strutturati.

Le tecniche di Data Mining si preoccupano di estrarre informazioni da fonti di diverso tipo, adattando varie soluzioni per ciascuna categoria sopra esposta. La tipologia dei dati da analizzare varia in base al contesto applicativo e alla metodologia di reperibilità dei dati stessi. Nel caso di un contesto aziendale, ad esempio, i dati sono solitamente strutturati in database. Essi consistono in modelli di archiviazione dati, organizzati in base a determinati schemi, in modo da accedervi in maniera semplice per poter interagire con essi. Affinchè le collezioni di dati siano identificati in database, essi devono essere:

- persistenti: la loro attività non è limitata all'uso che i programmi fanno di essi;

- condivise: i dati sono condivisi tra più applicazioni e più utenti, in modo da eliminare ridondanze all'interno di un contesto aziendale;

- consistenti: la quantità di dati potrebbe risultare considerevolmente elevata.

I dati memorizzati all'interno dei database sono di norma gestiti da software denominati *Sistemi per la Gestione di Basi di Dati* (DBMS). Questi ultimi consistono in prodotti software che permettono il monitoraggio delle collezioni di dati, tra cui la realizzazione, la modifica e l'interrogazione per poter accedere ad essi con l'obiettivo di fornire sicurezza, affidabilità e integrità

dei dati. Quest'ultima è ritenuta fondamentale ai fini di un'ottima gestione dei database poiché, dato che essi consentono l'accesso multi-utente, è necessario che l'accesso concorrente relativo a determinati dati sia regolamentato e monitorato.

Dal punto di vista dell'architettura interna al database, la rappresentazione con cui i dati vengono disposti e archiviati nel database definisce un modello, ovvero la struttura organizzativa dei dati inerenti all'ambito informativo di interesse. Inizialmente, il modello gerarchico e il modello reticolare costituivano i principali modelli di rappresentazione di dati all'interno di un database e sono ancora oggi utilizzati, anche se raramente. Il modello gerarchico permette di schematizzare i dati di un database in più tabelle, organizzate secondo una struttura ad albero. In questo modo, una tabella principale rappresenta la radice e tutte le altre seguono il modello gerarchico, in modo da rispettarne le regole e le correlazioni di struttura. La ricerca di un'informazione è ritenuta veloce ma poco flessibile in questo caso, poiché è necessario percorrere tutto l'albero per poter accedere al dato desiderato. Il modello di tipo reticolare costituisce invece un aggiornamento del modello gerarchico, poiché parte da quest'ultimo per meglio rendere la ricerca dell'informazione più flessibile. La struttura è rappresentata da un grafo non orientato, in cui non è più necessario definire una tabella radice, ma ogni tabella ha uguale rilevanza.

Il modello relazionale è il modello più diffuso per rappresentare la struttura di un database e ha reso i precedenti modelli poco significativi e ormai in disuso. Il modello relazionale definisce un preciso schema logico, che rende i dati indipendenti dalla struttura fisica del database in modo da determinare un livello di astrazione forte rispetto ai modelli passati. Questo garantisce che l'accesso alle informazioni possa essere indipendente dall'organizzazione stessa con cui i dati sono gestiti. Il concetto principale su cui il modello relazionale si basa è quello di *relazione*. Quest'ultima definisce una tabella, in cui vi sono tante colonne quanti sono gli attributi, definiti anche domini della relazione, e le righe quante sono le tuple, identificate in base al concetto di chiave. Una chiave è definita da uno o più attributi che determinano l'identità della tupla, in modo da ottenere tuple univoche nella relazione. Il modello relazionale ha avuto un forte impatto sul concetto di gestione dei dati, grazie soprattutto al livello di astrazione da esso definito, rendendo la ricerca veloce e soprattutto flessibile. Infatti al giorno d'oggi, in contesti aziendali, viene prevalentemente utilizzato il modello relazionale garantendo gestione e monitoraggio dei dati adatti alle esigenze.

3.2.2 Strategie di Data Mining

Le strategie di data mining principali [4] che delineano il metodo di analisi con cui si affrontano determinati problemi, sono suddivisibili fondamentalmente in due categorie:

1. strategie supervisionate;
2. strategie non supervisionate.

La prima categoria include tutte le tecniche che prevedono l'elaborazione di un modello di dati attraverso l'addestramento guidato e controllato che utilizza una collezione di dati già nota, denominata *training set*. Quest'ultimo solitamente è costituito da dati etichettati a priori attraverso cui si procede all'istruzione del modello con l'applicazione di algoritmi, il cui scopo è quello di apprendere dai dati alcune caratteristiche che indicano una determinata classe in funzione di un numero più o meno ristretto di attributi, per poi classificare o predire rispettivamente dati non etichettati o eventi futuri in base alla conoscenza acquisita. Nell'ambito del database relazionale, con il termine "classe" si indica un attributo i cui valori determinano l'appartenenza di una certa informazione.

L'apprendimento supervisionato si prefigge di svolgere principalmente due attività che è bene distinguere:

- classificazione;
- predizione.

Così come accennato precedentemente, il processo di classificazione consiste nella creazione di un modello accurato che permetta di assegnare con un certo grado di confidenza dati non etichettati ad una determinata classe, attraverso la scoperta di dettagliati profili che descrivano ogni valore di classe. Questo processo presuppone l'utilizzo e l'analisi di un training set, che consente di addestrare il modello per l'attributo classe già definito in base a correlazioni note, per poi classificare nuovi dati la cui etichetta di classe non è conosciuta a priori. Per testare la validità del classificatore creato, è possibile applicare varie tecniche di partizionamento, tra le quali spicca la tecnica del *cross validation* [16]. Essa consiste nel determinare un numero "n" di sottoinsiemi del dataset originale per poi utilizzare ciclicamente n-1 sottoinsiemi come training set e il rimanente come testset. Quest'ultimo contiene record già etichettati per verificare se, applicando il modello appena creato, le regole stabilite siano applicabili in maniera generale o meno e se il modello risulti altamente valido, in base allo studio di alcuni parametri, quali

accuratezza, *precisione* e *recall*. L'accuratezza determina quanti oggetti sono stati classificati correttamente rispetto all'insieme totale degli oggetti classificati, cioè quanto il classificatore risulti esatto o meno. La precisione testa quanto il classificatore sia preciso, ossia quanti elementi sono stati classificati in maniera corretta rispetto al totale degli elementi classificati secondo una determinata classe. Mentre il recall misura la capacità del classificatore a recuperare gli elementi corretti relativamente al totale degli elementi per quella classe.

Il concetto di classificazione può essere utilizzato relativamente a svariati settori, ad esempio, relativamente al settore della sanità è possibile fare uso della classificazione per predire quale trattamento è necessario adottare per un determinato profilo secondo una patologia; nel settore bancario è necessario distinguere profili a rischio o meno per l'assegnazione di prestiti o ad esempio, identificare e classificare un'email come spam e così via.

La predizione [16] invece si prefigge l'obiettivo di prevedere un comportamento futuro, inizialmente sconosciuto, in base a valori e informazioni di partenza correlati tra di loro. Molte tecniche utilizzate per la classificazione permettono di ricavare anche risultati predittivi. Nonostante questo, esistono delle differenze sostanziali tra i due modelli. La prima differenza consiste nel tipo di risultato che essi riescono ad ottenere. Infatti, la classificazione determina l'appartenenza di record o profili ad etichette di classe di tipo categorico, ossia valori discreti o nominali; la predizione, invece, solitamente ha lo scopo di predire valori continui. Inoltre, la classificazione determina valori già conosciuti, dipendenti dalle etichette di classe già stabilite nel training set, mentre la predizione ottiene valori nuovi e sconosciuti, ma correlati agli input.

La seconda categoria di attività tipiche del data mining include algoritmi descrittivi che permettono l'identificazione di caratteristiche e correlazioni tra i dati senza l'utilizzo di un training set come insieme di dati di input già ben definiti, ma servendosi esclusivamente del dataset da analizzare. Gli algoritmi appartenenti a questa categoria non hanno come obiettivo quello di classificare dati secondo determinate classi o predire situazioni nel futuro, ma permettono di cogliere nella struttura dei dati in esame informazioni interessanti, probabilmente nascoste, in modo da estrarre modelli interpretabili e utilizzabili in base al contesto applicativo.

3.2.3 Tecniche di classificazione

I requisiti che un buon classificatore dovrebbe soddisfare e che determinano le differenze tra le numerose tecniche esistenti relative alla classificazione, riguardano principalmente:

1. la precisione con cui il classificatore riesce ad etichettare nuovi dati;
2. la velocità di calcolo del classificatore;
3. l'interpretabilità del modello ottenuto;
4. scalabilità;
5. robustezza: ossia quanto un classificatore risulti corretto e propenso alla gestione di dati rumorosi e anomali.

La scelta della tecnica idonea dipende sia dal contesto applicativo, nonché dalle informazioni dei dati in questione, sia dall'interpretabilità e dall'utilizzo del modello ricavato. Le tecniche di classificazione principali [4, 8, 10] sono le seguenti:

- decision tree: questa tecnica consiste in uno strumento di supporto alle decisioni e consente di costruire un modello di classificazione secondo una struttura ad albero. Partendo dal nodo radice, tutti i nodi ad esclusione delle foglie rappresentano delle condizioni da verificare, susseguite in maniera sequenziale e gli archi che si diramano da essi rappresentano le possibili soluzioni al relativo test. Le diramazioni possono essere due, come nel caso di un albero decisionale binario, o multiple. I nodi foglia costituiscono l'etichetta da assegnare, il cui range e tipo di valore sono fortemente dipendenti dal contesto applicativo, dato che sono supportati sia valori categorici che numerici. Gli algoritmi utilizzati per realizzare il sopracitato processo di decisione sono svariati, tra cui spiccano ID3 e CART; nonostante la varietà degli algoritmi, applicano tutti una strategia greedy e ricorsiva. Con tecnica greedy si intende un approccio che consiste nel selezionare la decisione ottimale ad ogni step, senza la possibilità di tornare al passo precedente, mentre è definita una strategia ricorsiva poiché permette di suddividere il problema in partizioni più piccole, per ottenere un risultato e uno scorrimento tra i passi più rapidi. Uno dei vantaggi principali dell'utilizzo di questo modello consiste nell'elaborare i dati indipendentemente dal loro dominio, in modo tale da favorire la conoscenza esplorativa. La fase di decisione è molto veloce ma la qualità del risultato può dipendere fortemente dalla ricchezza dei dati. Un altro sostanziale vantaggio risiede nell'intuizione di un tale modello, che lo rende facilmente interpretabile;
- bayesian classification: consiste in un classificatore costruito secondo il teorema di Bayes. Quest'ultimo permette di calcolare la probabilità

condizionata di un evento A, dato un evento B, presupponendo la conoscenza delle probabilità indipendenti dei due eventi e della probabilità condizionata di B, dato A. Attraverso l'utilizzo di questo teorema è possibile calcolare la probabilità di assegnare una determinata etichetta all'elemento in esame. Il classificatore bayesiano risulta altamente performante e rapido ed una delle caratteristiche peculiari riguarda il fatto che ogni attributo viene analizzato singolarmente, in modo da verificare la sua probabilità di appartenenza ad una classe in maniera indipendente dagli altri attributi;

- support vector machines: utilizzato in vari ambiti, permette di rappresentare i dati di un training set come punti in uno spazio, separati in base alla categoria di appartenenza. Dopodichè, si costruisce un iperpiano che massimizzi la distanza tra due o più insiemi di decisione e i nuovi dati verranno classificati in base alla porzione di piano in cui si troveranno. Questa strategia risulta molto lenta nella costruzione del modello, ma piuttosto accurata nella classificazione;

- neural network: il modello di rete neurale si basa sulla struttura di un cervello umano, secondo la cui analogia, i neuroni rappresentano le unità elaborative e le sinapsi rappresentano i collegamenti tra le unità. Solitamente la struttura di una rete neurale è costituita da strati e in generale è possibile distinguerne tre, ossia strato di input, strato interno e strato di output. Ogni strato è costituito da "n" nodi e le connessioni tra uno strato e l'altro sono pesate, ovvero è assegnato loro un costo. Attraverso l'utilizzo di un training set, quindi secondo l'approccio supervisionato, è possibile istruire la rete neurale affinché l'errore derivante dalla differenza tra i nodi di output già definiti e gli output derivanti da una prima analisi secondo parametri casuali, diminuisca sempre più. Utilizzando il meccanismo denominato *backpropagation*, è possibile propagare indietro l'errore verso gli strati precedenti in maniera iterativa, affinchè tutti i nodi possano adattarsi e ricominciare. Nonostante il modello delle reti neurali possa risultare lento nella fase di apprendimento e poco interpretabile, alcuni vantaggi principali derivanti dal suo utilizzo consistono nell'elevata accuratezza che esso garantisce, nella robustezza nei confronti di valori anomali e rumori e nel supporto a valori sia continui che discreti.

3.2.4 Regressione

La predizione permette di determinare un evento futuro sconosciuto, ma correlato a un insieme di dati raccolti nel passato. Essa gestisce funzioni a valori

continui e le tecniche più utilizzate consistono in modelli matematici. Oltre al fatto che alcune delle tecniche sopra discusse relative al concetto di classificazione possono essere utilizzate anche a scopi predittivi, come ad esempio, le reti neurali e l'albero delle decisioni, in questo caso il metodo più utilizzato è quello della *regressione* [4]. Nonostante le varie tipologie di regressione, tra cui la regressione logistica, il concetto generale che sta alla base di questa metodologia consiste nello stimare inizialmente l'obiettivo in base alle variabili di ingresso del training set, ossia i predittori. Attraverso questi ultimi, i cui valori sono già noti a priori, è possibile delineare un modello lineare o non lineare, affinchè le variabili in uscita possano essere rappresentate dipendentemente dagli ingressi con maggiore probabilità possibile. Per mezzo di un test set, successivamente, è possibile testare l'efficacia e l'accuratezza del modello creato per poter prevedere nuovi insiemi di dati, i cui valori sono sconosciuti. Il metodo della regressione è solitamente utilizzato per predire valori continui attraverso il calcolo della distribuzione dei dati disponibili, differentemente dalla classificazione, la quale si occupa invece di etichettare valori in prevalenza discreti.

3.2.5 Tecniche descrittive

Il data mining non supervisionato consiste di tecniche che permettono l'estrazione di modelli interpretabili, intrinseci nella struttura dei dati in esame ma non deducibili a primo impatto. Essi si differenziano dalla categoria delle tecniche predittive e di classificazione principalmente per due motivi, rispettivamente in base all'output ricavato e all'input fornito:

1. permettono di descrivere correlazioni tra i dati strutturati già in possesso, per ricavare informazioni utili da poter applicare in futuro. A differenza dei metodi inerenti alla classificazione, essi non hanno l'obiettivo di estrapolare target dei record in esame in base ad etichette già note o di predire eventi e comportamenti futuri nel caso della predizione, ma cercano di analizzare informazioni aggiuntive basandosi soltanto sui dati a disposizione, per estrarre una conoscenza utile ed interpretabile;

2. il concetto che sta alla base delle tecniche non supervisionate è quello di caratterizzare i dati e creare un processo risolutivo senza avere come input un insieme di dati già etichettati o annotati a priori, a differenza del training set utilizzato nelle tecniche predittive.

Le tecniche descrittive possono essere impiegate in molti ambiti applicativi, ad esempio è possibile gestire al meglio la disposizione della merce all'interno

di un supermarket in base alla frequenza di acquisti dei clienti o in ambito assicurativo/bancario è possibile delineare dei profili simili che caratterizzano vari livelli di rischio o ancora, fornire servizi di pubblicità in base alle visite su siti di e-commerce e così via. Le tecniche principali appartenenti alla categoria di metodi descrittivi sono *Clustering* ed *Estrazione di regole di associazione*.

3.2.6 Clustering

Il clustering [8] è un processo che permette di suddividere l'insieme di dati in esame in più sotto-insiemi, denominati *cluster*. Essi hanno la caratteristica di essere differenti l'uno con gli altri, ma i cui elementi sono fortemente correlati e simili tra di loro per ciascun gruppo. A differenza della classificazione, questo tipo di analisi presuppone di non avere alcuna conoscenza sui cluster che verranno creati, rientrando quindi nella categoria di tecniche non supervisionate del data mining. Secondo la definizione precedente di clustering, è possibile identificare ogni gruppo come se fosse una classe, dato che i principi che regolano il processo di scissione dell'intero insieme di dati si basa sul concetto di somiglianza ed è possibile regolare il trattamento analitico dei dati stessi in maniera individuale. Il clustering può essere adottato in numerosi ambiti, tra i quali spiccano il concetto del cosiddetto *market segmentation*, che consiste nella suddivisione degli acquirenti in sotto-gruppi in modo da differenziarne le attività in base alle caratteristiche dei gruppi stessi, e il rilevamento di atteggiamenti sospetti relativamente a vari settori, attraverso il riconoscimento di valori non ritenuti affini rispetto ai cluster identificati. Un altro utilizzo molto frequente è quello del clusterizzare documenti simili in base al contenuto, che trova applicazione, ad esempio, nelle ricerche Web.

Le varie tecniche in merito alla clusterizzazione sono distinte in categorie relativamente al tipo di approccio e operatività, e sono:

- clustering partizionale: consiste nel partizionare il dataset originale in più cluster. Il concetto cardine utilizzato come criterio di partizione è quello di distanza. Il processo tende a spostare in maniera ciclica gli elementi in modo tale da minimizzare lo sparpagliamento degli elementi di uno stesso cluster e viceversa, massimizzare la distanza tra cluster differenti, garantendo che nessuno di essi sia completamente vuoto ma che abbiano tutti almeno un elemento al loro interno. Uno degli algoritmi più noti che segue questa tecnica è il cosiddetto *Algoritmo K-means* [16]. L'obiettivo di quest'ultimo consiste nel minimizzare la distanza intra-cluster. L'algoritmo prevede un numero "k" di cluster definito a priori ed opera secondo i seguenti quattro passi:

1. seleziona casualmente "k" punti dell'insieme iniziale che ricopriranno il ruolo di centroide. Quest'ultimo determina il valore medio del cluster;
2. calcola la distanza tra i punti del dataset e i centroidi scelti;
3. distribuisce gli elementi nei vari cluster in base alla minima distanza con i centroidi;
4. ricalcola la media dei nuovi cluster e ripete dal passo 2 fin quando la distanza tra gli oggetti rispetto ai punti centrali rimane invariata.

Il vantaggio principale nell'utilizzo del suddetto algoritmo consiste nella velocità di convergenza. Tra gli svantaggi più importanti invece, vi è il fatto di non garantire la soluzione migliore. Inoltre questo algoritmo riscontra difficoltà nel caso di cluster di forma e densità differenti e non riesce a gestire al meglio i valori anomali;

- clustering gerarchico: a differenza del metodo di partizionamento, esso permette di suddividere il dataset originale in cluster annidati, organizzati secondo una gerarchia e non richiede la definizione a priori del numero di classi, come si è potuto constatare nell'algoritmo K-means. Il criterio base utilizzato per suddividere o accorpare i cluster è solitamente la distanza o la densità. È possibile distinguere due principali strategie:

  1. approccio agglomerativo: prevede l'inserimento di ogni elemento all'interno di un cluster a sé stante, per poi successivamente accorpare i cluster fino a realizzare un unico insieme, denominato *dendogramma*, ovvero una rappresentazione grafica che tiene conto della distanza tra i cluster e dei passi che si sono susseguiti per il loro accorpamento;
  2. approccio divisivo: consiste nel considerare inizialmente un unico insieme con la totalità degli elementi all'interno, per poi suddividere successivamente quest'ultimo in tanti gruppi più piccoli fino a raggiungere una determinata situazione;

- clustering basato sulla densità: questo approccio utilizza come criterio di partizione dei cluster il concetto di densità. A differenza degli altri approcci, il parametro di accorpamento degli elementi appartenenti ad un dataset e la valutazione di nuovi elementi dipendono da quanto un cluster sia denso o meno, ovvero quanti dati siano presenti in un

determinato raggio. Tra i principali vantaggi di questa tecnica vi è la possibilità di ottenere gruppi con forme non esclusivamente sferiche e il trattamento efficiente di valori anomali.

3.2.7 Estrazione di regole di associazione

L'analisi associativa [1, 14] è una tecnica di tipo descrittivo che permette di estrapolare correlazioni, inizialmente nascoste, tra i dati all'interno di un database transazionale mediamente grande. Il concetto su cui si basa questo tipo di analisi è quello di determinare pattern frequenti, ossia schemi significativi facilmente interpretabili che ricorrono spesso all'interno di un dataset. Un database transazionale identifica un database in cui ogni record rappresenta un'operazione univoca denominata *transazione*, costituita da un insieme di *item*. Un item determina l'unità informativa a partire dalla quale si effettuerà l'analisi delle co-occorrenze e la generazione dei pattern più rilevanti determinate dalla coppia feature-valore della feature. In questo contesto, l'analisi associativa si prefigge di estrarre insiemi di item che siano fortemente correlati tra di loro e che in relazione ad alcune misure di interesse, possano soddisfare dei requisiti di occorrenza. Le regole di associazione [1] consistono in espressioni che regolano come i dati siano correlati tra di loro e giustificano nello specifico il verificarsi di un item in base all'occorrenza di uno o un insieme di altri item. Le regole sono espresse in base alla forma: $Corpo \Rightarrow Testa$, dove "Corpo" sta ad indicare l'antecedente della regola, ossia uno o più item che si collocano a sinistra dell'espressione, mentre "Testa" rappresenta il conseguente, ovvero l'itemset che ne consegue. La validità e la robustezza della regola sono esprimibili principalmente mediante tre misure statistiche di interesse:

1. supporto: esprime la frequenza percentuale con cui l'itemset occorre nel dataset rispetto al numero totale delle transazioni e nello specifico rivela quanto l'insieme che comprende corpo e testa sia ricorrente, indicando l'utilità della regola stessa;

2. confidenza: consiste nella probabilità condizionata della presenza di un item data l'occorrenza dell'altro, ovvero la probabilità del verificarsi della testa dato il corpo. Essa esprime la certezza e la forza delle co-occorrenze presenti nella regola;

3. lift: è un parametro che esprime l'indice di correlazione tra la testa e il corpo in una regola e misura lo scostamento dall'indipendenza delle occorrenze. Differisce dalla confidenza e supera i relativi limiti poiché tiene conto anche del supporto della testa, qualora i due itemset non

fossero stocasticamente indipendenti. I possibili valori che il lift può assumere sono:

a. $= 1$: indica l'indipendenza tra l'antecedente e il conseguente;
b. > 1: esprime la dipendenza tra corpo e testa, che aumenta al crescere positivo del discostamento da 1;
c. < 1: indica una dipendenza negativa, ovvero il verificarsi del corpo implica negativamente la testa.

Il processo inerente all'estrazione delle regole è una ricerca esplorativa che consiste principalmente nella ricerca dei pattern più frequenti. In questo contesto, il concetto di frequenza indica la possibilità di ricavare regole il cui supporto e la cui confidenza siano maggiori o uguali rispettivamente ad una soglia di supporto minimo e ad una soglia di confidenza minima. Si effettua quindi lo scarto dei pattern che non soddisfano tali vincoli e successivamente si procede alla scoperta e alla generazione delle regole associative relative agli itemset più frequenti.

Uno degli ambiti applicativi in cui l'analisi associativa espressa mediante regole ha avuto maggiore impatto sin dall'inizio, riguarda il concetto del *market basket analysis* [5, 14]. Quest'ultimo si riferisce all'analisi di associazioni tra i diversi prodotti che i clienti acquistano solitamente al supermercato in modo da delineare meglio i profili degli acquirenti, sia per applicare la cosiddetta *cross-selling*, che consiste in una strategia capace di incrociare informazioni su particolari clienti per proporre prodotti affiliati in base a dati prelevati in passato, sia per una migliore gestione degli scaffali del market. Nella Tabella 3.1 è riportato un esempio.

Tabella 3.1: *Ipotetico insieme di scontrini rilasciati da un supermarket*

TID	*items*
1	Wine, Milk, Coffee
2	Bread, Diapers
3	Coffee, Wine, Milk
4	Milk, Yogurt, Wine
5	Beer, Wine, Coffee

Relativamente all'esempio, è possibile dedurre una delle tante regole associative correlate e le relative misure di interesse: $Coffee \Rightarrow Wine$.

Questa regola di associazione indica la correlazione tra due item, ossia "caffè" e "vino". I parametri di interesse applicati al caso in questione riportano:

- il supporto di questa regola, che ne indica la frequenza relativamente al totale degli scontrini, è pari a 3/5, ovvero il 60% degli scontrini contiene l'occorrenza in questione;

- la confidenza, ovvero il rapporto tra l'unione degli itemset rispetto al totale delle transazioni che contengono il corpo, è uguale a $3/3 = 1$, che rappresenta quindi il 100% delle probabilità di trovare entrambi gli item, dato l'elemento "caffè";

- il lift è uguale al rapporto tra la confidenza della regola e il supporto della testa, ovvero $(3/3)/(4/5) = 1,25$, che rappresenta una correlazione positiva tra gli item.

3.3 Knowledge Discovery in Databases

Il concetto di data mining è lo step fondamentale di un processo più ampio che consiste non solo nell'estrazione di informazioni dai dati e/o nelle predizioni possibili di eventi futuri, ma affinchè le tecniche di data mining vengano applicate in maniera esauriente senza condurre a risultati ingannevoli, è necessaria l'adozione di alcuni passi che processino i dati per eliminare eventuali criticità e selezionarli, integrarli e organizzarli in base al tipo di analisi da effettuare. Con il termine *Knowledge Discovery in Databases* (KDD) si intende il processo che racchiude, in generale, i concetti di organizzazione dei dati, ricerca di schemi significativi (pattern), valutazione della conoscenza acquisita e la generazione di un'interpretazione più semplice possibile per gli utilizzi futuri.

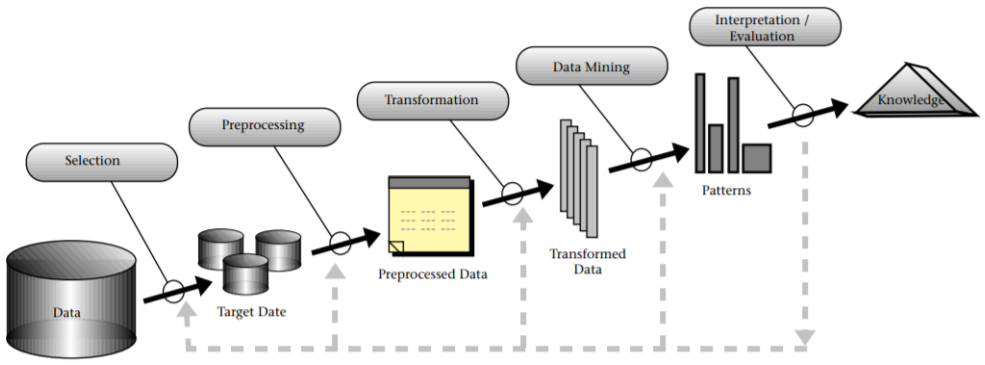

Figura 3.1: *Overview relativa agli step che compongono il processo KDD [7]*

Gli step che compongono il processo di Knowledge Discovery in Databases [7] sono raffigurati in Figura 3.1:

1. *Data selection*: acquisire, in base all'ambito applicativo in questione, una conoscenza tale da selezionare un set di dati significativo su cui eseguire l'analisi;

2. *Preprocessing*: questa fase include due sub-step principali per il trattamento dei dati selezionati nel passo 1, in modo da gestire problemi di criticità dei dati:

   - *Data cleaning*: questo step si occupa di "pulire" i dati risolvendo inconsistenze, incompatibilità e valori mancanti;
   - *Data integration*: questa sezione include le tecniche per integrare dati estratti da più sorgenti differenti, identificare e risolvere conflitti e gestire le ridondanze;

3. *Data transformation*: consiste in un'operazione di trasformazione dati in modo da ricavarne strutture idonee per l'applicazione di determinate tecniche di mining. Una delle operazioni solitamente utile inerente a questa fase è denominata *Data Reduction* che consiste nel ridurre il dataset originale in termini di volume, lasciandone invariata l'integrità;

4. *Data mining*: processo che prevede l'applicazione di alcune tecniche prevalentemente automatizzate con lo scopo di estrarre alcuni data-pattern di interesse in una particolare forma di rappresentazione o su un insieme di rappresentazioni diverse;

5. *Pattern evaluation and interpretation*: valutazione e identificazione dei modelli ritenuti più interessanti relativamente al contesto e possibile ritorno alle fasi iniziali per ulteriori iterazioni.

Di seguito verranno introdotte le sezioni più salienti circa il processo sopra esposto, specificandone meglio le tecniche e lo scopo.

3.3.1 Data Cleaning e principali problemi

Dopo aver estratto gli attributi essenziali all'analisi nella fase di "selection data" tenendo conto sia del dominio dei dati sia dell'obiettivo che si cerca di ottenere per mezzo del processo di data mining, è necessario applicare l'operazione di data cleaning affinchè vengano eliminate eventuali criticità e inconsistenze insite nel dataset. Con il termine "data cleaning" [8, 13] si fa riferimento ad un generico processo capace di garantire, con una certa soglia di affidabilità, la correttezza di una grande mole di dati. L'obiettivo di questa fase è quello di migliorare la qualità dei dati grezzi reperiti direttamente

dalla fonte, in modo da tener conto di vari fattori che potrebbero compromettere la validità dei risultati ottenuti a fine analisi. Se questo processo venisse trascurato, i dati allo stato iniziale non potrebbero indurre a risultati consistenti dopo l'applicazione degli approcci di data mining.

In relazione a un insieme di dati da analizzare, la possibilità di imbattersi in questo fenomeno è molto probabile e le motivazioni possono essere più o meno numerose in base al tipo di contesto applicativo. Le ragioni che portano ad avere valori mancanti nel dataset, rendendo alcuni attributi non accurati o incompleti, possono dipendere da problemi legati a malfunzionamenti della meccanica inerente agli strumenti adottati per rilevare i dati o alle ristrette potenzialità di questi ultimi; altri problemi possono insorgere semplicemente dal fatto che alcuni utenti preferiscono tralasciare alcuni campi nel riempimento di qualsivoglia form, perché ritenuti molto confidenziali o di scarsa importanza. Alcuni metodi noti per ovviare al problema sopra descritto, sono i seguenti:

- sostituire i campi mancanti con una variabile unica per tutti: questo metodo permette di eliminare i valori mancanti e sostituirli con una qualsiasi etichetta. Qualora i valori rilevati con questa etichetta dovessero essere troppi, si arriverebbe però ad alcune considerazioni errate tali da dover essere accuratamente gestite;

- eliminare le tuple che contengono valori mancanti: questo approccio risulta utile nel caso in cui i valori mancanti siano davvero pochi e ininfluenti;

- utilizzare un valore ritenuto più probabile mediante l'utilizzo di opportuni strumenti probabilistici, analizzando l'intero insieme di valori inerenti all'atributo interessato.

Un ulteriore grande problema relativo alla consistenza dei dati riguarda il concetto di rumore. Quest'ultimo indica un errore casuale che influisce negativamente sulla raccolta dei dati e se mantenuto, anche sull'applicazione delle tecniche di data mining. Le fonti di errore scaturiscono principalmente da errori impliciti introdotti da strumenti di misurazione o da errori casuali introdotti durante la fase di raccolta dei dati in database. Una delle tecniche maggiormente utilizzate consiste nell'operazione di "Binning". Questa tecnica tende a eliminare il problema del rumore uniformando il valore di un determinato dato attraverso l'analisi di tutti i valori, o di una parte, appartenenti allo stesso attributo o comunque aventi lo stesso significato. Dopo aver ordinato e raggruppato i dati in bin, attraverso criteri di vario tipo, si

procede ad uniformarli attraverso la loro sostituzione con un valore rappresentativo del bin stesso, come ad esempio la media. Un'altra tecnica consiste nel rilevare gli errori attraverso l'utilizzo di tecniche di clustering che, come discusso nel paragrafo 3.2.6, permettono di escludere valori non ritenuti consoni agli altri.

3.3.2 Data Integration e Data Transformation: problemi e soluzioni

Con il termine Data Integration si intende l'operazione di incorporare dati ottenuti da sorgenti diverse, al fine di ottenere un unico database completo e fornito di tutte le informazioni utili per l'analisi successiva. Questa operazione può però risultare soggetta ad alcuni problemi rilevanti.

Uno dei principali problemi scaturito da questa fase è quello della ridondanza. L'operazione di integrazione di dati [8] appartenenti a fonti differenti tra di loro, può condurre ad ottenere attributi di ugual significato o facilmente deducibili da altri. Una delle soluzioni potrebbe essere l'analisi di correlazione, che permette di misurare quanto gli attributi siano fortemente correlati tra di loro, in modo da eliminarne uno ritenuto totalmente ridondante. Un altro problema legato al concetto di data integration riguarda la semantica dei dati, la quale solitamente risulta differente tra le varie sorgenti in termini di informazioni e granularità di acquisizione.

Dopo aver integrato più fonti di dati in modo da avere una visione unica del dataset da analizzare, si procede con la fase di data trasformation, tenendo in considerazione successivamente la possibilità di ridurre il dataset risultante in termini di numero di attributi o di range di valori numerici possibili per determinati attributi. La fase di data transformation consiste nel trasformare e modificare la struttura dei dati in maniera tale da eliminare ulteriori criticità presenti nel dataset risultante dall'applicazione delle tecniche relative alla fase di data integration, ma soprattutto consiste di operazioni tali da rendere la forma dei dati idonea per le successive tecniche di data mining. I passi principali che costituiscono questa fase includono:

- la formulazione di nuovi attributi per mezzo delle informazioni risiedenti altrove all'interno del dataset, per favorire ambienti di analisi specifici;
- la gestione di valori anomali è solitamente un passo fondamentale, ma potrebbe non essere necessario in base al tipo di contesto, come ad esempio nel caso del rilevamento di frode, in cui, mediante l'utilizzo

delle tecniche di clustering, è possibile determinare quali valori siano o meno sospetti;

- il concetto di aggregazione, il quale consiste solitamente nel modificare la granularità temporale dei dati in modo da compattarli secondo tecniche specifiche e in base al tipo di analisi da effettuare nella fase di data mining;

- la discretizzazione è uno step necessario per la modifica della distribuzione dei dati prevalentemente di tipo numerico, in modo da normalizzarli e da delinearne una distribuzione discreta suddividendoli in range, con lo scopo di adottare al meglio alcune tecniche successive che non supportano l'elaborazione di variabili continue.

3.4 La classificazione associativa

La classificazione associativa [3, 12, 15] consiste nell'elaborazione di un modello astratto delineato per mezzo delle regole di associazione, capace di etichettare nuovi oggetti non noti a priori in base alla conoscenza dedotta in fase di addestramento, secondo l'utilizzo di un insieme di dati già etichettati, denominato *training set*. Le regole di associazione [1, 14] sono divenute uno strumento essenziale per costruire un classificatore accurato e, così come spiegato nel paragrafo 3.2.7, esse sono sono costituite da una parte conseguente, che identifica una classe, e una parte antecedente, costituita da un insieme di itemset più o meno complesso in base al tipo di caso in esame. Le regole di associazione permettono di identificare le correlazioni più significative tra vari attributi che occorrono spesso nell'insieme di dati da analizzare e l'accuratezza del modello associativo si è dimostrata nel corso degli anni migliore rispetto ad altri approcci, relativamente alla costruzione di un classificatore. Tra i principali vantaggi nell'utilizzo delle regole associative spicca l'efficienza del modello generato, dovuta anche al training set che si ha a disposizione, e l'interpretabilità semplice ed intuitiva delle regole. La fase di costruzione del modello è determinata principalmente da due step:

1. estrazione delle regole di associazione e ordinamento in base ad alcuni parametri di interesse;

2. applicazione di tecniche che riducono l'insieme di regole generate: nel caso di dataset di proporzioni significative, il numero di regole estratte potrebbe risultare rilevante a tal punto da richiedere la definizione di alcune tecniche di taglio per poterlo facilmente ridurre.

Il numero delle regole di associazione ricavate dipende principalmente dalla dimensione del dataset da analizzare e nel caso in cui quest'ultimo si riveli eccessivamente grande, la generazione delle regole potrebbe determinare un problema in termini di tempo e di calcolo, difficili da gestire. Partendo dall'assunto che un maggior insieme di regole tende a costituire un modello di classificazione completo e maggiormente dettagliato, è stato introdotto il concetto di *lazy pruning* (taglio pigro), che consiste in una tecnica da adottare sulla totalità di regole estratte per ridurne il numero al minimo possibile, in modo tale da eliminare soltanto le regole che non forniscono alcun supporto nell'etichettatura di nuovi dati. Questo concetto era già stato introdotto in passato da altri approcci, molti dei quali hanno però riscontrato un ostacolo significativo nel ricoprire il vasto insieme di regole, provocando così una perdita sostanziale di informazioni. Per evitare questo problema, è stato applicato il classificatore associativo allo stato dell'arte L^3 (*Live and Let Live*) [3] per la classificazione di dati strutturati, il quale permette la costruzione del modello mediante la tecnica del lazy pruning e mediante una rappresentazione in forma compatta delle regole, affinchè l'intero insieme venga mostrato nella sua interezza. Dopo aver quindi estratto le regole di associazione ed aver elaborato il modello di classificazione mantenendo un numero di regole il più elevato possibile, successivamente si procede alla classificazione di nuovi dati non etichettati. Quest'ultima operazione è fatta in modo tale da considerare in un primo tempo un insieme di regole definite "ad alta qualità" e solo qualora questo insieme non soddisfi la classificazione di alcuni dati, viene preso in considerazione anche il secondo insieme di regole, le quali non sono state utilizzate nella realizzazione del modello, ma che potrebbero comunque risultare utili.

3.4.1 L'algoritmo Live and Let Live

Relativamente alla fase di taglio delle regole di associazione ricavate durante la costruzione del modello, il classificatore L^3 utilizza un approccio denominato "lazy pruning", che consiste nell'eliminare soltanto le regole che classificano in maniera controproducente i dati etichettati appartenenti al training set, nella fase di addestramento. In questo modo, rispetto agli altri approcci utilizzati da altri classificatori, si ottiene un vasto insieme di regole solitamente efficace per gli obiettivi di classificazione. L^3 permette di suddividere il suddetto insieme di regole principalmente in tre categorie:

1. regole utilizzate: le regole appartenenti a questa categoria sono state impiegate nella fase di addestramento per costruire il modello di clas-

sificazione e rappresentano le più importanti proprietà relative ad ogni classe;

2. regole di scorta: questa categoria comprende tutte le regole che non sono state utilizzate nella fase di training, ma che potrebbero comunque risultare utili nel caso in cui la prima tipologia di regole non fosse pienamente efficace rispetto alla classificazione di nuovi elementi;

3. regole dannose: il classificatore elimina questo insieme di regole poiché esse si sono dimostrate deleterie nella fase di training, diminuendo l'accuratezza del modello ricavato.

In corrispondenza delle prime due categorie sopra discusse, il classificatore distingue rispettivamente le regole di primo livello e le regole di secondo livello. Le regole ricavate nella fase di estrazione vengono ordinate principalmente in base alla confidenza, al supporto e soprattutto alla loro lunghezza, dato che a parità delle prime due misure, viene considerata migliore la regola più lunga poiché ritenuta più accurata, denominata *specialistic rule*, piuttosto che una corta, denominata *general rule*. Quest'ultima categoria include le regole che verranno prese in considerazione qualora le rispettive regole specialistiche non riescano a classificare correttamente nuovi dati. Così come accennato precedentemente, è possibile associare al concetto di lazy pruning l'introduzione di una forma compatta dell'insieme di regole, affinchè esso venga rappresentato al meglio senza perdita di informazioni.

Una caratteristica molto importante del classificatore L^3 consiste nell'utilizzo di più soglie di supporto minimo per selezionare i pattern più frequenti nella fase di generazione delle regole. La scelta di un'unica soglia di supporto minimo potrebbe risultare non idonea qualora si avesse un insieme di dati ripartito in maniera non uniforme rispetto alle classi, poiché si rischierebbe di trascurare la classe minoritaria nel caso di una soglia di supporto troppo alta e viceversa, si otterrebbe un numero troppo elevato di regole nell'altro caso, provocando problematiche sia in termini di gestione che di tempo. Si ricorre quindi all'adozione di più soglie di supporto scelte e ponderate in base alla frequenza di ogni classe, in modo da ottenere la giusta proporzione regole/classe [3]. Per testare l'accuratezza del classificatore, viene adottata la tecnica del *cross validation*. Essa consiste nel partizionare il dataset in "n" parti in modo tale da utilizzare iterativamente n-1 partizioni come training set e 1 partizione come testset. Le misure utilizzate per testare alcune caratteristiche del classificatore sono deducibili dalla matrice di confusione ricavata per ognuno dei gruppi partizionati precedentemente e sono accuratezza, precisione e recall, di cui si è discusso nel paragrafo 3.2.2. Qualora le tempistiche di elaborazione nella fase di apprendimento siano troppo elevate

e qualora l'accuratezza del modello costruito risulti notevolmente bassa, è possibile tenere conto di una modifica delle soglie di supporto e/o confidenza minimi.

Capitolo 4

Analisi dei dati telematici e aziendali

In questo capitolo, verranno presentate e discusse le fasi del processo di analisi dei dati aggregati provenienti dalle black box installate sui veicoli degli assicurati e dei dati archiviati riguardanti gli assicurati stessi. Percorrendo i vari step, si arriverà a definire nel dettaglio la generazione dei dataset su cui applicare il classificatore L^3 [3], in modo tale da ottenere gli obiettivi prefissati dal lavoro di tesi, ovvero la caratterizzazione delle polizze rispetto ai livelli di rischio nel medio-lungo termine.

4.1 Struttura del processo applicativo

Il lavoro di tesi è strutturato principalmente in tre fasi, scandite dai passi richiesti dal processo di analisi e preparazione dei dati e dalla successiva applicazione delle tecniche di data mining in base ai casi di studio. Il sistema per l'analisi dei dati aggregati, rappresentato in Fig 4.1, prevede tre moduli principali:

1. *Data cleaning and preparation.* Questa fase è finalizzata alla preparazione, pulizia e modifica della struttura dei dati raccolti dalle scatole nere;

2. *Data analytics.* In questa fase vengono applicate alcune tecniche di data mining in base al tipo di analisi e ai casi di studio. In particolare, viene applicato il classificatore L^3 per produrre modelli associativi differenziati in base alla tipologia di input fornito e dai parametri scelti dall'utente;

3. *Expert-driven result visualization and validation.* La terza fase è incentrata sull'esplorazione dei risultati e sulla valutazione della loro bontà mediante l'analisi di misure statistiche e della coerenza dei modelli, e sulla visualizzazione dei contenuti e dei risultati generati attraverso interfacce idonee per il supporto alle decisioni, in modo da renderli facilmente fruibili ed utilizzabili in futuro.

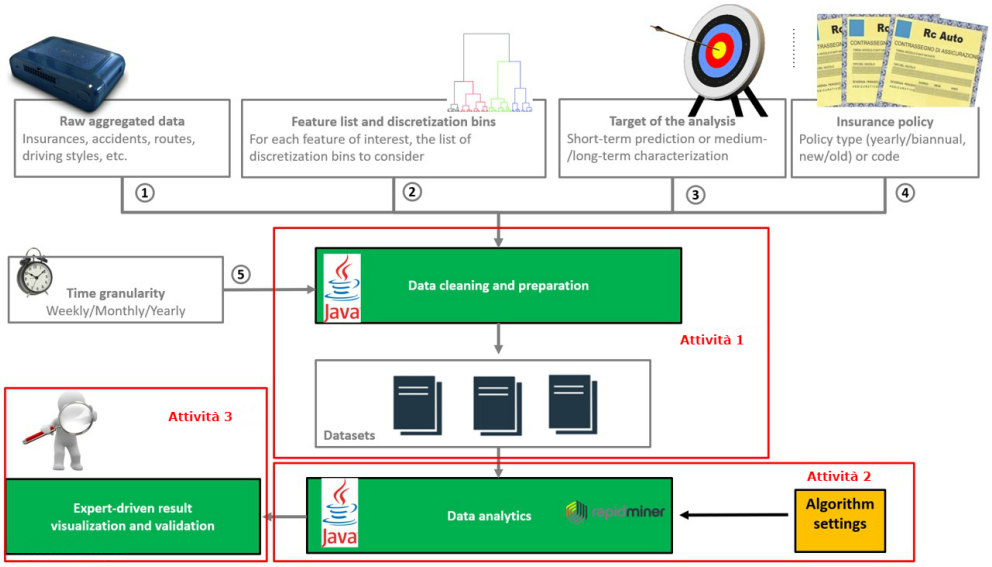

Figura 4.1: *Struttura del processo di analisi*

Il modulo Data cleaning and preparation è progettato in modo tale da rendere parametrico gran parte del processo di trasformazione dei dati forniti, minimizzando i costi di riconfigurazione del sistema e massimizzandone la flessibilità. Il modulo oggetto di questa attività produce un insieme di dataset in formato relazionale che soddisfano i requisiti indicati in input dall'utente. È stato quindi sviluppato un applicativo Java che permette all'utente di scegliere:

- il sottoinsieme di dati aggregati da analizzare;
- la lista di feature d'interesse per le analisi successive;
- le regole per l'eventuale trasformazione dei dati originali da valori numerici a intervalli discreti, nel caso in cui gli algoritmi di machine learning, applicati nei passi successivi, lo richiedano;
- la tipologia di polizze da analizzare;

- la granularità con cui aggregare i dati raccolti su base temporale, ovvero distinguere tra aggregazione settimanale, mensile e annuale, in base agli obiettivi e al caso di studio in esame.

I dataset prodotti nella fase precedente sono processati dal secondo modulo, il quale ha il compito di analizzare i dati preparati attraverso algoritmi di machine learning opportunamente configurati. I risultati ricavati dal modulo di analisi dipendono dal caso di studio e sono, rispettivamente, le predizioni del livello di rischio e le descrizioni delle correlazioni più significative presenti tra le feature che caratterizzano il dataset e i livelli di rischio. Il terzo modulo consente di validare i risultati ottenuti dai passi precedenti, mediante appositi processi statistici per garantirne la significatività, nonchè valutare la rilevanza, pertinenza e coerenza delle correlazioni identificate. La validazione include, inoltre, la scelta e la generazione di opportuni schemi di visualizzazione dei risultati raccolti, in funzione delle necessità e degli obiettivi da perseguire in ciascun caso di studio.

4.2 Il processo di Knowledge Discovery

Per l'applicazione delle tecniche di data mining sui dati delle polizze, integrati con quelli raccolti mediante la telematica, è stato necessario adottare una fase di preprocessing per gestire problematiche legate alla reperibilità dei dati e all'integrazione di vari dataset. Gli step caratterizzanti l'analisi del dataset in esame seguono il processo di Knowledge Discovery in Databases (KDD) [7], di cui si è discusso ampiamente nel capitolo 3, e sono:

1. *Selection:* dai dati raccolti e aggregati dalle scatole nere e dai dati operativi relativi a polizze e sinistri verranno dapprima selezionate le informazioni d'interesse, ovvero verrà estratto un sottoinsieme di feature di potenziale interesse per l'analisi svolta;

2. *Preprocessing:* in questa fase, i dati selezionati verranno processati per eliminare rumore, ridondanze e per renderli idonei all'analisi successiva; in particolare, verranno identificati e opportunamente gestiti i dati mancanti, per poi applicare passi di discretizzazione e uniformare i dati relativi alle medesime feature in maniera tale da rendere affidabili le analisi statistiche e i processi di generazione dei modelli predittivi;

3. *Transformation:* i dati preprocessati saranno poi trasformati nel formato idoneo per applicare gli algoritmi di machine learning, utili per

l'estrazione della conoscenza. In particolare, verranno applicati algoritmi di classificazione diversi a seconda dei casi di studio. Di conseguenza, relativamente al tipo di analisi da effettuare, i dati andranno preventivamente aggregati temporalmente e saranno memorizzati in vari dataset nel formato relazionale, differenziati in base alle tipologie di polizze;

4. *Data mining:* sui dataset generati nella fase precedente, verranno applicati algoritmi di machine learning di tipo supervisionato. In particolare, si applicheranno tecniche di classificazione che analizzano una collezione di dati storici per estrapolare modelli predittivi di una o più feature inerenti ai livelli di rischio;

5. *Interpretation/Evaluation:* questa fase si prefigge l'obiettivo di analizzare i risultati ottenuti dalle tecniche applicate nella fase precedente in modo da testarne accuratezza, validità e coerenza mediante l'utilizzo di opportune tecniche statistiche. Grazie all'applicazione del classificatore L^3, sarà possibile analizzare il contenuto dei modelli associativi prodotti, in modo tale da descrivere le correlazioni tra le feature non predittive e quelle predittive e capire i motivi per cui il modello produce determinate predizioni.

Al termine dell'intero processo di analisi, verranno estratti i pattern più significativi e altamente validi in base al contesto, in modo tale da definire l'effettivo valore aggiunto dei dati raccolti dalla telematica e formulare schemi indicativi facilmente interpretabili a supporto delle compagnie assicurative.

4.2.1 Selezione delle feature

Questa fase consiste nell'acquisizione di una conoscenza tale da riuscire a selezionare e mantenere informazioni importanti in base all'analisi da effettuare, rispetto alla totalità del dataset in esame. Attraverso alcuni studi preliminari sui dati, verificandone la consistenza e l'accuratezza, è stato possibile estrarre un sottoinsieme rilevante di attributi tra quelli presenti nel dataset in possesso. I principali motivi che hanno condotto allo scarto di alcuni attributi e di cui si discuterà in maniera approfondita nelle fasi successive, sono i seguenti:

- preponderanza di valori mancanti e/o nulli: questo fenomeno incide negativamente sull'analisi, causando l'estrazione di risultati fuorvianti e poco significativi;

- attributi ridondanti: alcuni attributi riportano valori esattamente identici ad altri, a seguito dell'integrazione di più dataset;

- attributi ritenuti non interessanti ai fini analitici: in base al caso di studio applicato, è stato necessario scartare alcune informazioni piuttosto che altre per la loro poca significatività se rapportate agli obiettivi preposti.

Nel dettaglio, in riferimento ai dati elencati nel capitolo 2, sono stati ritenuti altamente validi gli attributi rappresentati in Tabella 4.1, con la particolare attenzione rivolta agli indicatori del livello di rischio accuratamente selezionati.

Tabella 4.1: *Selezione delle feature ritenute importanti per l'analisi*

Categoria	**Attributo**
Aggregazioni temporali	-Giorno rilevamento dati -Data inizio del periodo di osservazione -Ora inizio del periodo di osservazione -Data fine del periodo di osservazione -Ora fine del periodo di osservazione -Giorno -Settimana -Anno -Data del sinistro
Percorrenze	-Numero totale di metri -Metri percorsi in autostrada -Metri percorsi in città -Metri percorsi su strade di periferia -Metri percorsi nella fascia diurna -Metri percorsi nella fascia notturna
Stile di guida	-Eccessi di velocità in città durante la fascia diurna

Continua nella prossima pagina

Continua dalla pagina precedente

Categoria	Attributo
	-Eccessi di velocità in autostrada durante la fascia diurna -Eccessi di velocità su strade extra-urbane durante la fascia diurna -Eccessi di velocità altrove durante la fascia diurna -Eccessi di velocità in città durante la fascia notturna -Eccessi di velocità in autostrada durante la fascia notturna -Eccessi di velocità su strade extra-urbane durante la fascia notturna -Eccessi di velocità altrove durante la fascia notturna -Accelerazioni brusche in città durante la fascia diurna -Accelerazioni brusche in autostrada durante la fascia diurna -Accelerazioni brusche su strade extra-urbane durante la fascia diurna -Accelerazioni brusche altrove durante la fascia diurna -Accelerazioni brusche in città durante la fascia notturna -Accelerazioni brusche in autostrada durante la fascia notturna -Accelerazioni brusche su strade extra-urbane durante la fascia notturna -Accelerazioni brusche altrove durante la fascia notturna -Decelerazioni brusche in città durante la fascia diurna -Decelerazioni brusche in

Continua nella prossima pagina

Continua dalla pagina precedente

Categoria	Attributo
	autostrada durante la fascia diurna -Decelerazioni brusche su strade extra-urbane durante la fascia diurna -Decelerazioni brusche altrove durante la fascia diurna -Decelerazioni brusche in città durante la fascia notturna -Decelerazioni brusche in autostrada durante la fascia notturna -Decelerazioni brusche su strade extra-urbane durante la fascia notturna -Decelerazioni brusche altrove durante la fascia notturna -Curvature ad alta velocità in città durante la fascia diurna -Curvature ad alta velocità in autostrada durante la fascia diurna -Curvature ad alta velocità su strade extra-urbane durante la fascia diurna -Curvature ad alta velocità altrove durante la fascia diurna -Curvature ad alta velocità in città durante la fascia notturna -Curvature ad alta velocità in autostrada durante la fascia notturna -Curvature ad alta velocità su strade extra-urbane durante la fascia notturna -Curvature ad alta velocità altrove durante la fascia notturna -Cambi repentini di direzione in città durante la fascia diurna

Continua nella prossima pagina

Continua dalla pagina precedente

Categoria	**Attributo**
	-Cambi repentini di direzione in autostrada durante la fascia diurna -Cambi repentini di direzione su strade extra-urbane durante la fascia diurna -Cambi repentini di direzione altrove durante la fascia diurna -Cambi repentini di direzione in città durante la fascia notturna -Cambi repentini di direzione in autostrada durante la fascia notturna -Cambi repentini di direzione su strade extra-urbane durante la fascia notturna -Cambi repentini di direzione altrove durante la fascia notturna
Informazioni sulle polizze	-Numero della polizza -Inizio copertura della polizza -Fine copertura della polizza -Data di scadenza polizza -Nuova polizza -Durata polizza (Semestrale o Annuale) -Tipo di polizza -Massimale
Informazioni sui veicoli	-Cavalli fiscali -Cavalli potenza -Alimentazione del veicolo -Tipo di veicolo -Utilizzo del veicolo -Anno di prima immatricolazione -Età del veicolo -Data dell'ultima voltura
Dettagli sull'assicurato	-Provincia

Continua nella prossima pagina

Continua dalla pagina precedente

Categoria	Attributo
	-Area territoriale -Regione -Codice geografico -Età della patente -Classe di merito -Classe agevolata -Flag proprietario-contraente -Anni senza assicurazione -Attestato di rischio -Distanza dall'ultimo sinistro -Tipo di danno -CAP -Sesso -Professione -Flag proprietario-contraente -Ritardo ottenimento patente -Età dell'assicurato -Numero di rinnovi -Anni di possesso del veicolo
Indicatori del livello di rischio	-Numero di incidenti con colpa grave causati -Numero di incidenti senza colpa grave causati -Numero di incidenti causati -Costo degli incidenti causati

Si conclude dalla pagina precedente

4.2.2 Preprocessing dei dati

La fase di preprocessing ha l'obiettivo di rimuovere le criticità e le inconsistenze dei dati, nonché gestire i valori mancanti e i dati affetti da rumore, per poi rendere il loro formato adatto alle tecniche da utilizzare nei passi successivi. Così come discusso nel capitolo 3, è possibile distinguere due sub-step inerenti a questa fase, ovvero *data cleaning* e *data integration*. La prima applica strategie per risolvere le inconsistenze insite nella reperibilità dei dati, mentre la seconda si occupa di gestire le problematiche scaturite dall'integrazione di dati derivati da sorgenti differenti. Congiuntamente allo

scarto delle informazioni durante la fase di selezione delle feature, sono stati affrontati i problemi della fase di data integration, tra i quali è possibile citare principalmente il fenomeno di ridondanza. L'integrazione di dati derivati da più sorgenti può indurre a questo fenomeno, scandito principalmente da due cause:

- attributi di significato uguale, ma con una intestazione differente nei vari dataset da integrare;
- attributi con una intestazione differente, i cui valori sono uguali nel significato ma diversi nel formato.

Questo ha favorito lo scarto di ulteriori attributi poiché ritenuti completamente inadatti a contribuire alla determinazione di utili risultati analitici. Inoltre, alcuni dati presentano incoerenze di formato che ne impediscono la corretta analisi, poichè alcuni algoritmi di classificazione gestiscono esclusivamente dati in formato discreto, per cui è necessario formattarne il contenuto per una migliore comprensione e gestione. Sono quindi state applicate tecniche di data cleaning per ripulire il dato da rumore o inconsistenze, e di discretizzazione, per trasformare i valori continui in intervalli discreti, sostituendo ogni valore con un'opportuna etichetta inerente all'intervallo di corrispondenza. In base alla preponderanza più o meno sostanziale dei valori mancanti per ogni attributo, sono stati adottati differenti metodi, quali:

- eliminare l'intera colonna: questa tecnica risulta valida soltanto nel caso in cui tutti o quasi tutti i valori relativi alla colonna in esame siano mancanti o uguali a null/zero;
- sostituire i campi mancanti con un valore altamente significativo per quell'attributo, come ad esempio la media. Questa tecnica viene utilizzata nel caso in cui i valori mancanti non siano eccessivamente tanti, dato che eliminare l'intero attributo potrebbe comportare in questo caso la perdita di informazioni interessanti;
- nel caso in cui i campi con valori nulli siano molto pochi, vengono eliminate le tuple.

I principali metodi di discretizzazione [11] dei dati per gestire il problema del rumore, ovvero eliminare gli outliers e discretizzare i valori continui in intervalli discreti, sono:

1. *Equi-width:* consiste nella divisione del range di valori assunti da una feature in intervalli di ampiezza fissa;

2. *Equi-depth:* permette di suddividere il range di valori assunti da una feature in intervalli contenenti lo stesso numero di valori;

3. *Entropy-based:* questa tecnica sfrutta il concetto di entropia calcolata rispetto a una classe di riferimento con lo scopo di suddividere il range di valori di una determinata feature in intervalli;

4. *Clustering:* utilizza metodi di clustering come criterio di suddivisione del range di valori di una feature in intervalli.

Inizialmente sono state analizzate le distribuzioni dei dati inerenti a tutti gli attributi, in modo tale da distinguere le feature categoriche da quelle numeriche, e inerentemente a quest'ultime, verificare per quali attributi si dovessero applicare le tecniche di discretizzazione secondo le strategie sopra descritte o ad hoc, in base al contesto. In particolare, sono stati generati gli intervalli relativi a ciascuna delle feature di seguito elencate, secondo le metodologie 1 e 2:

- informazioni relative alle percorrenze;

- informazioni relative allo stile di guida;

- cavalli potenza;

- età della patente;

- ritardo ottenimento patente;

- età del veicolo;

- anni di possesso del veicolo;

- costo degli incidenti causati.

Così come accennato precedentemente, per alcune altre feature sono state utilizzate delle discretizzazioni ad hoc, in base al tipo di attributo in esame. Per le feature numeriche con un numero molto limitato di valori, quali ad esempio "Anni senza assicurazione", "Attestato di rischio", "Distanza dall'ultimo sinistro", sono stati mantenuti i valori originali analogamente alla feature del "Massimale", dato che i singoli possibili valori sono altamente significativi per le analisi. Per la maggior parte delle feature relative agli stili di guida, i valori sono stati trasformati in valori booleani (nessuna infrazione o una infrazione o più) su base settimanale e mensile, mentre i relativi valori

sono stati discretizzati mediante le due tecniche equi-depth ed equi-width su base annuale.

Gli indicatori del livello di rischio sono stati discretizzati in valori booleani, in modo da indicare la presenza o meno di un incidente o più, in base al tipo di dataset in esame. Per tutte le altre feature di tipo categorico, quindi non numeriche o per le quali è necessario mantenere ogni valore poiché altamente significativo, il passo di discretizzazione sopra descritto non è applicabile e di conseguenza sono state mantenute inalterate. Per tali feature è stata verificata la correttezza dei valori disponibili, talvolta inconsistenti rispetto al formato richiesto. Per ovviare a tali problematiche, è stato uniformato il relativo contenuto per permettere un'analisi affidabile delle frequenze dei valori assunti.

4.2.3 Trasformazione dei dati

La fase di trasformazione ha l'obiettivo di generare i dataset su cui saranno costruiti i modelli di classificazione per realizzare gli obiettivi preposti. Essa consiste principalmente in un'aggregazione dei dati originali su tre differenti livelli di granularità: settimanale, mensile e annuale. Le aggregazioni settimanali e mensili sono state utilizzate per effettuare le predizioni del livello di rischio nel breve periodo, mentre le aggregazioni annuali per caratterizzare i livelli di rischio associati alle polizze. L'aggregazione settimanale ha permesso di produrre un dataset di tipo relazionale [2], come rappresentato nella Tabella 4.2, contenente un record per ogni chiave "Numero di polizza" - "Settimana".

Tabella 4.2: *Schema del dataset generato mediante aggregazione settimanale*

Numero della polizza	**Settimana**	**Classe di merito**	**...**

I valori numerici associati alle percorrenze, agli stili di guida e al numero di sinistri commessi dagli assicurati sono aggregati opportunamente per polizza e settimana dell'anno. Nel caso in cui in una specifica settimana dell'anno un veicolo non abbia circolato, il relativo record non è stato inserito per considerare solo i periodi in cui il veicolo è stato attivo. Tuttavia, nella fase di analisi sono state considerate rappresentazioni differenti, includendo anche tuple in cui l'assicurato non ha circolato.

Durante la fase di trasformazione è stato inoltre introdotto un nuovo attributo, ovvero l'attributo "Mese", in modo tale da poter realizzare l'aggregazione mensile. Quindi, analogamente al caso precedente, questa aggrega-

zione ha prodotto un dataset di tipo relazionale, illustrato nella Tabella 4.3, contenente un record per ogni chiave "Numero della polizza" - "Mese".

Tabella 4.3: *Schema del dataset generato mediante aggregazione mensile*

Numero della polizza	**Mese**	**Classe di merito**	**...**

Infine, per costruire modelli predittivi specifici per ogni polizza, è stato possibile aggregare le feature del dataset ricavato dalle fasi precedenti rispetto al numero di polizza, con granularità annuale. Di conseguenza, è stato generato un dataset relazionale tale che abbia come chiave identificativa di ogni tupla, la feature "Numero della polizza", illustrato in Tabella 4.4.

Tabella 4.4: *Schema del dataset generato mediante aggregazione annuale*

Numero della polizza	**Età**	**Classe di merito**	**...**

Il caso di studio relativo a questo lavoro di tesi ha come oggetto l'ultimo schema generato in modo tale da fornire la caratterizzazione delle polizze a lungo termine rispetto ai quattro livelli di rischio definiti in Tabella 4.1 (Indicatori del livello di rischio). Dopo aver ottenuto il dataset annuale, è stato possibile partizionarlo ulteriormente utilizzando i valori di due feature, quali:

1. Nuova polizza: permette di distinguere le polizze nuove da quelle già rinnovate;
2. Durata polizza: permette di distinguere le polizze semestrali da quelle annuali.

determinando quattro distinti dataset:

- polizze nuove annuali;
- polizze nuove semestrali;
- polizze vecchie annuali;
- polizze vecchie semestrali.

Ogni caratterizzazione quindi descrive le correlazioni più significative e le co-occorrenze tra i livelli di rischio sopra definiti e i valori delle feature presenti all'interno dei dati, distinguendo le varie tipologie di polizze per attenzionare al meglio le varie casistiche.

4.3 Caso di studio

La costruzione dei modelli caratterizzanti i livelli di rischio è basata sull'applicazione di un algoritmo di classificazione associativo allo stato dell'arte, denominato L^3 [3]. L'algoritmo, progettato per predire il valore ignoto di una determinata feature, denominata classe, in base ai valori noti delle feature restanti, genera un modello interpretabile basato su regole di associazione [1]. Per raggiungere l'obiettivo prefissato nel caso di studio in esame, l'algoritmo viene applicato allo scopo di generare ed esplorare il contenuto del modello ricavato. Quindi, durante la fase di analisi dei dati, si costruisce un modello basato su regole di associazione a partire dal dataset delle polizze annuali, fissando come obiettivo della predizione i valori di una delle feature definite come livelli di rischio. Il modello è composto da un insieme di pattern che descrivono le correlazioni più significative tra il livello di rischio e le restanti feature presenti nei dati, dalle quali, grazie a un'approfondita esplorazione dei modelli, vengono estratti gli schemi più rilevanti per definire la caratterizzazione delle polizze.

Nel dettaglio, dopo aver ricavato il dataset annuale dal modulo di data cleaning and preparation, vengono generati tanti dataset quante sono le categorie di feature, discusse nel paragrafo 4.3.1, per poi partizionarli una prima volta relativamente alle tipologie di polizze e una seconda volta riguardo ai quattro livelli di rischio definiti. Questa seconda fase di partizionamento permette di caratterizzare le feature di ogni categoria relativamente a ciascun livello di rischio per volta. Vengono quindi generati diversi modelli del classificatore L^3, uno per ciascuna combinazione tra gli indicatori del livello di rischio e le tipologie di polizze.

Lo scopo principale della generazione dei modelli associativi consiste nell'estrazione di tutte le combinazioni frequenti di item, ovvero le combinazioni di coppie "feature" - "valore della feature", che sono frequentemente associate alla medesima polizza. Ad esempio, la regola:

(*(Provincia, Torino), (Numero totale di metri, 30000)* $\Rightarrow$ *(Numero di incidenti causati, 1)*

indica la co-occorrenza di tre itemset. Il segno di "$\Rightarrow$" separa il corpo della regola, ovvero l'itemset che si trova a sinistra, dalla testa della regola, ovvero l'itemset che si trova a destra. Così come discusso ampiamente nel capitolo 3, vi sono tre parametri fondamentali che misurano le proprietà della regola, quali:

1. supporto: indica la frequenza della co-occorrenza degli itemset presenti

nella regola rispetto al totale delle transazioni;

2. confidenza: misura la probabilità condizionata di occorrenza della testa, dato il corpo;

3. lift: nel caso in cui la distribuzione dei dati non sia uniforme rispetto alle classi, questa misura tiene conto anche della frequenza della testa della regola, indicandone livello e tipologia di correlazione.

L'algoritmo L^3 produce un sottoinsieme di regole mirate alla classe da predire e aventi elevati valori di supporto e confidenza. Nel caso di studio in esame, per caratterizzare i livelli di rischio si è scelto di considerare le regole aventi nella testa, item descrittivi del livello di rischio. Di conseguenza, così come riportato dalla regola sopra definita, tutte le regole estratte dal classificatore in questione avranno come testa della regola un item del tipo "livello di rischio analizzato" - "valore", mentre nel corpo saranno presenti combinazioni di item appartenenti ad una o più feature arbitrarie.

4.3.1 Definizione delle categorie di feature

Per valutare la correlazione tra i livelli di rischio e le diverse tipologie di feature presenti nel dataset ricavato dalle fasi precedenti, sono state determinate tre categorie di feature, come illustrato nella Tabella 4.5, quali:

1. Caratteristiche delle polizze: informazioni specifiche sulla polizza, sul veicolo a cui la polizza è associata e sull'assicurato intestatario della polizza;

2. Stile di guida: informazioni quantitative rilevate dalle scatole nere installate sui veicoli riguardo allo stile di guida degli assicurati;

3. Percorrenze: informazioni quantitative reperite dalle scatole nere riguardanti i percorsi effettuati.

La caratterizzazione dei livelli di rischio viene effettuata sia separatamente per ciascuna categoria, al fine di correlare i valori delle feature di una categoria specifica con il livello di rischio considerato, sia combinando le feature di categorie differenti; la seconda tipologia, denominata *Mixed-feature*, ha l'obiettivo di correlare i valori delle feature appartenenti a categorie differenti per uno specifico livello di rischio. Per applicare il classificatore ed effettuare le analisi successive, viene ricavato un sottoinsieme rappresentativo, ovvero il 30%, delle feature appartenenti a ciascuna delle categorie sopra descritte.

La selezione si basa sia sulla rilevanza nel dominio specifico sia sulla significatività del dato in relazione alla presenza di eventuali valori mancanti, nulli e inconsistenti, sulla base delle conclusioni tratte durante l'attività relativa al modulo "Data cleaning and preparation".

Tabella 4.5: *Sottoinsiemi delle feature per ciascuna categoria*

Categoria	**Attributo**
Caratteristiche delle polizze	-Durata polizza -Classe di merito -Nuova polizza -Provincia -Sesso -Ritardo ottenimento patente -Anno di prima immatricolazione -Professione -Età dell'assicurato -Numero di incidenti con colpa grave causati -Numero di incidenti senza colpa grave causati -Numero di incidenti causati -Costo degli incidenti causati
Stile di guida	-Eccessi di velocità in città durante la fascia diurna -Eccessi di velocità in autostrada durante la fascia diurna -Accelerazioni brusche in città durante la fascia diurna -Accelerazioni brusche in autostrada durante la fascia diurna -Decelerazioni brusche in città durante la fascia diurna -Curvature ad alta velocità in città durante la fascia diurna -Curvature ad alta velocità in autostrada durante la fascia diurna -Cambi repentini di direzione in città durante la fascia diurna

Continua nella prossima pagina

Continua dalla pagina precedente

Categoria	**Attributo**
	-Cambi repentini di direzione in autostrada durante la fascia diurna -Durata polizza -Nuova polizza -Numero di incidenti con colpa grave causati -Numero di incidenti senza colpa grave causati -Numero di incidenti causati -Costo degli incidenti causati
Percorrenze	-Numero totale di metri -Metri percorsi in autostrada -Metri percorsi in città -Metri percorsi in periferia -Metri percorsi nella fascia diurna -Metri percorsi nella fascia notturna -Durata polizza -Nuova polizza -Numero di incidenti con colpa grave causati -Numero di incidenti senza colpa grave causati -Numero di incidenti causati -Costo degli incidenti causati
Mixed-feature	-Numero totale di metri -Metri percorsi in autostrada -Eccessi di velocità in città durante la fascia diurna -Eccessi di velocità in autostrada durante la fascia diurna -Eccessi di velocità in città durante la fascia notturna -Accelerazioni brusche in città durante la fascia diurna

Continua nella prossima pagina

Continua dalla pagina precedente

Categoria	Attributo
	-Accelerazioni brusche in città durante la fascia notturna -Accelerazioni brusche in autostrada durante la fascia notturna -Curvature ad alta velocità in città durante la fascia diurna -Cambi repentini di direzione in città durante la fascia diurna -Cambi repentini di direzione in autostrada durante la fascia diurna -Cavalli fiscali -Data del sinistro -Durata polizza -Classe di merito -Nuova polizza -Alimentazione del veicolo -Sesso -Ritardo ottenimento patente -Anno di prima immatricolazione -Età del veicolo -Età dell'assicurato -Numero di incidenti con colpa grave causati -Numero di incidenti senza colpa grave causati -Numero di incidenti causati -Costo degli incidenti causati

Si conclude dalla pagina precedente

4.3.2 Generazione dei profili caratterizzanti i livelli di rischio

Relativamente alla definizione delle categorie di feature stabilita nel paragrafo 4.3.1, i modelli generati rispetto alle prime tre categorie di feature permettono di ricavare una descrizione delle polizze associate ad un determinato livello di rischio mediante profili specifici per una particolare categoria. Ad

esempio, considerando le feature appartenenti alla categoria "Stile di guida", si considerano solo le regole aventi nella testa una combinazione di feature pertinenti esclusivamente allo stile di guida. Un esempio di regola di questa tipologia è riportato di seguito:

(*Numero di cambi repentini di direzione in tratti urbani* > *50*), (*Numero di decelerazioni improvvise in autostrada* > *10*) ⇒ (*Numero di incidenti causati* > *0*)

con le relative misure di interesse:

- supporto = 1.5%;
- confidenza = 85%.

La regola di esempio sopra riportata descrive un profilo riguardante le polizze che causano incidenti. Il suddetto profilo è basato su due caratteristiche peculiari dello stile di guida, misurabili mediante l'utilizzo delle scatole nere: le frenate brusche e le decelerazioni improvvise. La regola indica che la combinazione di due specifici comportamenti inappropriati ripetuti nell'arco di un anno è associata a polizze che causano incidenti nell'85% dei casi. Quindi, tale comportamento può essere considerato come indicativo di un'effettiva propensione al rischio. Le regole estratte dal sottoinsieme "Mixed-feature" invece, combinano le caratteristiche di categorie di feature differenti. Ad esempio, la regola seguente combina una misura legata alle percorrenze con una legata allo stile di guida:

(*Numero di Km percorsi in tratti urbani* > *20000*), (*Numero di decelerazioni improvvise in autostrada* > *10*) ⇒ (*Numero di incidenti causati* > *0*)

con le relative misure di interesse:

- supporto = 1.8%;
- confidenza = 70%.

La regola sopra riportata permette di associare ad un livello di rischio alto i titolari di polizze che percorrono un numero significativo di km in tratti urbani e che effettuano un numero di decelerazioni brusche elevato in tratti autostradali.

La fase di generazione dei vari modelli di classificazione consiste essenzialmente in due step, quali:

1. estrazione delle regole di associazione;
2. minimizzazione dello scarto di regole ritenute deleterie per la classificazione di nuovi dati.

Un aspetto molto importante che ha caratterizzato particolarmente i risultati riguarda due parametri di interesse da dover impostare, relativi alla soglia di supporto minimo e alla soglia di confidenza minima, dei quali si è discusso ampiamente nel capitolo 3. La distribuzione dei dati può inficiare la qualità del processo di estrazione delle regole di associazione, dato che tutti i dataset analizzati sono caratterizzati da un significativo sbilanciamento tra i livelli di rischio associati alle polizze, in base al tipo di dataset in esame. Infatti, la maggioranza delle polizze è associata ad un livello di rischio basso, in quanto nell'anno in questione non ha causato/subito sinistri. Al contrario, una porzione minoritaria del dataset presenta elevati livelli di rischio ed è quindi meritoria di particolare attenzione, in quanto i profili associati a tali livelli di rischio sono quelli maggiormente critici per le compagnie assicurative.

Il processo di generazione delle regole di associazione può risultare poco efficace nel caratterizzare la classe minoritaria, ovvero il livello di rischio alto, in quanto la soglia di supporto minimo scelta, che rappresenta una frequenza minima di co-occorrenza degli item contenuti nella regola, è indipendente dal numero totale di polizze associate a ciascun livello di rischio [3]. Dopo aver quindi effettuato numerose prove affinchè si trovasse un buon compromesso tra le regole appartenenti alla classe maggioritaria e quelle appartenenti alla classe minoritaria, si è scelto di applicare la soglia di supporto globale dell'1%. Nel caso in cui i dati siano distribuiti in maniera bilanciata tra le varie classi, la medesima soglia viene applicata in maniera univoca per tutte le regole, altrimenti la soglia applicata sulle regole relative al livello di rischio minoritario viene incrementata in proporzione alla percentuale di sbilanciamento esistente rispetto alla classe maggioritaria.

4.3.3 Interpretazione dei pattern e definizione dei KPI

Il classificatore L^3 adotta una rappresentazione di tipo compatto e ordinato delle regole in base ad alcuni parametri, in modo tale da ricoprire l'intera totalità dell'insieme di regole ricavate durante la fase di apprendimento, come discusso nel capitolo 3. La selezione dei pattern frequenti di maggiore interesse si svolge principalmente in due fasi. La prima consiste nel selezionare le regole in base a quattro fattori principali, quali:

1. supporto;

2. confidenza;
3. lift;
4. lunghezza.

Il classificatore ordina l'insieme delle regole in base alla tripla combinazione di supporto, confidenza e lunghezza delle regole. Quest'ultimo parametro viene preso in considerazione come uno dei criteri di scelta poichè data una regola lunga, essa descrive caratteristiche più specifiche grazie alla presenza di un numero più o meno considerevole di itemset coinvolti. Successivamente, viene adottato un criterio di scelta basato sull'informazione ottenuta e sulla coerenza degli item relativamente al tipo di contesto, in modo tale da ricavare correlazioni probabilmente nascoste tra feature interessanti per fornire un servizio di valore aggiunto. Alla luce delle regole ottenute, vengono definiti degli indicatori chiave in modo tale da descrivere al meglio ciascun livello di rischio date alcune rilevanti combinazioni di feature, con lo scopo di fornire schemi facilmente interpretabili e altamente usufruibili in futuro.

Capitolo 5

Risultati sperimentali

In quest'ultimo capitolo, sarà discussa la configurazione del classificatore con cui sono stati generati i modelli associativi relativi alle varie categorie di feature e sarà possibile apprezzare alcuni risultati.

5.1 Approccio

L'utilizzo del classificatore L^3 ha permesso la costruzione di vari modelli determinati dalla combinazione dei valori di due punti essenziali, quali categorie delle feature e tipologie delle polizze. Di conseguenza, le regole di associazione sono state estratte in modo tale da identificare i vari casi singolarmente e caratterizzare le polizze di una determinata tipologia rispetto ai quattro livelli di rischio definiti. La distinzione di più categorie di feature ha permesso di attenzionare l'analisi sia analizzando feature esclusivamente della categoria a cui essi appartengono, sia combinando caratteristiche di categorie distinte per studiarne l'impatto.

Per generare i modelli associativi, è stato necessario impostare dei parametri d'interesse, focalizzando l'attenzione sulle soglie da applicare per istruire il classificatore affinchè determinasse quali fossero o meno i pattern più frequenti. Il numero e la distribuzione delle regole rispetto alle varie etichette di classe da determinare costituiscono due fattori fortemente dipendenti dalla scelta dei parametri delle soglie. Per cui, sono state impostate le seguenti soglie globali:

- soglia di supporto minimo: 1%;
- soglia di confidenza minima: 50%.

Nonostante, in linea generale, siano state adottate le suddette soglie, è stato necessario in alcuni casi aumentare la soglia di supporto minimo, poichè

dato in ingresso al classificatore un dataset piuttosto ampio, esso consente di ricavare un numero sostanziale di regole. Questo comporta problemi a livello computazionale in termini di tempo, per cui è stato necessario modificare la soglia di supporto minimo in modo tale da ricavare un numero più ristretto di regole, rendendolo gestibile secondo tempistiche non eccessive. Il classificatore permette di ricavare due categorie di regole, ovvero le regole di primo livello e le regole di secondo livello. Per analizzare i risultati dei vari modelli, sono state prese in considerazione le regole della prima categoria, in modo da valutare un insieme di regole ad alta qualità, utilizzate nella fase di apprendimento.

5.2 Risultati ricavati

Dopo aver ottenuto le regole di associazione per ciascun caso di analisi, sono state selezionate le regole di primo livello ritenute più interessanti in base ad alcuni parametri statistici ricavabili dai modelli costruiti. Dopodichè, è stata applicata una traduzione delle regole da una forma compatta, generata dal classificatore, ad una forma standard, facilmente comprensibile e ne sono state valutate la bontà e la validità in base al contesto applicativo. Questo processo è stato ripetuto più volte, in modo da ottenere un numero discreto di caratterizzazioni significative. Nella Tabella 5.1, sono rappresentati alcuni risultati, distinti per categoria di feature e tipologia di polizze.

Tabella 5.1: *Selezione di alcuni risultati ottenuti*

Categorie	**Polizze**	**Regole di associazione**
Mixed-feature	Nuove/Semestrali	-Classe di merito = bassa, Anno di prima immatricolazione = mediamente recente, Metri in autostrada = basso, Ritardo ottenimento patente = nessuno, Accelerazioni brusche di giorno in città = medio, Età dell'assicurato = adulto, Eccessi di velocità di notte in città = basso, Età del veicolo = mediamente recente ⇒ *Numero di incidenti con colpa grave causati = 1* -Anno di prima immatricolazione

Continua nella prossima pagina

Continua dalla pagina precedente

Categorie	**Polizze**	**Regole di associazione**
		= non recente, Età dell'assicurato = adulto, Cavalli fiscali = basso ⇒ *Numero di incidenti senza colpa grave causati = 1*
		-Classe di merito = bassa, Alimentazione del veicolo = benzina, Accelerazioni brusche di giorno in città = medio, Accelerazioni brusche di notte in città = basso, Frenate brusche di giorno in città = elevato, Cavalli fiscali = basso ⇒ *Numero di incidenti causati = 1*
	Nuove/Annuali	Numero totale di metri = medio/alto, Sesso = femmina, Curvature ad alta velocità di giorno in città = medio/alto, Ritardo ottenimento patente = alto ⇒ *Numero di incidenti con colpa grave causati = 1*
	Vecchie/Semestrali	Accelerazioni brusche di giorno in città = nessuna, Accelerazioni brusche di notte in autostrada = nessuna, Cavalli fiscali = basso, Età del veicolo = non recente ⇒ *Numero di incidenti con colpa grave causati = 0*
	Vecchie/Annuali	Eccessi di velocità di giorno in autostrada = nessuno, Cavalli fiscali = basso, Alimentazione del veicolo = diesel, Sesso = femmina, Anno di prima immatricolazione = mediamente recente ⇒ *Numero di incidenti senza colpa grave causati = 0*
Caratteristiche delle polizze	Nuove/Semestrali	Classe di merito = bassa, Alimentazione del veicolo = diesel, Sesso = maschio, Anno di prima

Continua nella prossima pagina

Continua dalla pagina precedente

Categorie	**Polizze**	**Regole di associazione**
		immatricolazione = mediamente recente, Ritardo ottenimento patente = nessuno, Età dell'assicurato = adulto, Professione = impiegato, Provincia = Barletta ⇒ *Numero di incidenti con colpa grave causati = 1*
	Nuove/Annuali	-Classe di merito = bassa, Sesso = femmina, Età dell'assicurato = mezza età, Anno di prima immatricolazione = recente, Professione = altro, Ritardo ottenimento patente = alto, Provincia = Pesaro, Alimentazione del veicolo = gpl ⇒ *Numero di incidenti con colpa grave causati = 1*
		-Ritardo ottenimento patente = basso, Sesso = femmina, Provincia = Napoli, Anno di prima immatricolazione = recente, Età dell'assicurato = adulto, Alimentazione del veicolo = benzina, Classe di merito = alta, Professione = medico ⇒ *Costo degli incidenti causati = medio/alto*
	Vecchie/Annuali	-Classe di merito = bassa, Provincia = Torino, Anno di prima immatricolazione = mediamente recente ⇒ *Numero di incidenti con colpa grave causati = 0*
		-Sesso = femmina, Età dell'assicurato = adulto, Professione = altro, Alimentazione del veicolo = benzina, Ritardo ottenimento patente = alto, Classe di merito = media, Anno di prima immatricolazione = vecchia,

Continua nella prossima pagina

Continua dalla pagina precedente

Categorie	**Polizze**	**Regole di associazione**
		Provincia = Napoli ⇒ *Numero di incidenti con colpa grave causati = 1*
		-Alimentazione del veicolo = diesel, Sesso = maschio, Anno di prima immatricolazione = recente, Classe di merito = media, Professione = agente di commercio, Provincia = Frosinone ⇒ *Numero di incidenti con colpa grave causati = 1*
	Vecchie/Semestrali	Anno di prima immatricolazione = non recente, Alimentazione del veicolo = diesel, Sesso = maschio, Professione = operaio ⇒ *Numero di incidenti senza colpa grave causati = 0*
Stile di guida	Nuove/Semestrali	-Decelerazioni brusche di giorno in città = elevato, Curvature ad alta velocità di giorno in città = elevato, Curvature ad alta velocità di giorno in autostrada = basso, Accelerazioni brusche di giorno in città = basso, Frenate brusche di giorno in città = elevato, Frenate brusche in autostrada di giorno = basso ⇒ *Numero di incidenti con colpa grave causati = 1*
		-Frenate brusche di giorno in autostrada = nessuno, Decelerazioni brusche di giorno in città = mediamente basso, Curvature ad alta velocità di giorno in città = mediamente basso ⇒ *Numero di incidenti senza colpa grave causati = 0*

Continua nella prossima pagina

Continua dalla pagina precedente

Categorie	**Polizze**	**Regole di associazione**
	Vecchie/Annuali	Eccessi di velocità di giorno in città = medio, Curvature ad alta velocità di giorno in autostrada = basso ⇒ *Numero di incidenti con colpa grave causati = 0*
	Vecchie/Semestrali	Curvature ad alta velocità di giorno in autostrada = basso, Eccessi di velocità di giorno in città = medio, Decelerazioni brusche di giorno in città = elevato, Curvature ad alta velocità di giorno in città = elevato, Frenate brusche di giorno in autostrada = basso, Frenate brusche di giorno in città = medio, Accelerazioni brusche di giorno in città = basso ⇒ *Numero di incidenti con colpa grave causati = 1*
	Nuove/Annuali	-Accelerazioni brusche di giorno in autostrada = nessuno, Decelerazioni brusche in città di giorno = mediamente basso, Frenate brusche in città di giorno= mediamente basso, Curvature ad alta velocità in città di giorno = mediamente basso ⇒ *Numero di incidenti senza colpa grave causati = 0* -Eccessi di velocità in autostrada = nessuno, Decelerazioni brusche in città di giorno = mediamente alto, Curvature ad alta velocità in autostrada di giorno = basso, Frenate brusche in autostrada di giorno = basso, Curvature ad alta velocità in città di giorno = mediamente alto, Frenate brusche

Continua nella prossima pagina

Continua dalla pagina precedente

Categorie	**Polizze**	**Regole di associazione**
		in città di giorno = elevato, Eccessi di velocità in città di giorno = basso ⇒ *Costo degli incidenti causati = alto*
Percorrenze	Nuove/Semestrali	Numero totale di metri = medio/alto, Numero metri percorsi in autostrada = medio, Numero metri percorsi in periferia = medio/alto, Numero metri percorsi di notte = mediamente basso, Numero metri percorsi di giorno = elevato, Numero metri percorsi in città = medio/alto ⇒ *Numero di incidenti con colpa grave causati = 1*
	Vecchie/Annuali	-Numero metri percorsi in autostrada = medio, Numero totale di metri = medio/alto, Numero metri percorsi di notte = basso/medio, Numero metri percorsi in città = basso, Numero metri percorsi di giorno = medio/alto, Numero metri percorsi in periferia = alto ⇒ *Numero di incidenti con colpa grave causati = 1* -Numero metri percorsi in città = medio, Numero metri percorsi di giorno = medio, Numero totale di metri = medio/alto, Numero metri percorsi in periferia = basso/medio, Numero metri percorsi di notte = basso/medio, Numero metri percorsi in autostrada = basso ⇒ *Costo degli incidenti causati = alto*
	Vecchie/Semestrali	Numero metri percorsi in autostrada = basso, Numero metri

Continua nella prossima pagina

Continua dalla pagina precedente

Categorie	Polizze	Regole di associazione
		percorsi in città = basso/medio, Numero metri percorsi di giorno = basso/medio, Numero metri percorsi di notte = basso ⇒ *Numero di incidenti senza colpa grave causati = 0*

Si conclude dalla pagina precedente

Le regole di associazione ricavate permettono di esplorare le correlazioni, non note a priori, tra feature della stessa categoria o di categorie differenti, delineando la combinazione di determinati intervalli di valori relativamente ad un vasto insieme di dati. La tipologia "Mixed-feature" è da ritenersi la più interessante, poichè consente di mettere in relazione più categorie differenti in modo tale da specificare più dettagliatamente i profili delle polizze. Prendendo come riferimento un sottoinsieme delle correlazioni più significative ricavabili dall'insieme di regole estratte e accuratamente selezionate, dalle quali è possibile definire certi indicatori chiave (KPI) con cui misurare l'impatto delle correlazioni per scopi futuri, è possibile trarre le seguenti conclusioni:

- le polizze con classe di merito bassa associate a soggetti con una delle seguenti caratteristiche:
  - poca esperienza di guida;
  - patente ottenuta tardivamente;
  - residenza in zone con elevata incidenza dei sinistri;

  rappresentano profili ad alto rischio. Monitorare dunque la combinazione dei fattori sopra indicati, anzichè considerarli separatamente, consente alla compagnia assicurativa di gestire in modo efficace il rischio e di conseguenza, migliorare i processi di pricing;
- le polizze con classe di merito bassa associate a soggetti con stili di guida poco regolari, relativamente, ad esempio, ad un numero elevato di cambi repentini di direzione, rappresentano profili ad alto rischio. Per cui, monitorare la classe di merito in combinazione con gli stili di guida consente una più accurata gestione del rischio;

- le polizze con classe di merito bassa associate a veicoli di potenza limitata non costituiscono tipicamente profili ad alto rischio; al contrario, le polizze associate a veicoli di età avanzata e intestate a soggetti con stili di guida poco regolari, relativamente, ad esempio, ad un numero eccessivo di curvature ad alta velocità di giorno in aree urbane, sono potenzialmente ad alto rischio. Considerare quindi l'età e la potenza del veicolo in combinazione con la classe di merito e gli stili di guida consente di raffinare ulteriormente la valutazione sul livello di rischio della polizza.

Capitolo 6

Conclusioni e sviluppi futuri

Nel corso degli anni, le compagnie assicurative hanno mostrato forte interesse verso il concetto di scatola nera, la quale è divenuta un elemento essenziale al momento di una nuova stipulazione di contratto. Per incentivarne l'utilizzo e delineare così una forte riduzione delle frodi assicurative, le compagnie forniscono delle agevolazioni qualora si decida di adottarne una. I principali vantaggi che derivano dall'installazione della black box a bordo degli autoveicoli consistono nella possibilità di delineare in maniera dettagliata la dinamica di un eventuale sinistro, di rintracciare un veicolo nel caso in cui esso venisse sottratto e di personalizzare le polizze assicurative in base alle caratteristiche del conducente. Relativamente all'ultimo punto, le compagnie si sono rivelate sempre più intenzionate a sfruttare la grande mole di dati reperita dalle scatole nere, in modo tale da ricavarne importanti conoscenze ai fini di realizzare servizi a valore aggiunto per i clienti.

In questo lavoro di tesi, sono stati analizzati i dati relativi alle polizze assicurative, descritti nel capitolo 2, con lo scopo di caratterizzare i profili dei clienti in base al loro livello di rischio di causare sinistri e di fornire correlazioni nascoste tra i dati in modo tale da estrarre indicatori per realizzare servizi ad hoc. La caratterizzazione è stata realizzata attraverso l'applicazione del classificatore basato su regole di associazione L^3 (*Live and Let Live*). Quest'ultimo è stato applicato per ricavare vari modelli associativi per poi successivamente esplorarne il contenuto ed estrapolarne le correlazioni tra le varie feature, mediante la rappresentazione di regole di associazione. Nel capitolo 4 è stato discusso l'intero processo di analisi dei dati, prestando una significativa attenzione alla fase di preprocessing. In quest'ultima fase è stata realizzata la pulizia dei dati e la trasformazione del loro formato per renderli idonei all'applicazione delle tecniche di data mining e inoltre, sono stati ottenuti i dataset preparati per l'analisi successiva.

I dataset realizzati sono stati differenziati relativamente ai seguenti fat-

tori:

- tipologia di polizze;
- categoria di feature;
- livelli di rischio.

La caratterizzazione dei livelli di rischio è stata effettuata sia per ciascuna categoria, al fine di correlare i valori delle feature di una categoria specifica con il livello di rischio considerato, sia combinando le feature di categorie differenti. Quindi, dopo aver applicato il classificatore L^3, sono stati ricavati i modelli associativi distinti in base ai tre punti sopra citati. Le correlazioni estratte e selezionate hanno permesso di delineare profili dettagliati di polizze in base ai livelli di rischio definiti, in modo tale da identificare associazioni non note, utili come criterio di decisione per valutarne l'impatto e la propensione ai sinistri. Lo scopo del lavoro di tesi si riassume nella volontà di supportare le compagnie assicurative nel trattamento e nella gestione della significativa mole di dati reperita dalle scatole nere. Nel dettaglio, le correlazioni più significative estratte dai modelli associativi sono state schematizzate secondo Key Performance Indicators, in modo tale che le compagnie possano sfruttarle sia per la profilazione dei clienti, differenziando l'erogazione di servizi per gruppi, sia per valutare l'impatto futuro di alcune feature caratterizzanti, relativamente ai livelli di rischio.

Bibliografia

[1] R. Agrawal, T. Imieliński, and A. Swami. Mining association rules between sets of items in large databases. In *Acm sigmod record*, volume 22, pages 207–216. ACM, 1993.

[2] P. Atzeni, S. Ceri, S. Paraboschi, and R. Torlone. *Basi di dati: modelli e linguaggi di interrogazione (seconda edizione)*. McGraw-Hill, 2006.

[3] E. Baralis, S. Chiusano, and P. Garza. A lazy approach to associative classification. *IEEE Transactions on Knowledge and Data Engineering*, 20(2):156–171, 2008.

[4] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[5] S. Brin, R. Motwani, J. D. Ullman, and S. Tsur. Dynamic itemset counting and implication rules for market basket data. *Acm Sigmod Record*, 26(2):255–264, 1997.

[6] R. Elmasri and S. Navathe. *Fundamentals of database systems*. Addison-Wesley Publishing Company, 2010.

[7] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37, 1996.

[8] J. Han, J. Pei, and M. Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

[9] D. J. Hand, P. Smyth, and H. Mannila. *Principles of Data Mining*. MIT Press, Cambridge, MA, USA, 2001.

[10] S. B. Kotsiantis, I. D. Zaharakis, and P. E. Pintelas. Machine learning: a review of classification and combining techniques. *Artificial Intelligence Review*, 26(3):159–190, 2006.

[11] H. Liu, F. Hussain, C. L. Tan, and M. Dash. Discretization: An enabling technique. *Data mining and knowledge discovery*, 6(4):393–423, 2002.

[12] B. L. W. H. Y. Ma and B. Liu. Integrating classification and association rule mining. In *Proceedings of the fourth international conference on knowledge discovery and data mining*, 1998.

[13] E. Rahm and H. H. Do. Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.*, 23(4):3–13, 2000.

[14] P.-N. Tan, M. Steinbach, and V. Kumar. Association analysis: basic concepts and algorithms. *Introduction to Data mining*, pages 327–414, 2005.

[15] F. Thabtah. A review of associative classification mining. *The Knowledge Engineering Review*, 22(1):37–65, 2007.

[16] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, 2nd edition, 2005.