POLITECNICO DI TORINO

Corso di Laurea in Ingegneria Informatica

Tesi di Laurea Magistrale

Predizione del livello di rischio di polizze assicurative mediante tecniche di classificazione

Relatore
prof. Luca Cagliero

Correlatore:
prof.ssa Elena Baralis

Laureando
Gianluca Pasciucco

Anno accademico 2018-2019

Il successo è l'abilità di passare da un fallimento all'altro senza perdere l'entusiasmo.

W. Churchill

Ringraziamenti

Sebbene il conseguimento di una laurea magistrale sembri un traguardo puramente individuale, per me non è stato così. Senza il supporto di alcune persone, tutto ciò non sarebbe stato possibile e ci tengo a ringraziarle tutte.

Desidero ringraziare il prof. Cagliero per aver guidato questo progetto con estrema attenzione, passione e professionalità; si è sempre dimostrato pronto ad aiutarmi, a colmare le mie perplessità e a indirizzarmi verso la strada giusta, a ogni ora del giorno e della notte. Ringrazio la prof. ssa Baralis che tramite una sapiente capacità di insegnamento mi ha fatto appassionare ai temi di data science e supervisionando l'intero progetto ha saputo indicare le direzioni più interessanti per completare le nostre ricerche.

Ringrazio immensamente i miei genitori.
La motivazione maggiore che mi ha guidato in questi anni è stata quella di rendervi orgogliosi di me e di non sprecare nemmeno una goccia dei vostri sacrifici, spero che questa laurea possa essere una piccola ricompensa. Mi avete insegnato cos'è l'amore e vi ringrazio per avermi regalato la serenità dello stare in famiglia, facilitando di molto il mio compito di figlio e di studente. Vi ringrazio per essermi stati sempre vicini, sempre presenti, ho apprezzato ogni singola attenzione che mi è stata data e ho cercato di fare tesoro di ogni critica ricevuta. Ringrazio mia sorella Susanna che non si ferma mai; riesci a trasmettermi sempre ottimismo e ci sei ogni volta che ne ho bisogno, spero vivamente che dopo questa laurea magistrale tu la smetta di dire che sono un fannullone. Ringrazio i miei nonni che mi hanno dimostrato sempre un affetto incondizionato e una generosità immensa, vi sono grato per avermi insegnato cosa significa l'umiltà e il sacrificio, gran parte di questo percorso è merito vostro.

Antonio, Carlo e Francesco: definirvi solamente amici sarebbe ingiusto, potervi definire fratelli sarebbe un onore.
Ringrazio Antonio. In pochi anni le gioie e i dolori che hanno fatto da sfondo alla nostra amicizia hanno creato un rapporto indissolubile e speciale. Spronarci a vicenda durante le sessioni d'esame e riuscire a superare le difficoltà insieme, ha raddoppiato la soddisfazione. Sapere di poter contare sui tuoi consigli, anche quando eravamo gli unici due a non gioire, mi ha fatto sentire una persona fortunata.
Ringrazio Carlo. Ripercorrendo i ricordi dall'infanzia, non riesco a ricordare un solo momento felice in cui tu non sia presente. La tua ambizione, la tua dedizione e la tua voglia di non accontentarsi sono un modello che porterò sempre con me nell'ambito lavorativo e personale, certo del successo che potranno generare. Ringrazio Francesco. Passare del tempo con te, senza ridere, è un'impresa impossibile. La tua energia, la tua passione, la tua sincerità incondizionata e la tua capacità di

gestire le grandi responsabilità che, talvolta, la vita può assegnare, provocano in me una grande ammirazione e ti rendono una preziosa fonte di ispirazione.

Grazie ai miei amici e le mie amiche di Stigliano, conservo gelosamente i vostri consigli, le serate passate insieme e le nostre risate, le ritengo un tesoro inestimabile.

Grazie ai miei compagni di avventura, Luca e Simone, che scherzosamente definisco grandi artisti delle lamentele gratuite, senza i quali le lezioni non sarebbero state le stesse e io sarei stato sicuramente uno studente peggiore. Avere amici come voi ha reso il Politecnico più facile.

Grazie, infine, a chi ha condiviso con me un giorno, un mese, un anno, sette anni e mezzo, ognuno di voi mi ha insegnato qualcosa e mi ha reso una persona migliore.

Indice

Elenco delle tabelle

Elenco delle figure

Capitolo 1

Introduzione

Viviamo in un mondo dove le tecnologie si evolvono rapidamente e coinvolgono, con vere e proprie rivoluzioni, interi settori economici o sociali. Fino ai primi anni 2000 la tecnologia GPS, che ci consente di conoscere la nostra posizione esatta in ogni angolo del pianeta, era un'esclusiva dedicata a pochi settori, soprattutto militari, con costi elevatissimi per un utente medio. Come spesso accade, le tecnologie sviluppate in ambito militare e sui prototipi usati nelle competizioni automobilistiche, con il passare degli anni e con l'abbassamento dei costi, vengo trasferite alle produzioni di serie destinate all'uso civile.

Il settore assicurativo è stato uno dei primi a intuire le potenzialità di un tracciamento GPS in grado di registrare gli spostamenti di un determinato automobilista, di aiutarlo in caso di incidente segnalando ai soccorsi la posizione esatta dell'accaduto, o in caso di furto segnalando alle forze dell'ordine la posizione dell'auto in tempo reale. Tutto questo è stato possibile grazie all'introduzione della cosiddetta "scatola nera" (o black-box), un piccolo apparecchio da installare sull'auto di un assicurato, in grado di registrare tutti gli spostamenti di un'automobile, nonché il comportamento alla guida del conducente. In Italia, a differenza di altri paesi in cui è obbligatoria da decenni, l'installazione della scatola nera è ancora facoltativa, tuttavia sono stati imposti per legge degli sconti da applicare sulle polizze RC Auto che scelgono di utilizzare la black-box, motivo per cui negli ultimi anni si è assistito ad un'impennata nel numero dei guidatori che scelgono questo sistema. Grazie alla diffusione sempre più ampia, il volume di dati a disposizione delle compagnie assicurative è cresciuto sempre di più, diventando un vero e proprio patrimonio da proteggere e da studiare.

Questo lavoro di tesi ha avuto come obiettivo primario quello di cercare una risposta ad una domanda ambiziosa: è possibile prevedere un incidente? Attraverso l'uso delle più moderne tecniche di machine learning e data mining si è cercato di costruire un modello che potesse essere in grado di predire il livello di rischio di una determinata polizza. Questo tipo di tecniche è alla base di molti dei recenti successi nel campo della medicina, della gestione delle emergenze, della tutela dell'ambiente e permette di scovare delle connessioni (potenzialmente) nascoste tra i dati in nostro possesso e degli specifici eventi.

Il dataset analizzato nell'ambito di questo lavoro di tesi memorizza quattro categorie di informazioni aggregate: i percorsi effettuati dai veicoli assicurati, gli stili

di guida adottati dai guidatori dei veicoli, le caratteristiche delle polizze stipulate e dei guidatori a cui è associata ciascuna polizza, i sinistri commessi. Per ogni categoria sono disponibili dati aggregati che riassumono caratteristiche salienti relative ad una specifica categoria (ad es., per i percorsi effettuati viene registrato il numero di km percorsi in tratti urbani). Le percorrenze tengono traccia di tutti gli spostamenti del veicolo e sono il conteggio dei metri percorsi e della durata (in secondi) di tali distanze, in una determinata settimana. I dati vengono poi differenziati per tipo di strada percorsa (autostrada, urbana, extraurbana, ..) per giorno della settimana e per fascia oraria. Lo stile di guida invece viene misurato attraverso cinque tipi di eventi che esprimono dei comportamenti scorretti alla guida: eccesso di velocità, accelerazioni e frenate brusche, cambi di direzione repentini e curve percorse a velocità elevata. Tramite la scatola nera è possibile rilevare questi 5 tipi di eventi che poi vengono aggiunti al report settimanale e differenziati, come nel caso delle percorrenze, per giorno della settimana, fascia oraria e tipo di strada. Ottengono di fatto una fotografia comportamentale dell'automobilista. Grazie a questo tipo di eventi si può già iniziare a notare la differenza tra un guidatore prudente e uno meno prudente. Vengono infine integrati i dati anagrafici dell'assicurato (sesso, residenza, età, professione, ecc.), le caratteristiche tecniche del veicolo (alimentazione, potenza, ecc.) e i dati di un eventuale sinistro (data, costo, veicoli coinvolti, ecc.).

L'analisi del dataset che memorizza i dati storici relativi a polizze, stili di guida, percorrenze e sinistri è finalizzato ad identificare pattern ricorrenti in grado di discriminare le caratteristiche delle polizze a rischio sinistri da quelle non a rischio. Per analizzare un dataset che integra informazioni eterogenee come quelle sopra descritte si è resa necessaria un'importante fase di pre-processing, in cui grazie ai consigli degli esperti di dominio, sono stati rimossi gli attributi potenzialmente meno interessanti per l'analisi, sono stati rimossi i dati incongruenti e infine sono stati aggregati i dati (inizialmente su base giornaliera) in tre diverse aggregazioni: settimanale, mensile e annuale. Lo scopo di questa differenziazione è quello di cercare di far emergere un pattern comportamentale via via più definito, che rappresenti quasi una media del comportamento di un guidatore, che può essere quindi più evidente a livello annuale o mensile, meno evidente a livello giornaliero o settimanale. La fase di pulizia e di aggregazione del dataset ha richiesto anche lo sviluppo di un'applicazione specifica per consentire di utilizzare i vari tool di analisi anche su elaboratori di piccole dimensioni come quelli di uso comune.

Il dataset preparato è stato analizzato mediante algoritmi di machine learning supervisionati per costruire modelli predittivi del livello di rischio; in particolare, sono stati costruiti modelli di classificazione a partire da un insieme di dati storici relativi a polizze aventi livello di rischio noto. Tali modelli sono applicabili per predire il livello di rischio di una nuova polizza. I livelli di rischio considerati nelle analisi svolte sono il rischio di commettere almeno un incidente nel periodo futuro considerato (ad es. il prossimo anno) e il costo dei sinistri causati. Le tecniche di classificazione considerate per l'analisi sono le Support Vector Machines, i Decision Tree, i distance-based classifiers (k Nearest Neighbors) e due algoritmi Bayesiani (Naive Bayes e WAODE). Le analisi sono state effettuate sia su scala settimanale sia su scala annuale. Il modello creato deve essere validato, cioè deve esserne provata la qualità che può essere descritta tramite quattro parametri principali: accuratezza,

richiamo, precisione, f-score. Dagli esperimenti effettuati si è notato come i modelli di classificazione non siano stati in grado di predire con sufficiente precisione le polizze ad altro rischio incidente su base settimanale, mentre sono stati ottenuti risultati promettenti sulle analisi effettuate su base annuale.

Questa tesi è composta da 8 capitoli. Il *capitolo 2*, che segue questa introduzione, descrive il contesto dell'analisi, le caratteristiche di una scatola nera e i dati a disposizione. Il *capitolo 3* spiega la metodologia dell'analisi e i riassume i passi seguiti dall'inizio alla fine dell'esperimento. Il *capitolo 4* tratta il machine learning e il data mining, spiegando le tecniche utilizzate all'interno di questo lavoro di tesi e le motivazioni relative alle scelte operate. Il *capitolo 5* descrive lo stato dell'arte della ricerca in ambito di predizione e analisi di polizze (assicurative e non). Il *capitolo 6* illustra le funzionalità del software sviluppato che si occupa della fase di pre-processing dei dati. Il *capitolo 7* descrive come sono state combinate le varie tecniche di classificazione e validazione, l'applicazione dei metodi scelti, viene analizzata la differenza tra i vari algoritmi e l'influenza dei diversi parametri sull'esito degli esperimenti. Nel *capitolo 8* vengono riepilogati i risultati e vengono suggeriti alcuni possibili sviluppi futuri.

Capitolo 2

Contesto dell'analisi

Il settore assicurativo è stato uno dei primi a intuire le potenzialità di un tracciamento GPS in grado di registrare gli spostamenti di un determinato automobilista e di aiutarlo in caso di incidente segnalando ai soccorsi la posizione esatta dell'accaduto, o in caso di furto segnalando alle forze dell'ordine la posizione dell'auto in tempo reale. Tutto questo è stato possibile grazie all'introduzione della cosiddetta "scatola nera" (o black-box), un piccolo apparecchio da installare sull'auto di un assicurato, in grado di registrare tutti gli spostamenti di un'automobile, nonché il comportamento alla guida del conducente. Può essere utilizzato dalla compagnia assicuratrice anche per profilare ulteriormente i clienti e offrire quindi delle polizze ulteriormente personalizzate e fatte su misura per il tipo di utilizzo.

2.1 Cos'è una scatola nera

Una scatola nera è un dispositivo satellitare che viene installato sul veicolo assicurato. Grazie ad un modulo GPS, un accelerometro e un microprocessore è in grado di registrare 5 tipi di eventi legati allo stile di guida (accelerazione brusca, frenata brusca, eccesso di velocità, percorrenza curva a velocità elevata, cambiamento repentino di direzione) e di misurare le percorrenze chilometriche differenziate per giorno della settimana, fascia oraria (diurna o notturna), tipo di strada percorsa (autostrada, urbana, suburbana, altro).

La compagnia assicurativa beneficia di numerosi vantaggi derivanti dall'installazione della scatola nera:

- Una gestione più efficiente delle pratiche assicurative genera dei risparmi sui costi dei sinistri a vantaggio delle compagnie di assicurazione e, conseguentemente, dei loro assicurati che possono usufruire di migliori condizioni. La compagnie assicurativa ne incentiva l'installazione tramite uno sconto fisso che può arrivare fino al 10% sul premio assicurativo.
- Personalizzazione dell'offerta, grazie all'analisi del comportamento di guida dell'assicurato (analisi del chilometraggio, dello stile di guida, della frequenza di utilizzo del veicolo, della velocità, ecc.). È possibile offrire una riduzione ulteriore del premio assicurativo per gli automobilisti più prudenti.

- Semplificazione del processo di gestione di un sinistro e ottimizzazione del rilevamento e accertamento delle frodi.
- Fidelizzazione dei migliori clienti con polizze «su misura» arricchite da servizi a valore aggiunto (assistenza stradale più efficiente, rilevazione automatica di un incidente, localizzazione del veicolo in caso di furto, comunicazione diretta tra Centrale Operativa e veicolo tramite dispositivo di bordo o cellulare).

2.1.1 Caratteristiche tecniche

Una scatola nera è un piccolo dispositivo sviluppato appositamente per il mercato assicurativo. L'installazione può essere effettuata direttamente dal Cliente, seguendo le istruzioni fornite. I dati analitici rilevati diventano pertanto uno strumento fondamentale per l'elaborazione di statistiche di utilizzo del veicolo e permettono anche la ricostruzione telematica della dinamica dell'incidente. Contiene al suo interno diversi componenti tra cui cinque componenti chiave: modulo GPS, modulo GSM, connessione CAN-bus, piattaforma accelerometrica e il bus wireless.

Figura 2.1: Esempio di scatola nera

2.2 Modalità di raccolta dei dati

Il dataset oggetto degli esperimenti ha lo scopo di fornire i dati relativi alle percorrenze effettuate dal veicolo assicurato in un determinato giorno e tiene traccia, per ogni giorno presente nei record, degli eventi legati allo stile di guida (se presenti). I file vengono prodotti su base settimanale con dettaglio giornaliero, quindi per ogni giorno per i quali sono stati raccolti i dati delle percorrenze sarà prodotto un file che le contiene, con relativi dati che descrivono il comportamento di guida. Per "viaggio" si intende la distanza percorsa (e il tempo impiegato) tra l'accensione del motore e il successivo spegnimento. Un viaggio a cavallo tra due giorni, verrà diviso in due singoli viaggi ognuno afferente al giorno di riferimento.

2.3 Dataset delle percorrenze

Contiene i dettagli relativi alle percorrenze del veicolo. Tiene traccia sia dello spazio percorso che del tempo impiegato.

N.	Field	Type	Comment/Notes
1	Tipo	String	Sempre 2
2	Contract	String	Numero polizza csu un blank in testa.
3	N. voucher	String	Fisso 10 csu blank in testa
4	Settimana	String	Settimana fiscale ISO 8601 standard format (es. 2017-W06 è la settimana che va dal 06/02 al 12/02).
5	Start Date	String	YYYY-MM-DD Riferita alla data del primo viaggio effettuato nella settimana di riferimento
6	Start Time	String	HH:MM:SS. Riferita all'ora del primo viaggio effettuato nella settimana di riferimento
7	End Date	String	YYYY-MM-DD Riferita alla data dell'ultimo viaggio effettuato nella settimana di riferimento. Il numero di giorni analizzato (end date ? start date) può essere minore di 7 in quanto corrispsude al numero di giorni in cui il Cliente ha utilizzato il mezzo.
8	End Time	String	HH:MM:SS. Riferita all'ora dell'ultimo viaggio effettuato nella settimana di riferimento

9	Number of trips	Number	Numero di viaggi compiuti nella settimana di riferimento compresi quelli spezzati a cavallo del primo e dell'ultimo giorno Dalle 00.00.00 del Lunedì al 23.59.59 della Domenica. NOTE: i viaggi effettuati a cavallo di due giorni devsuo csusiderarsi come 2 viaggi distinti
10	Metri percorsi	Number	Numero di metri percorsi nei viaggi effettuati nella settimana di riferimento
11	Metri percorsi su autostrade	Number	Nella settimana di riferimento, numero di metri percorsi su autostrade. otal Metri percorsi in the period in subject su motorways
12	Metri percorsi su strade urbane	Number	Nella settimana di riferimento, numero di metri percorsi su strade urbane.
13	Metri percorsi su altro tipo	Number	Nella settimana di riferimento, numero di metri percorsi su altre strade (non autostrade e strade urbane)
14	Metres travelled su tipo sconosciuto	Number	Nella settimana di riferimento, numero di metri percorsi su strade scsuosciute (nsu autostrade e strade urbane)
15	Tempo di viaggio	Number	Secondi percorsi nei viaggi compiuti nella settimana di riferimento compresi quelli spezzati a cavallo del primo e dell'ultimo giorno

16	Tempo di viaggio su autostrade	Number	Nella settimana di riferimento, Secondi percorsi nei viaggi su autostrade
17	Tempo di viaggio su strade urbane	Number	Nella settimana di riferimento, Secondi percorsi nei viaggi su strade urbane
18	Tempo di viaggio su altro tipo	Number	Nella settimana di riferimento, Secondi percorsi nei viaggi su altre strade (non autostrade e strade urbane)
19	Tempo di viaggio su tipo sconosciuto	Number	Nella settimana di riferimento, Secondi percorsi nei viaggi su strade scsuosciute (nsu autostrade e strade urbane)
20	Metri percorsi Lun	Number	Totale metri percorsi nella giornata di Lunedì
21	Metri percorsi Mar	Number	Totale metri percorsi nella giornata di Martedì
22	Metri percorsi Mer	Number	Totale metri percorsi nella giornata di Mercoledì
23	Metri percorsi Gio	Number	Totale metri percorsi nella giornata di Giovedì
24	Metri percorsi Ven	Number	Totale metri percorsi nella giornata di Venerdì
25	Metri percorsi Sab	Number	Totale metri percorsi nella giornata di Sabato
26	Metri percorsi Dom	Number	Totale metri percorsi nella giornata di Domenica
27	Tempo di viaggio Lun	Number	Totale Secondi percorsi nella giornata di Lunedì
28	Tempo di viaggio Mar	Number	Totale Secondi percorsi nella giornata di Martedì

29	Tempo di viaggio Mer	Number	Totale Secondi percorsi nella giornata di Mercoledì
30	Tempo di viaggio Gio	Number	Totale Secondi percorsi nella giornata di Giovedì
31	Tempo di viaggio Ven	Number	Totale Secondi percorsi nella giornata di Venerdì
32	Tempo di viaggio Sab	Number	Totale Secondi percorsi nella giornata di Sabato
33	Tempo di viaggio Dom	Number	Totale Secondi percorsi nella giornata di Domenica
34	Metri percorsi di giorno	Number	Totale metri percorsi nella fascia oraria diurna (06 - 24)
35	Metri percorsi di notte	Number	Totale metri percorsi nella fascia oraria notturna (24 - 06)
36	Tempo di viaggio di giorno	Number	Totale Tempo di viaggio nella fascia oraria diurna (06 - 24)
37	Tempo di viaggio di notte	Number	Totale Tempo di viaggio nella fascia oraria notturna (24 - 06)

Tabella 2.1: Dataset delle percorrenze

2.4 Dataset dello stile di guida

Riporta tutti i dettagli allo stile di guida del guidatore. La scatola nera è in grado di rilevare 5 tipi di eventi:

- Eventi di eccesso di velocità: L'eccesso di velocità è considerato rispetto ai limiti di velocità[1] nominali su tre classi di strada (urbana, autostradale, altro) e considerando la velocità media su un tratto di strada di circa 2 km di lunghezza.
- Eventi di Accelerazione: considerato rispetto al superamento di una soglia di accelerazione longitudinale al veicolo.

[1] a cui viene aggiunta una tolleranza del 10%

- Eventi di Decelerazione: considerato rispetto al superamento di una soglia di decelerazione longitudinale al veicolo.
- Eventi di Curvatura: considerato rispetto al superamento di una soglia di accelerazione trasversale al veicolo per un periodo di tempo prolungato (curva percorsa ad alta velocità).
- Cambio repentino di direzione: considerato rispetto al superamento improvviso di una soglia di accelerazione trasversale al veicolo

N.	Campo	Tipo	Commenti/Note
1	Tipo	Stringa	Il valore è sempre 4 (= detail trip record)
2	Codice Evento	Numero	Codice del tipo di evento classificato nel record: 11 = Eventi di eccesso di velocità 21 = Eventi di Accelerazione 23 = Eventi di Decelerazione 25 = Eventi di Curvatura 27 = Cambio repentino di direzione
3	Tipo di strada	Stringa	U = urban; M = motorway; O = other (extra- urban, non-motorway); X = unknown
4	Fascia Oraria	Stringa	D = Diurna 6 - 24 (orario CET/CEST) N = Notturna 24 - 6 (orario CET/CEST)
5	Giorno della settimana	Stringa	LUN, MAR, MER, GIO, VEN, SAB, DOM
6	Numero di eventi	Numero	Numero di eventi riscontrati per il tipo di evento classificato nella settimana di riferimento.

Tabella 2.2: Dataset dello stile di guida

Il numero di eventi è da considerarsi quello effettuato nella settimana di riferimento. Per ogni tipo evento, tipo di strada, fascia oraria e giorno della settimana, deve essere scritto un record separato.

2.5 Dataset dei dettagli di polizza

Al dataset delle percorrenze e dello stile di guida vengono aggiunti i dettagli relativi ad una determinata polizza. Sono principalmente informazioni anagrafiche del

contraente della polizza, informazioni tecniche riguardo al veicolo e i dettagli di un eventuale sinistro.

N.	ATTRIBUTO	Nota /commento
1	CR	Codice del ramo di business
2	NPLZA	Numero della polizza
3	DINZIO	Data di inizio copertura della polizza
4	DFINE	Data di fine copertura della polizza
5	DSCAD	Data di scadenza della polizza
6	CV_FISC	Numero di cavalli fiscali del veicolo
7	KW	Potenza del veicolo in kW
8	COD_VEICOLO	Codice del tipo di veicolo
9	ID_BOX	Id della scatola nera
10	NPRS	Numero di persone trasportabili
12	SEX	Sesso (M/F)
13	PROV	Provincia di residenza dell'intestatario della polizza
14	AREA	Area geografica di provenienza
15	REGIONE	Regione di residenza dell'intestatario della polizza
16	COD_ISTAT	Codice ISTAT del comune di residenza
17	AFF_NUOVO	Si o No a seconda che sia una nuova polizza oppure una polizza rinnovata
18	RINNOVO	Polizza annuale / semestrale
19	ANZ_ASS	Classe non agevolata, polizza principale nella compagnia, polizza principale in altra compagnia
20	CLASSE_CU	Classe della polizza (da 1 a 16) - 99 viene usato per le polizze aziendali
21	ANZ_CLIENTE	Numero di rinnovi della polizza
22	MASSIMALE	massimale della polizza
23	AGEV	Classe agevolata o no
24	ETA_PAT	A che età è stata conseguita la patente
25	MESE_SCAD	Mese di scadenza della polizza
26	LAST_FIVE	Nell'attestato di rischio è indicato se negli ultimi 5 anni è stato assicurato e per quanti anni
27	INC_IN_ATT	Numero di sinistri presente nell'attestato di rischio
28	DIST_LAST	Quanti anni fa si è verificato l'ultimo sinistro
29	TIPO_DANNO	Tipo di danno
30	TOT_INC	Numero di sinistri totali nel periodo di riferimento della polizza

31	COSTO_TOT	Costo totale del sinistro
32	PAGATO_NO_REC	L'assicurazione si rivale sull'assicurato che non paga
33	RECUPERATO	L'assicurazione si rivale sull'assicurato che paga
34	DATA_INCIDENTE	Data del sinistro
35	YEARS_OWN	Anni di possesso del veicolo
36	COSTO	Costo totale del sinistro
37	PAG	Costo del sinistro già pagato
38	RISERVATO	Riservato
39	CAP	CAP di residenza del proprietario del veicolo
40	RINUNCIA	Rinuncia di rivalsa
41	COMP	Compagnia di assicurazione
42	N_CAUS	Numero totale dei sinistri, con colpa, causati
43	ALIM	Alimentazione del veicolo assicurato
44	BOX	Black box si/no
45	DESC_BOX	Descrizione del tipo di black box
46	COST	Costo totale dei sinistri, con colpa, causati
47	LICENSE_AGE	Età della patente
48	PROF	Professione dell'assicurato
49	TIPO_V	Tipo di veicolo
50	USAGE	Tipo di uso del veicolo
51	PROP_CONTR	Il proprietario è il contraente (1=Si, 0=No)
52	DIST_PAT	L'età in cui si è conseguita la patente - 18 (
53	AGE_C	Età del contraente
54	FIRST_IMM	Anno di prima immatricolazione del veicolo
55	FIRST_IMM_DATE	Data di prima immatricolazione del veicolo
56	CAR_AGE	Età dell'auto
57	ULT_VOLT	Data ultima voltura del veicolo
58	YEARS_OWN	Anni di possesso del veicolo
59	COSTO1	Non card claim cost - Costo degli incidenti gravi (che causano disabilità >9%)
60	N1	Non card claim number - Numero degli incidenti gravi (che causano disabilità >9%)
61	COSTO2	Debtor-card claim cost - Costo degli incidenti non gravi (che causano disabilità <9%)
62	N2	Debtor-card claim number - Numero degli incidenti non gravi (che causano disabilità <9%)

63	COSTO3	Costo degli incidenti "subiti" (senza colpa) dall'assicurato
64	FORF1	Forfait incassati dalle altre assicurazioni coinvolte per assicurati che subiscono il sinistro
65	N3	Numero degli incidenti "subiti" (senza colpa) dall'assicurato

Tabella 2.3: Dataset degli attributi di polizza.

2.6 Dataset completo

Il dataset completo è composto quindi dal dataset delle percorrenze, il dataset dello stile di guida e il dataset dei dettagli relativi alla polizza.

Dataset	**N. totale records**	**N. polizze**	**N. sinistri**	**Dimensione**
2015	15.552.427	74.566	4.178	15.9 GB
2016	15.684.077	83.479	3.870	11.1 GB

Tabella 2.4: Statistiche dei due dataset completi

Le polizze vengono ulteriormente divise in 4 gruppi a seconda che le polizze siano nuove o vecchie, cioè se sia un rinnovo o una prima sottoscrizione, e se le polizze siano a rinnovo semestrale o annuale. La motivazione di questa scelta risiede nella differenza di correlazione con il livello di rischio di polizze nuove e polizze vecchie, all'interno del quale si distinguono ulteriormente le modalità di rinnovo. I dataset verranno quindi divisi dopo la fase di preprocessing degli attributi in 4 dataset separati per ogni anno. Su questi nuovi dataset saranno applicate tutte le tecniche di machine learning separatamente e si vedrà come lo stesso algoritmo possa essere più o meno performante al variare dei 2 attributi sopracitati. Si creano quindi 4 combinazioni di attributi:

- NA: polizze Nuove Annuali
- NS: polizze Nuove Semestrali
- VA: polizze Vecchie Annuali
- VS: polizze Vecchie Semestrali

Capitolo 3

Metodologia

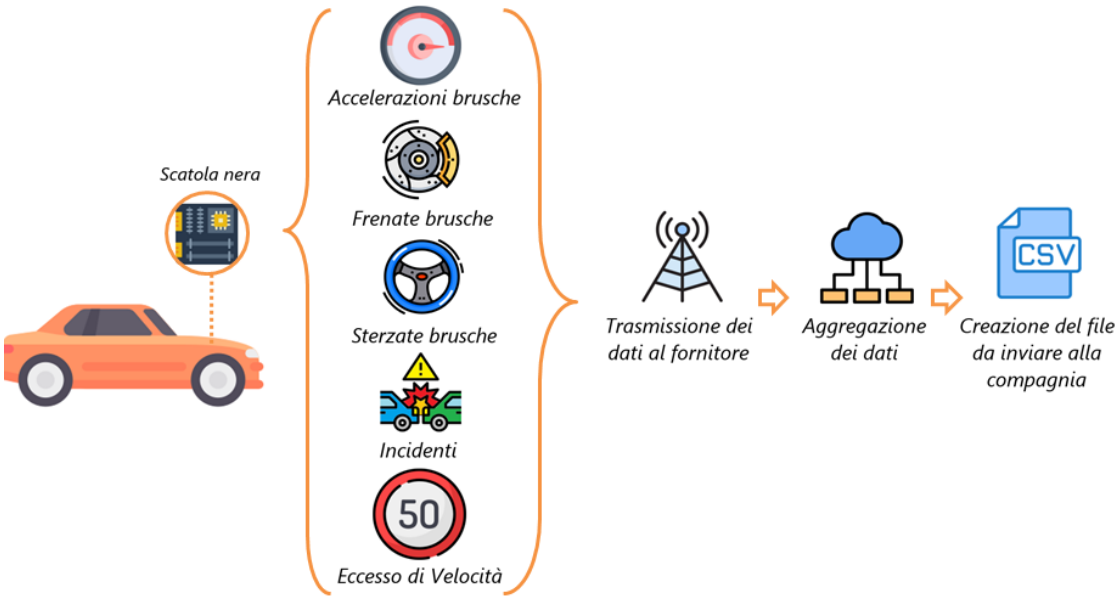

Figura 3.1: Dalla scatola nera ai dati

La scatola nera che viene montata sull'automobile del cliente ha la funzione di raccogliere le informazioni relative alle percorrenze, agli eventi di guida (es. frenate e accelerazioni brusche o superflue, sbandate, ecc.) e ad eventuali incidenti. Grazie a dei sensori interni, l'apparecchio è in grado di misurare i suddetti parametri e inviarli automaticamente, in un formato strutturato (es. JSON o XML) al gestore del servizio che li conserva, li aggrega e produce un file CSV (Comma Separated Value) da inviare alla compagnia assicurativa. La scatola nera invia le letture di dati aggregati secondo una certa temporizzazione definita dal produttore. Tali dati vengono trasmessi alla compagnia che li memorizza in un dataset di log, che tiene traccia dei dati telematici potenzialmente di interesse per le successive analisi. Un estratto di esempio del file in possesso della compagnia potrebbe essere il seguente:

Tramite un processo di cleaning dei dati, l'attenzione è stata rivolta a rendere la base dati coerente ed eliminare eventuali errori del dataset, quali dati mancanti o incongruenze nei valori degli attributi. La fase di processing dei dati è

NumPolizza, Data, KmPer, KmPerAut, N_Fren, N_SupVel, N_Accel, Incidenti, ...
154784, 10-DIC-2017, 28.3, 12.3, 7, 0, 3, 0, ...
178648, 10-DIC-2017, 37.2, 18.4, 1, 4, 5, 0, ...

stata effettuata mediante un software sviluppato appositamente ed è stata dedicata soprattutto all'aggregazione temporale dei dati, inizialmente aggregati su base giornaliera, scegliendo le tre aggregazioni settimanale, mensile e annuale. Come dimostrato durante le varie fasi di analisi, la ricerca di eventuali pattern comportamentali nello stile di guida ha dato risultati molto diversi in base alla granularità presa in esame.

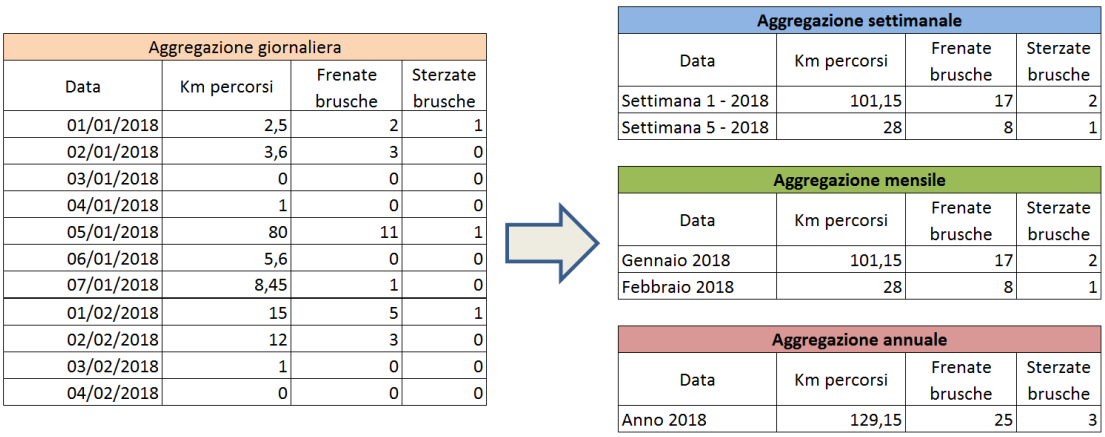

Figura 3.2: Schema di aggregazione

Il processo di analisi può essere diviso in 4 parti principali: preparazione dei dati, trasformazione dei dati (windowing), addestramento (training) del modello di classificazione su polizze con livello di rischio noto, applicazione del modello a nuove polizze con livello di rischio ignoto.

- La preparazione dei dati ha l'obiettivo di preparare i dati all'analisi successiva (gestione dati mancanti, discretizzazione, normalizzazione, selezione degli attributi d'interesse, ecc.). La discretizzazione e la selezione degli attributi possono essere variate tramite il settaggio di uno specifico file di testo.

- La trasformazione dei dati ha l'obiettivo di modellare la serie storica dei dati d'interesse in un formato idoneo all'applicazione dei modelli di classificazione, attraverso un processo basato su una finestra a scorrimento che raccoglie per ogni istante di tempo obiettivo della predizione gli n istanti precedenti. Questa tecnica, chiamata Windowing, viene effettuata tramite il software RapidMiner[1].

- L'addestramento ha l'obiettivo di applicare algoritmi di classificazione utili a predire il valore dell'attributo di classe (il livello di rischio in un dato istante)

[1]https://rapidminer.com/

in base ai valori dei restanti attributi (ad es. livello di rischio negli istanti precedenti, km percorsi negli istanti passati). Sono stati testati algoritmi presenti nella suite RapidMiner appartenenti a tre categorie differenti, ovvero alberi di decisione (DecisionTree, Random Forest), Bayesiani (Naive Bayes, WAODE), distance-based classifiers (K-NN) e classificatori a margine massimo (SVM). Ogni file creato nel passo di trasformazione viene usato come input del processo di classificazione, scegliendo uno alla volta i vari algoritmi.

- L'applicazione del modello è finalizzato ad applicare i modelli generati al passo precedente per predire il livello di rischio (ignoto) di una nuova polizza. I modelli saranno validati mediante tecniche standard di valutazione delle performance di classificazione, con particolare attenzione verso la capacità dei modelli di predire correttamente livelli di rischio alto (potenzialmente più critici per la compagnia).

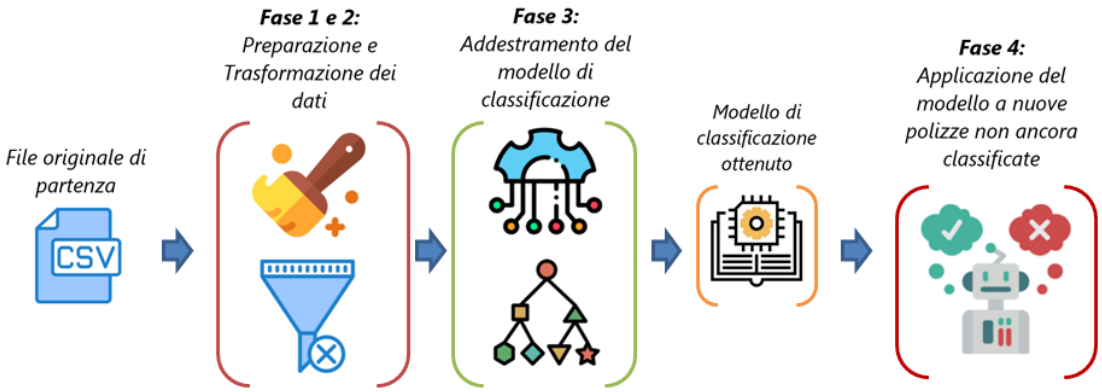

Figura 3.3: Processo completo di analisi

Le quattro fasi appena citate saranno approfondite nei capitoli successivi, elencando le motivazioni che hanno portato a tali scelte e i risultati ottenuti nelle differenti situazioni.

Capitolo 4

Machine Learning e Data Mining

Per *Machine Learning* si intende un ramo dell'intelligenza artificiale che si occupa della creazione di sistemi in grado di imparare dai dati, senza ricevere regole esplicite dal programmatore. I modelli di machine learning sono in grado di apprendere dai dati senza l'aiuto esplicito di un essere umano. Questa è la differenza principale fra i modelli di machine learning e i classici algoritmi. Negli algoritmi classici siamo noi a specificare il modo in cui individuare la soluzione, specificando una serie di passi da eseguire per passare dai dati iniziali al risultato desiderato. [4] [7]

L'apprendimento automatico viene diviso in tre gruppi (o paradigmi) in base al *segnale* utilizzato per l'apprendimento e il *feedback* disponibile al sistema di apprendimento.[8]Queste categorie sono:

- Apprendimento supervisionato
- Apprendimento non supervisionato
- Apprendimento per rinforzo

4.1 Apprendimento supervisionato

Nell' *apprendimento supervisionato*, l'algoritmo di apprendimento utilizza dei dati già classificati, al fine di predire il valore della categoria per i nuovi elementi. Il dataset di training dell'algoritmo contiene quindi sia i dati di input sia il risultato, e sulla base di esso, avviene il processo di *apprendimento*. A seconda di quale sia l'output desiderato, l'apprendimento supervisionato si suddivide ulteriormente in problemi di classificazione e problemi di regressione.

Per classificazione s'intende l'assegnazione di un oggetto ad una specifica classe di appartenenza. Questo tipo di problemi si occupa di assegnare un'etichetta discreta ad un oggetto, tramite la costruzione di un modello basato sull'osservazione dei dati di training. Tramite l'analisi di dati già classificati, il sistema, se ben addestrato, può essere capace di predire un'etichetta per dei dati ancora non classificati. (Esempio: si danno in input al sistema le analisi del sangue di tutti i malati di una certa patologia, dopo l'addestramento, ricevendo dei dati ancora non

etichettati di un nuovo paziente, il sistema deve essere in grado di predire la classe di appartenenza: malato/sano).

L'analisi della regressione è una tecnica usata per analizzare una serie di dati che consistono in una variabile dipendente e una o più variabili indipendenti. Lo scopo è stimare un'eventuale relazione funzionale esistente tra la variabile dipendente e le variabili indipendenti.Con la regressione si intende predire un valore numerico continuo, diversamente dalla classificazione in cui si cercano di predire dei valori discreti. Può essere applicata, per esempio, alla predizione di prezzi di un immobile, date le sue caratteristiche e il quartiere.

4.1.1 Algoritmi di apprendimento

4.1.2 Alberi decisionali

Un albero di decisione (Decision Tree) rappresenta un modello predittivo dove ogni variabile è rappresentata da un nodo interno, collegato ad uno o più nodi figli tramite un arco. La scelta di quale arco percorrere dipende dal valore della variabile relativa a quel nodo, rispetto al valore predetto per la variabile obiettivo. L'output può essere rappresentato anche graficamente sotto forma di un vero e proprio albero, in modo tale da evidenziare le scelte compiute dall'algoritmo.[7]

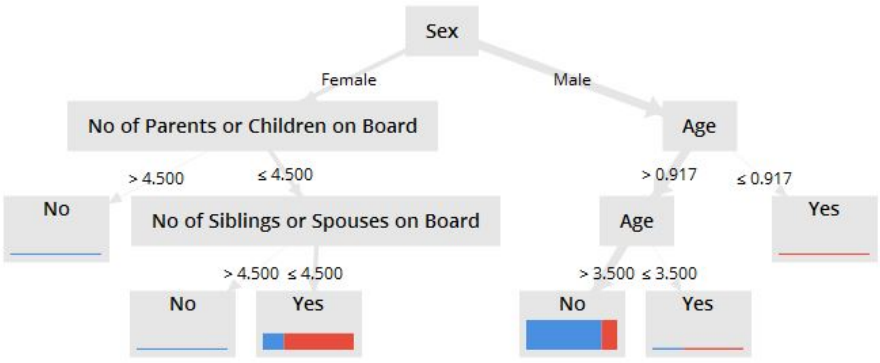

Figura 4.1: Esempio di Decision Tree

4.1.3 SVM - Support Vector Machine

Le SVM fanno parte di una famiglia di metodi di apprendimento *supervised* per la regressione e la classificazione. Sono conosciuti anche come classificatori a massimo margine, poiché cercano di minimizzare l'errore empirico di classificazione e massimizzare il margine geometrico. Nel caso di insiemi separabili linearmente, l'SVM ha il compito di costruire l'iperpiano di separazione (iperpiano ottimo) che renda massima la distanza tra gli elementi che appartengono a due classi differenti, rappresentati dai punti nel corrispondente iperspazio.(Figura 4.2) La tecnica di addestramento SVM può rappresentare funzioni non lineari complesse mettendo a disposizione un algoritmo estremamente efficiente. [6]

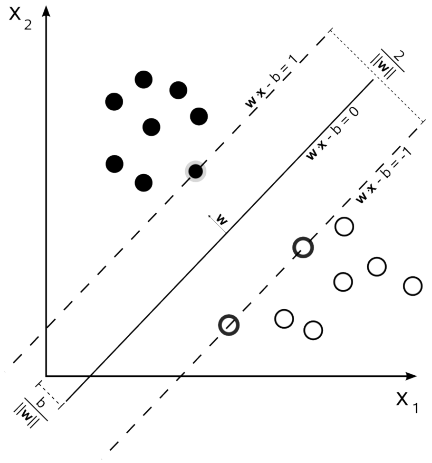

Figura 4.2: Esempio di separazione lineare, usando le SVM.

4.1.4 K-nearest neighbors(k-NN)

È l'algoritmo di classificazione più semplice. Si basa sulle caratteristiche degli oggetti vicino a quello considerato. Quando si è nella fase di decisione della classe a cui assegnare l'oggetto in esame, si guardano i sui k vicini. Se k=1 l'oggetto verrà assegnato alla classe del suo vicino, se k=3 l'oggetto verrà assegnato alla classe di cui i suoi vicini rappresentano la maggioranza. Se siamo in un contesto di classificazione binario, conviene usare un numero k dispari per evitare situazioni di parità.

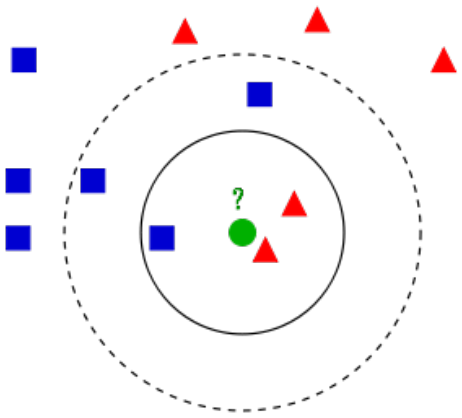

Figura 4.3: Esempio di k-NN (di Di A. Ajanki)

4.1.5 Random Forest

Una foresta casuale (Random Forest) è un classificatore d'insieme che offre la possibilità di ottenere una combinazione di modelli considerando, per la previsione, a ogni iterazione, diversi sottoinsiemi delle variabili, ottenendo così le stime da combinare. Una strategia di questo tipo prevede l'utilizzo di alberi come classificatori originali e la scelta casuale di variabili da inserire in ciascun modello. In realtà il termine Foresta Casuale ha un significato più generale e si riferisce a un qualsiasi classificatore ottenuto come combinazione di un insieme di classificatori ad albero.

4.1.6 Classificatori Bayesiani

I classificatori Bayesiani (Naive Bayesian Classifier) si basano sul teorema di Bayes secondo cui l'effetto di una variabile su una classe è indipendente dai valori delle altre variabili. Questa condizione è chiamata indipendenza condizionale e ha come obiettivo la semplificazione dei calcoli, motivo per cui l'algoritmo prende il nome di "naive[1]".[7]

4.2 Apprendimento non supervisionato

Nell'*unsupervised learning*, invece, il processo di apprendimento avviene in autonomia da parte dell'algoritmo, senza la conoscenza aggiunta dall'esperto di dominio o dall'analista. L'algoritmo cerca di imparare da solo a interpretare i dati. Il clustering e l'estrazione di regole associative fanno parte di questa famiglia. Le regole di associazione sono uno dei metodi per estrarre dai dati delle relazioni (potenzialmente) nascoste. Queste regole permettono di individuare collegamenti in ampi insiemi di dati e vengono spesso utilizzate per analizzare grandi quantità di dati alla ricerca di associazioni utili. Strutture come i supermercati o i siti di e-commerce da molti anni utilizzano i dati di acquisto dei clienti per realizzare pubblicità mirate e per migliorare l'organizzazione dei prodotti. Tramite le regole d'associazione si può descrivere una correlazione tra due fatti, scrivendo $A \rightarrow B$ si intende dato X è presente Y. La bontà di una regola è definita da due parametri: il supporto e la confidenza. Il supporto identifica quante volte X e Y appaiono percentualmente nell'insieme dei dati, la confidenza invece indica in percentuale quante volte la regola

4.3 Validazione

Per misurare l'accuratezza del modello creato nella fase di classificazione, bisogna effettuare un altro passo fondamentale che è quello della validazione. Attraverso la validazione si può testare il modello di predizione su una porzione di dataset che non ha fatto parte dei dati su cui il modello è stato addestrato (training set) e prende il nome di test set. Applicando il modello sul test set, si possono valutare alcuni indicatori quali la precisione, il richiamo, l'f-score ecc., che indicano la qualità del modello appena creato. Attraverso la matrice di confusione si possono esaminare i parametri sopracitati per tutte le classi della predizione.

4.3.1 Windowing

Il windowing è usato tipicamente per convertire dei dati di una serie storica in un dataset in formato *cross-section*. Con il termine cross-section si intende un tipo di studio basato su un campionamento trasversale. Gli studi cross-section forniscono

[1]Naive(o naïf): Ingenuo, schietto, primitivo.

solo indirettamente un'evidenza circa gli effetti di tempo e devono essere usati con grande cautela quando si traggono conclusioni circa il cambiamento. [2]

Dopo aver applicato il *windowing* su una serie storica, possiamo applicare degli algoritmi di predizione per prevedere un valore futuro.

4.3.2 Trasformazione della serie storica

Una serie storica consiste di un insieme di osservazioni, su un certo fenomeno, ordinate nel tempo. Ciascun valore può rappresentare una quantità rilevata in un istante temporale, ad esempio la temperatura corporea misurata ogni ora, il prezzo di un'azione registrato settimanalmente in alla chiusura della Borsa, oppure può derivare dall'accumulo di quantità rilevate su un intervallo temporale, ad esempio il consumo mensile di energia elettrica, il prodotto interno lordo ai prezzi di mercato misurato trimestralmente.

L'obiettivo dell'analisi delle serie storiche è quello di studiare le relazioni tra tali variabili casuali attraverso la costruzione di modelli miranti o all'ottenimento di previsioni di breve periodo o alla scomposizione della serie storica in un insieme di componenti latenti.[1].

4.3.3 Sliding Window Validation

La Sliding Window Validation è il metodo di validazione che si adatta particolarmente bene ai dataset su cui è stato effettuata l'operazione di windowing. La *finestra* è composta da un intervallo di dati (consecutivi nel tempo) usati come training window e una parte che viene usata come test del modello imparato dal training set.

Dopo aver scelto uno tra i metodi di classificazione, bisogna validare il modello. Tramite la Sliding Window Validation si può operare un tipo di validazione che fa *scorrere* in avanti la finestra di training, insieme a quella di test. La classificazione avviene valutando solo un certo intervallo di tempo e il test avviene su una finestra di tempo successiva. L'idea alla base è quella di cercare di prevedere il verificarsi di un particolare evento in un certo istante di tempo futuro, in base all'osservazione di eventi del passato.

I parametri principali che influenzano e determinano l'analisi sono:

- Training Window: Determina la quantità di eventi passati che voglio includere nella fase di addestramento del modello.
- Testing Window: La lunghezza del periodo su cui vado a fare la predizione.
- Horizon: Rappresenta *l'orizzonte* della mia predizione, ovvero la distanza tra l'ultimo evento facente parte della training window e il primo della testing window.
- Step: Determina lo step di scorrimento della finestra.

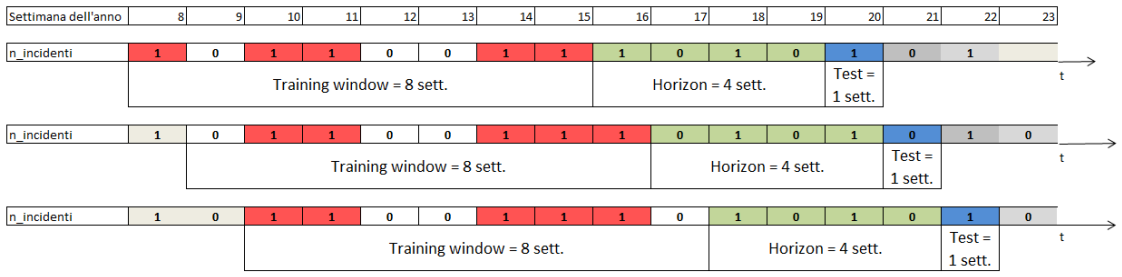

Figura 4.4: Sliding Window Validation

4.3.4 Expanding window

Con il metodo dell'expanding windows la finestra di training aumenta di dimensione a ogni step di training. Si parte con la finestra di una certa dimensione e si decide a ogni step di quanto deve ingrandirsi. Ciò consente di avere un training set cumulativo, cioè ogni volta che la finestra si ingrandisce, vengono considerati sempre più istanti di tempo per il training in contrapposizione con la Sliding Window Validation in cui la finestra di training ha una dimensione fissa.

4.3.5 Cross-Validation

La Cross-Validation è una tecnica usata per validare dei modelli predittivi partizionando il campione originale in un training set per addestrare il modello e un test set per validarlo. Nella k-fold cross-validation, il dataset originale è diviso in k partizioni uguali. Il criterio di campionamento può essere deciso tra Stratified-sampling (ogni campione rispetta la distribuzione delle classi nel dataset originale), Shuffled-sampling (campioni partizionati in maniera casuale), Linear Sampling (il partizionamento è fatto in maniera lineare, in base alla disposizione dei records nel dataset). Delle k partizioni create, k-1 vengono utilizzate come training set, per addestrare il modello, mentre la k-esima partizione funge da test set. Questo processo viene ripetuto k volte, variando ogni volta il campione di test e alla fine i k risultati ottenuti sono combinati per produrre una singola stima del modello creato.

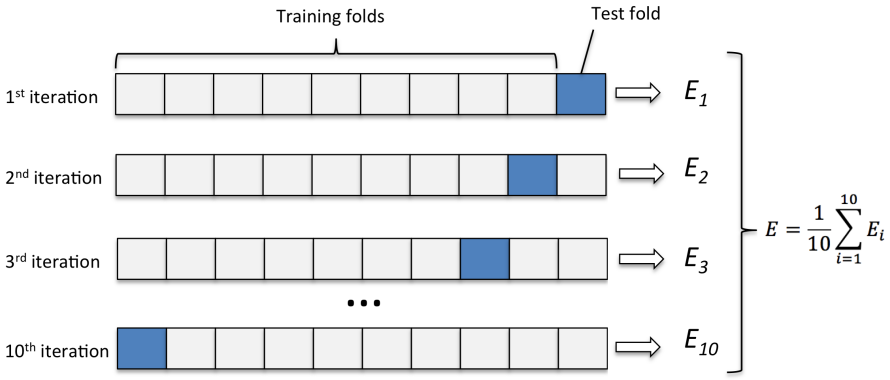

Figura 4.5: Cross-Validation con k=10

Il vantaggio di questa tecnica è che tutti i records sono usati sia per il training che per la validazione mentre ogni record è usato per la validazione esattamente una volta. In questo modo si evita il fenomeno dell'overfitting, ovvero quando il modello si adatta a caratteristiche che sono specifiche solo del training set, ma che non hanno riscontro nel resto dei casi; perciò, in presenza di overfitting, le prestazioni (cioè la capacità di adattarsi/prevedere) sui dati di allenamento aumenteranno, mentre le prestazioni sui dati non visionati saranno peggiori.

4.3.6 Hold-out

Con il metodo Hold-out dividiamo il dataset di partenza in due set, il training set e il test set (solitamente più piccolo del training set), e su questi due set effettuiamo una sola "run". Cioè il modello viene costruito sul training set e viene testato sul test set. Si contrappone quindi alla cross validation dove invece si effettuano k "run" di apprendimento e altrettante di validazione, con l'operazione di media finale di tutti i modelli creati. In un certo senso questo metodo può essere considerato come il tipo più semplice di cross-validation.

4.3.7 Trasformazione della serie storica

Una serie storica consiste di un insieme di osservazioni, su un certo fenomeno, ordinate nel tempo. Ciascun valore può rappresentare una quantità rilevata in un istante temporale, ad esempio la temperatura corporea misurata ogni ora, il prezzo di un'azione registrato settimanalmente in alla chiusura della Borsa, oppure può derivare dall'accumulo di quantità rilevate su un intervallo temporale, ad esempio il consumo mensile di energia elettrica, il prodotto interno lordo ai prezzi di mercato misurato trimestralmente.

L'obiettivo dell'analisi delle serie storiche è quello di studiare le relazioni tra tali variabili casuali attraverso la costruzione di modelli miranti o all'ottenimento di previsioni di breve periodo o alla scomposizione della serie storica in un insieme di componenti latenti.[1].

4.4 Tecniche di campionamento e meta-algoritmi

4.4.1 Oversampling e Undersampling

L'oversampling (sovracampionamento) e l'undersampling (sottocampionamento) rappresentano due tecniche che si rendono utili quando la distribuzione delle classi nel dataset è molto sbilanciata. Nel caso di una classificazione binomiale, se le classi nel training set sono molto sbilanciate (es. 99% incidente_no, 1% incidente_si) il processo di apprendimento può essere distorto, perché il modello tende a focalizzarsi sulla classe prevalente e ignorare gli eventi rari. L'uso di metodi di campionamento consiste nella modifica di un set di dati sbilanciati attraverso alcuni meccanismi in modo da fornire una distribuzione equilibrata.

Le tecniche più comuni sono il random oversampling che attua un campionamento con ripetizione delle osservazioni appartenenti alla classe rara e il random undersampling che, al contrario, effettua un campionamento senza ripetizione tra le osservazioni appartenenti alla classe maggioritaria. In altre parole, il random oversampling è un metodo che mira a bilanciare la distribuzione di classe attraverso la replicazione casuale di esempi appratenti alla classe minoritaria. L'oversampling aumenta la predisposizione del modello all'overfitting, cioè il fenomeno per cui il modello si comporta bene con i dati presenti nel dataset bilanciato con oversampling, ma non funziona in maniera appropriata su dati reali. L' undersampling invece comporta la perdita di informazioni poiché il campionamento della classe maggioritaria avviene in maniera casuale.

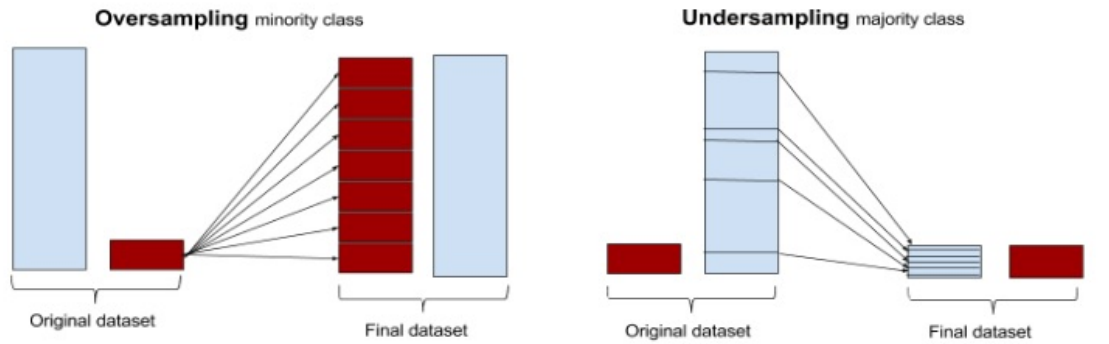

Figura 4.6: Oversampling e undersampling

4.4.2 Bagging

Il bagging fa parte dei cosiddeti meta-algoritmi del machine learning. E' un modo per diminuire la varianza della predizione generando dati addizionali nel training set, usando combinazioni con ripetizione per produrre dei multiset della stessa cardinalità dei dati originali. Aiuta a ridurre l'overfitting e si adatta ad essere utilizzato principalmente con gli alberi decisionali [2], ma può essere implementato con tutti gli algoritmi.

4.5 Misure di qualità

Per valutare in modo oggettivo l'efficacia di un modello di classificatore, sono state considerate le misure standard di qualità definite a partire dalla matrice di confusione. In particolare, la matrice di confusione riporta sulle righe i valori predetti dal classificatore (per semplicità i livelli di rischio basso e alto, rispettivamente), mentre sulle colonne sono indicati i valori attesi. L'accuratezza misura la percentuale di dati classificati correttamente (indipendentemente dalla classe considerata). Nel contesto specifico, indica la percentuale di polizze per cui il classificatore ha

[2] vd. 4.1.1

predetto correttamente il livello di rischio (alto o basso). Siccome nel dominio analizzato le due classi non hanno la stessa importanza, è utile valutare l'efficacia del classificatore nel predire una classe specifica.

La precisione di una classe (ad es. rischio alto) indica la percentuale di dati a cui è stata predetta correttamente la classe in esame; nel nostro contesto, ci focalizziamo sulla precisione del classificatore sulla classe ad alto rischio, in quanto indica il rapporto tra il numero di polizze a cui è stato assegnato correttamente un rischio alto e il numero totale a cui è stato assegnato un rischio alto. La capacità di assegnare correttamente un rischio alto implica una corretta gestione delle situazioni a rischio. Il richiamo di una classe indica la indica la percentuale di dati a cui è stata predetta la classe in esame rispetto al numero totale di dati appartenenti a tale classe. Nell'esempio precedente, il richiamo indica la percentuale di polizze a rischio alto a cui è stata effettivamente assegnata una classe di rischio alto. L'F-score di una classe è una misura comunemente usata per tenere conto sia della precisione e del richiamo di un classificatore nel valutarne l'efficacia rispetto a una classe; si calcola come la media armonica delle due misure precedentemente descritte.

accuracy: 50.87% +/- 8.09% (micro average: 50.89%)

	true incidente_si	true incidente_no	class precision
pred. incidente_si	78	75	50.98%
pred. incidente_no	91	94	50.81%
class recall	46.15%	55.62%	

Figura 4.7: Esempio di matrice di confusione

Capitolo 5

Stato dell'arte

Il settore assicurativo mondiale conta più di migliaio di compagnie con un giro d'affari che supera i mille miliardi di dollari ogni anno. Da sempre le compagnie assicurative hanno dedicato particolare attenzione all'innovazione tecnologica e grazie allo sviluppo, soprattutto negli ultimi anni, di tecnologie improntate alla sensoristica, alla localizzazione dei veicoli, alla profilazione dei clienti, la quantità di dati in loro possesso è aumentata esponenzialmente. L'aumento della varietà dei dati, nonché del loro volume, ha offerto una solida base per tutta una serie di studi e ricerche scientifiche che mirano ad estrarre conoscenza dai dati e a sfruttarli per fini di prevenzione, predizione e analisi di sinistri. Nonostante le tecniche di machine learning non siano recentissime, con l'avvento delle tecnologie Big Data si sono aperti nuovi fronti di analisi e spesso i risultati, inizialmente guardati con sospetto, sono diventati sempre più entusiasmanti. Gli studi e le ricerche analizzate durante lo svolgimento di questa tesi si sono rivelati spesso simili nel fine perseguito e nelle tecniche utilizzate ma, altrettanto spesso, diversi nei dati utilizzati e nel tipo di fonte dei medesimi dati, abbracciando ambiti come l'analisi del traffico veicolare in determinate città al fine di predire un sinistro, il rilevamento di frodi assicurative sia in ambito medico che veicolare, la predizione di un premio assicurativo per un determinato cliente. L'argomento della predizione degli incidenti è attualmente molto caldo. La rivista "Accident analysis & prevention" (Ed. Elsevier) raccoglie ogni mese i paper più importanti relativi al settore della predizione degli incidenti in ogni ambito (medico, legale, economico, educativo, comportamentale, legato ai trasporti).

Lo studio [10] di Vassilijeva et al. condotto nel 2017 ha cercato di classificare il livello di rischio di 540.000 polizze basandosi su 17 attributi di polizza (sesso, età dell'auto, alimentazione del veicolo, ...) tramite le reti neurali, riuscendo a identificare solo 81 polizze con un rischio di incidente >12.5%. Uno dei problemi maggiori secondo i ricercatori che hanno condotto questo esperimento è lo sbilanciamento delle classi (incidente-si, incidente-no) che non consente alle reti neurali di ottenere performance di rilievo; come si evidenzierà più avanti, anche in questa tesi si è dovuto affrontare questo tipo di problema. I Tra tutti i lavori analizzati in letteratura lo studio [9] ha usato come dati di ingresso quelli dei rilevatori di traffico real-time sulle strade di Atene dal 2006 al 2011, incrociandoli con i dati meteorologici e i dati relativi ai tipi di veicoli che attraversavano quel tratto stradale. Tramite le SVM (Support Vector Machine) lo studio aveva lo scopo di creare un modello in

grado di predire il livello di rischio di incidente per i motocicli. Si è dimostrato che, tramite l'uso delle serie storiche, si riusciva a predire un coinvolgimento di un motociclo in un sinistro con una accuracy del 63-64%. Nello studio [5] si utilizzano dei parametri GPS registrati sui veicoli di 1500 guidatori per 2 anni, per profilare il livello di rischio di ogni guidatore. È lo studio che più si avvicina al nostro ma ha una quantità di dati in ingresso molto minore e non utilizza nessuna tecnica di machine learning per il clustering. Le osservazioni vengono fatte soprattutto da un punto di vista statistico e hanno lo scopo di classificare in due categorie in base al numero di metri percorsi e al numero di viaggi effettuati. Sotto i 110 viaggi per mese e i 2.000 km al mese la classificazione dei guidatori che compiono un incidente è corretta rispettivamente del 55.7% e 67%.

Per quanto riguarda l'argomento della rilevazione delle frodi, la letteratura è molto ricca a riguardo. In "Fraud Detection and Frequent Pattern Matching in Insurance claims using Data Mining Techniques" di Aayushi Verma et al.[11] si utilizza il mining basato sulle regole di associazione per scoprire eventuali pattern che nascondano una frode. Viene applicata la tecnica di clustering *k-means* e vengono ricercati eventuali outliers basati sia su criteri clinici che su anomalie nei report dei pazienti. Su un dataset di 275.000 casi vengono riscontrati 210 casi sospetti. Lo studio [3] ha utilizzato i dati dei sinistri di Taiwan e ha dimostrato come le polizze che prevedevano una copertura più alta in caso di incidente fossero quelle più coinvolte in incidenti. La conclusione a cui sono arrivati gli autori è che le persone meno abili alla guida tendono ad acquistare polizze con una copertura maggiore e quindi la correlazione è data dalla poca abilità di guidare e non sembra legata (in quel caso) ad un eventuale tipo di truffa.

Il problema di assegnare una categoria di rischio a una determinata polizza (automobilistica o sanitaria) è abbastanza frequente in letteratura e per alcuni settori come il fraud detection, utilizzando il machine learning, si è arrivati a risultati interessanti, tuttavia nel settore assicurativo mancano ancora delle ricerche che dimostrino i rischi collegati ad un determinato stile di guida e soprattutto, anche le ricerche che più si orientano alla predizione di rischio legata ai sinistri, soffrono spesso lo sbilanciamento delle classi che abbassano notevolmente le capacità di "intuizione" e apprendimento del modello. Questo lavoro di tesi ha potuto beneficiare di un accesso diretto ai dati in possesso della compagnia assicurativa, cosa che in letteratura è abbastanza rara, come dimostrato anche dalla scarsità di studi simili riscontrati.

Capitolo 6

Software di pre-processing dei dati

6.1 Informazioni tecniche

Requisiti di sistema:

- Sistema operativo: Windows 7/8/10
- Librerie installate: Oracle Java SE 8 o superiore
- RAM: >1GB

File di input:

- Dataset non processato in formato .csv (per la parte di processing) [datset/-dataset1.csv nella cartella di prova]
- Dataset con aggregazione annuale discretizzata (per la parte di estrazione delle regole di associazione) [dataset/dataset2.csv nella cartella di prova]
- File di scelta degli attributi [settings/chooseFields.txt nella cartella di prova]
- File di settaggio delle discretizzazioni settimanale (W), mensile (M) o annuale (Y)[file nella cartella settings]
- File di settaggio delle features da includere nell'analisi di correlazione [vari esempi nella cartella featureSelection]

File di output:

- Un file .csv per ogni aggregazione scelta (parte di processing)
- Un file .txt contenente tutte le regole d'associazione estratte
- Un file .txt contenente le regole d'associazione estratte che superano una certa soglia di lift (impostato nella parte di analisi estrazione delle association rules)

I dettagli relativi al formato dei file verranno specificati ulteriormente nelle successive sottosezioni.

6.2 Scopo dell'applicazione

Prima di poter effettuare qualsiasi tipo di analisi, si è resa subito necessaria un'operazione di preparazione del dataset che si occupasse di rispondere a quattro criticità principali riscontrate nei dataset:

- **Scelta della granularità temporale da mantenere in tutto il dataset:** il dataset di partenza ha una granularità giornaliera, ma è composto da due dataset che in origine avevano due granularità temporali diverse (uno settimanale e uno giornaliera).

- **Eliminazione dei duplicati e gestione dei dati mancanti:** Dire nel paragrafo che il dataset fornito la compagnia conteneva dati ridondanti prodotti da operazioni di join su tabelle intermedie e dati mancanti dovuti a registrazioni non effettuate, errori di inserimento o valori non disponibili. Dopo aver verificato che il numero di dati mancanti era limitato e aver identificato i dati duplicati, si è provveduto all'eliminazione di tali dati.

- **Propagazione dell'attributo "DATA _INCIDENTE':** questo attributo indica, per la riga del rispettivo giorno, se sia avvenuto o no un sinistro. L'operazione di propagazione va a settare come "1' tutti i giorni della settimana di riferimento (insieme agli altri attributi relativi ai costi del sinistro).

- **Selezione parziale degli attributi del dataset**: questa fase consiste nella selezione di un sottoinsieme di attributi più significativi su cui addestrare i modelli di classificazione. I restanti attributi, considerati meno rilevanti o ridondanti, saranno rimossi e ignorati nelle successive analisi.

L'applicazione sviluppata parte dalla necessità di rispondere alle quattro criticità principali sopracitate, ma espande ulteriormente gli strumenti di modifica del file per rispondere a una necessità di flessibilità e soprattutto ad un'esigenza di utilizzo anche da parte di un utente non esperto di programmazione. L'applicazione è idealmente divisa in due parti: la prima di processing del file, la seconda di estrazione delle regole associative. L'intera applicazione è sviluppata in linguaggio *Java* ed è stato utilizzato il plugin di Eclipse "Window Builder"[1] per sviluppare l'interfaccia utente.

Il software sviluppato si pone come solido punto di partenza per eventuali analisi future già previste e si è reso utile per perseguire scopi di ricerca differenti, in altri lavori di Tesi paralleli a questo, nati all'interno dello stesso progetto di ricerca. Nella fase finale di sviluppo, si è deciso di aggiungere alcune funzionalità in grado di consentire anche ad un utente meno esperto di estrarre delle regole associative[2] tramite l'algoritmo L3.

[1] www.eclipse.org/windowbuilder

[2] vd. Capitolo 4.2

6.3 Processing del file

La parte di processing del file, che prepara il dataset per le analisi successive, consente di effettuare tre macro operazioni:

- Aggregazione temporale: consente di scegliere con quale granularità aggregare i dati: settimanale, mensile o annuale. Le tre differenti aggregazioni possono avere un ruolo chiave a seconda del tipo di analisi effettuata.
- Selezione degli attributi: si può scegliere un sottoinsieme dei 164 attributi.
- Discretizzazione dei valori degli attributi: si possono discretizzare i valori degli attributi secondo degli intervalli decisi dall'utente.

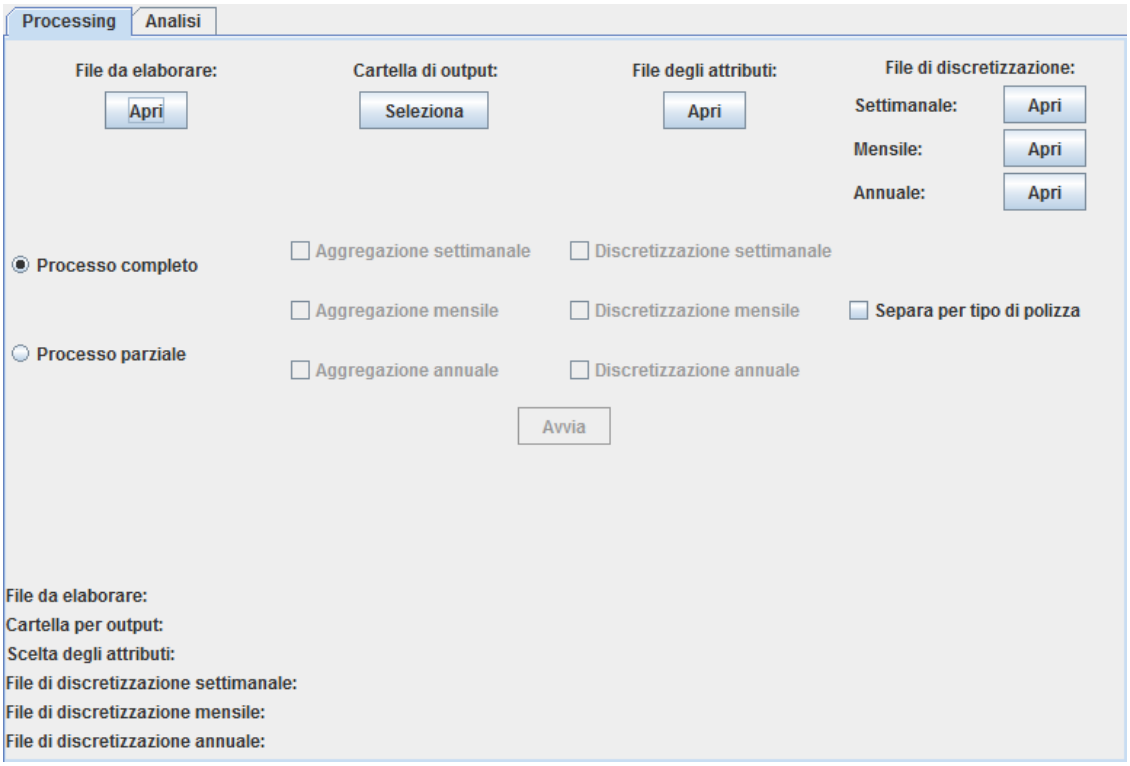

Figura 6.1: Home dell'applicazione

6.3.1 Manuale utente

Per utilizzare la fase di pre-processing occorre settare due file di testo inclusi nella directory principale dedicati alla selezione degli attributi e alla discretizzazione. Servono per indicare quali attributi si vogliono mantenere e quali si vuole eliminare. Mandando in input un file .csv contenente il dataset completo, vengono generati due file .txt in cui c'è la lista degli attributi del dataset, inizialmente settati tutti a 1 (che indica la selezione di quel particolare attributo), mentre il file di discretizzazione setta a 0 tutti gli attributi (inizialmente non voglio discretizzare nessun attributo). Il formato utilizzato è il seguente:

- File di selezione: Nome_attributo [0/1]
- File di discretizzazione: Nome_attributo; [0/1]; numero intervalli di discretizzazione; intervalli numerici; nomi delle etichette di ogni intervallo

```
NUMERO_SECONDI_VIAGGIATI_06 0
NUMERO_SECONDI_VIAGGIATI_07 0
NUMERO_METRI_PERCORSI_DIURNI 1
NUMERO_METRI_PERCORSI_NOTTURNI 1
NUMERO_SECONDI_VIAGGIATI_DIURNI 0
NUMERO_SECONDI_VIAGGIATI_NOTTURN 0
```

(a) Selezione attributi

```
NUMERO_SECONDI_VIAGGIATI_06;1;3;3600,7200;<1ora,1-2ore,>2ore
NUMERO_SECONDI_VIAGGIATI_07;1;3;3600,7200;<1ora,1-2ore,>2ore
NUMERO_METRI_PERCORSI_DIURNI;0
NUMERO_METRI_PERCORSI_NOTTURNI;0
NUMERO_SECONDI_VIAGGIATI_DIURNI;1;3;3600,7200;<1ora,1-2ore,>2ore
NUMERO_SECONDI_VIAGGIATI_NOTTURN;1;3;3600,7200;<1ora,1-2ore,>2ore
```

(b) Discretizzazione attributi

Figura 6.2: Esempio di selezione e discretizzazione

Scrivendo 0 o 1 dopo l'attributo desiderato indico se voglio conservare/discretizzare l'attributo (1) o no (0). Tramite questi due file si può cambiare il numero di attributi selezionati o gli intervalli di discretizzazione e le etichette prima di ogni analisi. Ovviamente i parametri inseriti devono essere corretti e coerenti tra loro, altrimenti il programma segnala un errore. (Es. il numero di etichette deve essere uguale al numero di intervalli). Dopo aver settato i parametri di interesse, tramite la Home dell'applicazione, vanno settati i vari file su cui si andranno a effettuare le varie operazioni di aggregazione/discretizzazione. Bisogna selezionare obbligatoriamente il file da processare (contenente il dataset originale), il file della selezione degli attributi (generato al passo precedente) e la cartella in cui andranno a finire tutti i file generati. Opzionalmente si possono impostare anche i file di discretizzazione.

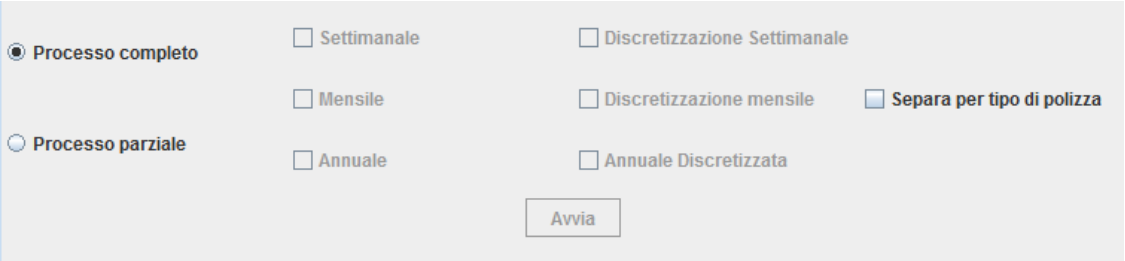

Figura 6.3: Esempio di Processo Completo

Selezionando l'opzione "Processo completo", il file verrà processato e saranno generate tutte le aggregazioni disponibili, cioè settimanale, mensile e annuale; in più, se al passo precedente sono stati settati anche i file di discretizzazione, saranno generate le discretizzazioni (nel caso in cui si carichi solo il file di discretizzazione settimanale, verrà generato un file discretizzato solo per quella aggregazione). Nel caso in cui non si vogliano ottenere tutte le aggregazioni, bisogna selezionare l'opzione "Processo parziale", e aggiungere manualmente le aggregazioni desiderate. Le selezioni riguardo alla discretizzazione saranno disponibili solo se è stato caricato il rispettivo file. È disponibile anche l'opzione "Separa per tipo di polizza" che divide prima il dataset per tipo di polizza (NA, NS, VA, VS) e poi effettua la discretizzazione (eventuale) su ognuno dei quattro dataset generati.

Al termine delle impostazioni, cliccando sul pulsante "Avvia" si manda in esecuzione il processo generando quindi i file di output delle varie aggregazioni. Il processo esegue i seguenti passi in ordine:

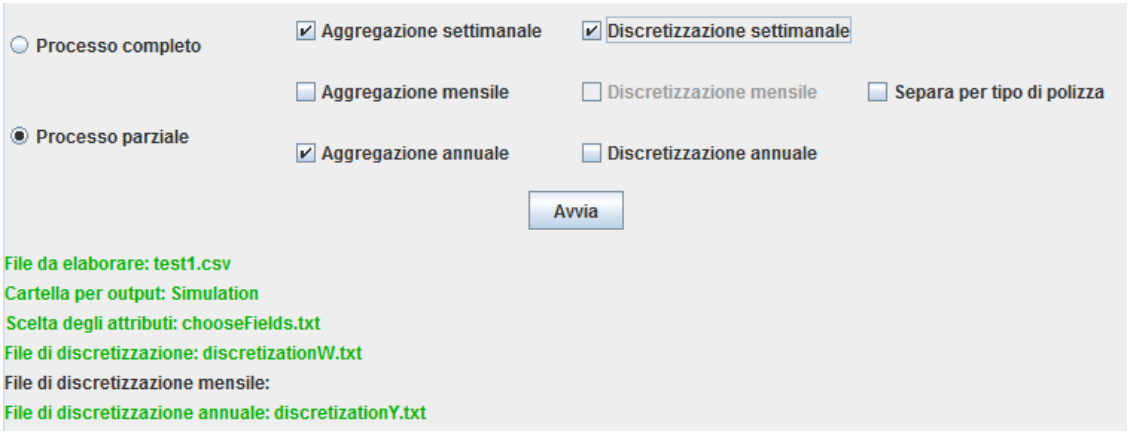

Figura 6.4: Esempio di Processo Parziale

1. Lettura dei file di attributi e discretizzazione: a partire dai due file creati vengono implementate due mappe che manterranno l'informazione su quali attributi utilizzare, quali no e come discretizzarli.

2. Cleaning: Vengono rimossi i gli attributi che sono stati indicati con "0" nel file di scelta degli attributi e vengono rimossi i record che presentano dei campi vuoti per degli attributi importanti del dataset (Numero di polizza, N. settimana, ecc.)

3. Aggregazione settimanale: viene subito fatta l'aggregazione settimanale in modo da ridurre il più possibile la dimensionalità del file prima di poter effettuare le operazioni successive, gli attributi legati alle percorrenze restano gli stessi mentre quelli legati ai costi dell'incidente vengono sommati su tutti i giorni della settimana.

4. Propagazione di DATA_INCIDENTE: come spiegato nel Capitolo 2, quando si verifica un incidente in un determinato giorno, il campo acc_sc di quel record riporterà la data del giorno, al contrario di un giorno senza incidenti che avrà il campo DATA_INCIDENTE vuoto. La data viene sostituita semplicemente da un "1" che indica l'avvenuto sinistro e ogni riga della settimana contenente il giorno dell'incidente verrà posta a 1 nel campo di DATA_INCIDENTE.

5. Aggiunta attributo mese e aggregazione mensile: prima di poter aggregare i dati mensilmente va aggiunto il campo Mese che inizialmente non è presente nel dataset. Viene calcolato approssimativamente con la formula $N_{Mese}=N_{Settimana}*7/30+1$ limitando il risultato a 12 per correggere il risultato dell'ultima settimana che risulterebbe nel mese 13.

6. Aggregazione mensile e annuale: dopo aver aggiunto l'attributo relativo al mese è possibile aggregare quindi su base mensile e successivamente settimanale. Vengono sommati tutti i dati relativi alle percorrenze, agli eventi di guida, ai costi di eventuali sinistri poiché calcolati su base settimanale. Ogni file viene iterato solamente una volta, mantenendo quindi la complessità O(n).

7. Discretizzazione: se abilitata, viene effettuata la discretizzazione sui file aggregati appena prodotti iterando su tutti i file solo una volta e cambiando i

vari attributi in base alla mappa creata a partire dal file di discretizzazione caricato.

8. (Opzionale) Divisione per tipo di polizza: se abilitata, divide i dataset appena creati in 4 partizioni ognuno (NA, NS, VA, VS)

6.4 Estrazione delle regole di classificazione

Tramite il tab "Analisi" presente sulla finestra principale, è possibile passare alla modalità in cui è possibile estrarre delle regole di classificazione tramite l'algoritmo L3 implementato in Java [3]. L'utente prima di avviare l'analisi, deve decidere quali parametri di supporto, confidenza e lift utilizzare. Il software si occuperà di estrarre le regole d'associazione in base alle impostazioni settate e le scriverà all'interno di un file .txt.

6.4.1 Manuale utente

Come input occorre caricare un file con discretizzazione annuale, preparato eventualmente tramite il processo di aggregazione disponibile nel tab di Processing, e bisogna settare i parametri di supporto e confidenza. Il secondo file da inserire in input serve per selezionare le features a cui si è interessati, cioè di quali caratteristiche si vogliono studiare le correlazioni. È composto da una lista di attributi che eventualmente possono essere modificati rispetto ai file originali. Per selezionare invece l'attributo di rischio che si vuole analizzare, bisogna sceglierlo nel pannello "Scelta dell'indicatore di rischio". Le sigle corrispondono rispettivamente a: Numero Incidenti Causati non gravi (NCD), Numero Incidenti gravi causati (NNC), Costo Incidenti Causati (CCC), Numero Totale Incidenti causati (NCC). Vengono generati due file, uno contiene tutte le regole estratte, mentre un secondo file (desinenza "-lift") contiene le regole filtrate per lift desiderato.

Il processo di generazione delle regole di associazione può risultare poco efficace nel caratterizzare la classe minoritaria (livello di rischio alto), in quanto la soglia di supporto minimo scelta (ad es. 1%), 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. Mentre le regole caratterizzanti un livello di rischio basso con un supporto superiore al 1% sono potenzialmente numerose, la stessa soglia applicata su profili relativi al livello di rischio alto può filtrare la maggior parte dei casi potenzialmente rilevanti.

Per ovviare al problema dello sbilanciamento nei dati tra le polizze appartenenti a diversi livelli di rischio, si può scegliere di applicare la soglia di supporto dell'1% sulla porzione di dati relativi a ciascun livello di rischio. Per calcolare il valore assoluto di supporto da applicare come soglia per una regola associata di un dato livello di rischio, si pesa la soglia fissa (1%) per la frequenza di occorrenza del livello

[3] www.dbdmg.polito.it/wordpress/research/associative-classification/

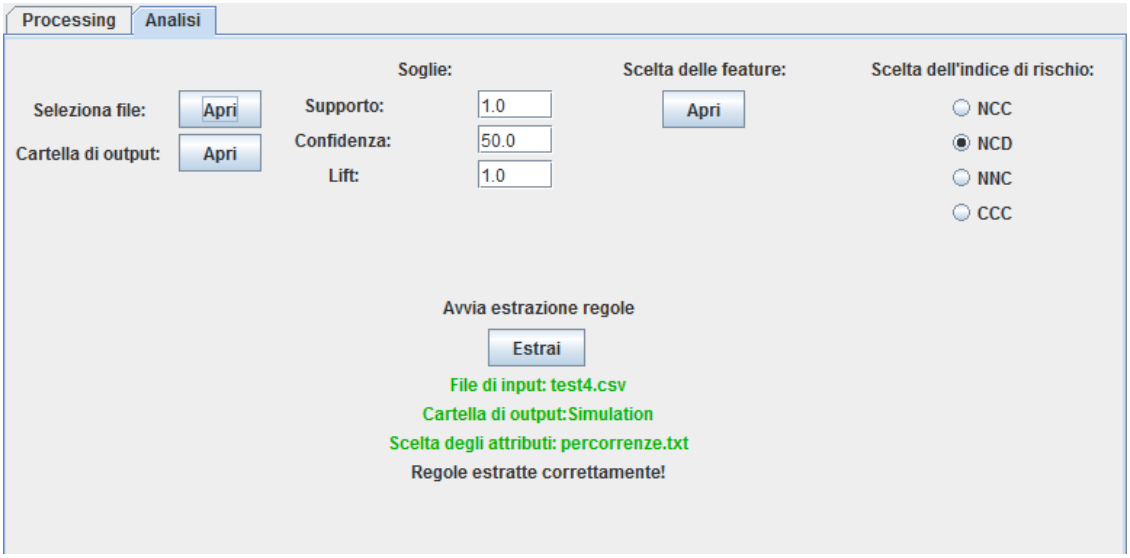

Figura 6.5: Tab di estrazione delle regole associative

di rischio nel dataset originale. In caso i livelli di rischio risultino esattamente bilanciati, la medesima soglia viene applicata per tutte le regole (indipendentemente dal livello di rischio). Al contrario, in caso vi sia un significativo sbilanciamento la soglia applicata sulle regole relative al livello di rischio minoritario viene incrementata in proporzione al percentuale di sbilanciamento esistente rispetto alla classe maggioritaria.

Capitolo 7

Risultati sperimentali

Sono state applicate tecniche standard di validazione per valutare l'efficacia e l'efficienza dei modelli di predizione sui dati analizzati. In particolare, sono state considerate tecniche basate su una finestra a scorrimento (sliding window), su una metodologia di tipo cross validation e tecniche di oversampling e undersampling. Variare il tipo di approccio all'analisi consente di studiare e ricercare diversi tipi di possibili pattern e correlazioni tra lo stile di guida e gli incidenti.

Il file contenente l'intero dataset viene processato in modo tale da separare i dati di ogni singola polizza, creando un file separato per ogni assicurato. Ciò consente di costruire un file dedicato per ogni polizza, che può essere sottoposto a una procedura di windowing che prenderà in analisi solo il numero di settimane di guida presenti nel file della singola polizza e con esse costruirà la successione di "finestre di training" da esaminare. In caso di buchi tra le settimane dovuti al mancato utilizzo del veicolo, essi non saranno colmati con dati fittizi (inserendo quindi settimane con 0 eventi di guida e 0 km percorsi) in grado di alterare l'eventuale correlazione temporale di alcuni comportamenti. Questa scelta consente di considerare le settimane di guida come indipendenti rispetto al periodo dell'anno, ma dipendenti solo dall'utilizzo del veicolo da parte dell'assicurato.

Esempio: In caso di dati disponibili, per una determinata polizza, per le settimane che vanno dalla n.10-2016 alla n.20-2016 e dalla n.30-2016 alla n.40-2016, considereremo la guida come continua, mettendo così in evidenza lo stile di guida, che non cambia in base alla settimana dell'anno, ma è caratteristico del guidatore. In questo caso i dati saranno trattati come se fossero 20 settimane contigue, senza interruzioni nel mezzo.

Gli attributi scelti in fase di windowing rispondono a due tipi di analisi: univariata e multivariata. Nell'analisi univariata, andremo ad effettuare il windowing solo sull'attributo che registra se si è verificato un incidente o meno, senza aggiungere ulteriori dati di polizza. Nell'analisi multivariata invece, andremo ad aggiungere al windowing altri attributi relativi a una determinata settimana (percorrenze, eventi di guida, ecc.) o relativi in generale alla polizza (età dell'assicurato, età del veicolo, ecc.). Per maggiori dettagli sulle tecniche utlizzate, si rimanda il lettore al Capitolo 4.

7.1 Validazione basata su sliding window

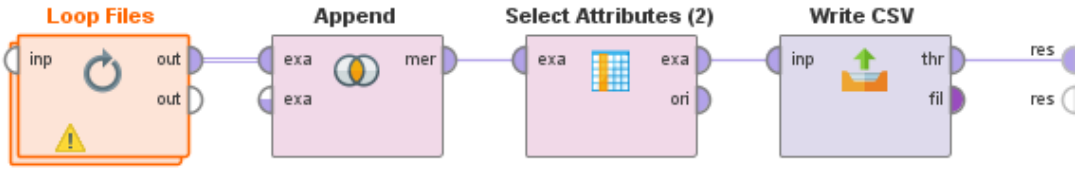

Figura 7.1: Processo di loop nella cartella dei file singoli di ogni polizza

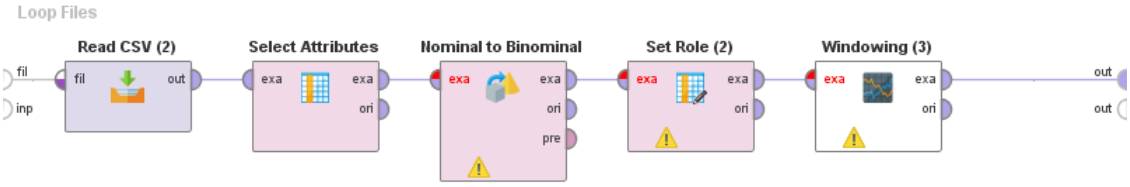

Figura 7.2: Dettaglio del sottoprocesso “Loop Files” di figura 7.1

Successivamente al processing dei singoli file e all’applicazione del windowing (Sez. 4.3.1) su ognuno di essi, tutti i file creati vengono inseriti nuovamente in un file unico, concatenandoli uno dopo l’altro. Lo scopo di questa operazione è quello di costruire un singolo file composto dai comportamenti dei vari assicurati, messi in un ordine temporale coerente. Vengono effettuate due scelte riguardanti i parametri che regolano la costruzione dei file di windowing che quindi portano alla creazione di due diversi file su cui effettuare gli esperimenti successivi. Un primo file è caratterizzato da una finestra di training di 12 settimane, un horizon di 4 settimane, e un test set di 4 settimane, cerca quindi di predire il verificarsi di un incidente nel secondo mese successivo all’ultimo dei 3 mesi di analisi. Nel secondo tipo di analisi scelta, la finestra di training prende in esame solo 4 settimane e tenta di predire un incidente nel secondo mese successivo a quello di analisi (tra il mese di training e quello di test ci sono 4 settimane di horizon). La creazione del modello di predizione viene effettuata tramite i 5 metodi elencati nel Capitolo 3, i file creati al passo precedente vengono quindi usati come input per ognuno dei 5 metodi di classificazione. Il modello generato da ogni metodo viene validato separatamente tramite la tecnica dello Sliding Window Validation. Tramite l’omonimo modulo disponibile su Rapid Miner effettueremo la validazione in due modalità ulteriori, con e senza l’opzione di *cumulative training* attivata; se abilitata, questa opzione permette di aggiungere la vecchia finestra di training alla nuova in esame, invece di sostituirla del tutto.

Usando tutte le combinazioni di metodi di classificazione e parametri di validazione, vengono testati i due file già sottoposti a windowing (con training set a 4 e 12 settimane). La qualità e le performance dei modelli creati vengono valutate tramite la matrice di confusione prodotta da RapidMiner, con particolare attenzione alla classe relativa al verificarsi dell’incidente. Tutti i modelli di classificazione basati su analisi univariata non sono in grado di predire livelli di rischio elevato con sufficiente precisione, procedendo in questo modo si sta cercando una correlazione tra più incidenti effettuati dalla stessa polizza, cosa che risulta abbastanza rara nel nostro dataset e in generale. Per ovviare alla limitata precisione dei modelli

generati sulla classe d'interesse, sono stati valutati modelli basati su analisi multivariati, ovvero la predizione del livello di rischio si basa non solo sui valori di rischio passati ma anche sui valori di variabili potenzialmente correlate (ad es. gli stili di guida, le percorrenze dei veicoli, ecc.), tuttavia i risultati non migliorano se non di pochi punti percentuali. Questo tipo di performance può trovare una spiegazione nella granularità settimanale del dataset che può impedire di mettere in evidenza determinati pattern, magari più evidenti guardando i dati con granularità mensile o annuale.

7.2 Validazione basata su cross validation e windowing

Tramite l'uso della sliding window validation (Sez. 4.3.3), abbiamo provato a cercare eventuali correlazioni tra i comportamenti passati di un certo assicurato e il futuro immediato. Si è cercato di dimostrare se i comportamenti dei mesi precedenti a un incidente potessero evidenziare una condotta comune, confinando l'analisi ad una polizza per volta. Con la Cross Validation (Sez. 4.3.5) invece, la finestra di training non scorre più lungo le settimane e non distingue più le polizze, ma analizza le righe che riportano un incidente nella settimana corrente e cerca eventuali correlazioni tra gli eventi di guida e le percorrenze delle settimane precedenti. Per la cross validation viene usata la modalità K-folds con k=10, il tipo di campionamento dei folds è stratified. Analizzando separatamente le polizze per tipo (NA,NS,VA,VS) risultano dei modelli che predicono il verificarsi di un incidente con scarsa precisione (valore massimo di 2,36%) ma con un richiamo mediamente più alto (valori oscillanti tra 35%con Naive Bayes e 45% con WAODE) rispetto all'analisi effettuata con la sliding window validation.

7.3 Validazione basata su cross validation e oversampling / undersampling

Il terzo tipo di analisi utilizza un approccio diverso dai primi due perché non effettua l'operazione di windowing. Le tecniche utilizzate per la classificazione sono le stesse elencate precedentemente; alla granularità settimanale analizzata nei punti precedenti, viene aggiunta anche l'analisi sul dataset aggregato annualmente, per controllare se e quanto influisca la granularità temporale sulla predizione del livello di rischio. Sono state sperimentate delle tecniche di oversampling e undersampling (che hanno come obiettivo quello di riequilibrare le classi analizzate. Vedi Sez. 4.4.1) che hanno migliorato le capacità del modello.

7.3.1 Risultati ottenuti con dati aggregati su base settimanale

Nell'aggregazione settimanale senza windowing troviamo tutti gli attributi presenti nel dataset originale tranne quelli che indicano le somme spese per gli incidenti e

il numero di incidenti causati poiché costituirebbero un indizio importante per il modello che vuole capire quali polizze sono a rischio incidente e quali no. Vengono separate le polizze per tipo di rinnovo e per tipo di contratto e viene effettuata nuovamente la creazione del modello e la sua validazione, per ogni anno in esame, per ogni tipo di algoritmo di classificazione.

Un problema riscontrato durante l'analisi è stato quello dello sbilanciamento eccessivo delle due classi, i casi che non riportavano nessun incidente, a livello settimanale, erano circa il 99% del dataset; queste condizioni rendono più difficile capire quali sono le caratteristiche che portano ad un incidente poiché i casi da cui imparare sono troppo pochi. Per risolvere tale problema, sono state utilizzate le tecniche di undersampling (Sez. 4.4.1) per equilibrare la classe "Livello di rischio basso" con la classe "Livello di rischio alto" in proporzione 50/50. Il processo (Figura 7.3) viene quindi arricchito di due parti fondamentali, la prima filtra le righe contenenti un incidente e la seconda le accoda ad un numero uguale di casi di non incidente (campionati casualmente), le due parti vengono quindi unite e vanno a formare un unico dataset su cui effettuare gli esperimenti. Per ogni tipo di dataset (VA,VS,NA,NS) viene costruito un dataset più piccolo, di dimensioni pari a due volte i casi di incidente.

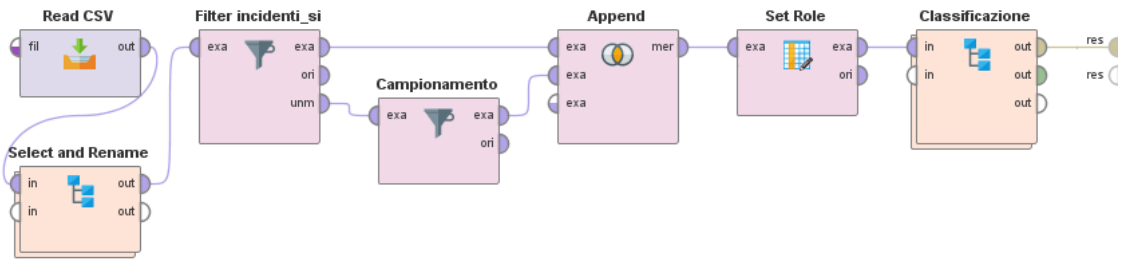

Figura 7.3: Processo di undersampling e analisi

Di seguito vengono elencati tutti i risultati ottenuti, differenziati per tipo di database e per anno. La classe "Livello di rischio alto" è la nostra classe di interesse, ovvero quella che riporta l'avvenuto incidente. L' f-score (nota anche come F_1-score o F-measure, letteralmente "misura F") è una misura dell'accuratezza di un test. La misura tiene in considerazione precisione e richiamo del test, dove la precisione è il numero di veri positivi diviso il numero di tutti i risultati positivi, mentre il richiamo è il numero di veri positivi diviso il numero di tutti i test che sarebbero dovuti risultare positivi (ovvero veri positivi più falsi negativi). L'F1 viene calcolato tramite la media armonica di precisione e recupero. Di seguito vengono riportati i risultati relativi alle analisi effettuate sull'anno 2016.

VS-2016		Rischio alto			Rischio basso	
	Accuracy	**Precision**	**Recall**	**F-score**	**Precision**	**Recall**
NB	52.6%	52.7%	51.4%	52%	52.5%	53.8%
DT	53.2%	54%	43.8%	48,3%	52.7%	62.7%
k-NN	52.3%	54%	31.8%	40,0%	51.6%	72.8%
SVM	52.2%	52.4%	48.7%	50,5%	52%	55.6%
WAODE	54%	54%	52.5%	53,2%	53.8%	55.4%

Tabella 7.1: Polizze Vecchie Semestrali - Anno 2016

VA-2016		Rischio alto			Rischio basso	
	Accuracy	**Precision**	**Recall**	**F-score**	**Precision**	**Recall**
NB	47.5%	45.6%	46.3%	45,9%	49.4%	48.6%
DT	53%	54%	15.8%	24,4%	52.9%	87.6%
k-NN	52%	50.1%	33.5%	40,2%	52.8%	69%
SVM	48.3%	45%	33%	38,1%	50.1%	62.4%
WAODE	51.2%	49%	33.8%	40,0%	52.3%	67.4%

Tabella 7.2: Polizze Vecchie Annuali - Anno 2016

NS-2016		Rischio alto			Rischio basso	
	Accuracy	**Precision**	**Recall**	**F-score**	**Precision**	**Recall**
NB	48.9%	48.8%	46.8%	47,8%	48.9%	51%
DT	51.1%	68.8%	4.2%	7,9%	50.6%	98.1%
k-NN	52.7%	54%	35.4%	42,8%	52%	70%
SVM	45.3%	44.7%	40%	42,2%	45.7%	50.6%
WAODE	46.6%	46%	41%	43,4%	47%	52.1%

Tabella 7.3: Polizze Nuove Semestrali - Anno 2016

NA-2016		Rischio alto			Rischio basso	
	Accuracy	**Precision**	**Recall**	**F-score**	**Precision**	**Recall**
NB	50.6%	50.6%	50.6%	50,6%	50.5%	53.2%
DT	52.1%	66.7%	2.4%	4,6%	51.9%	98.9%
k-NN	49%	45.5%	27.2%	34,0%	50.4%	69.4%
SVM	51.6%	50%	34.3%	40,7%	52.3%	67.8%
WAODE	56.4%	62%	26%	36,6%	55%	85%

Tabella 7.4: Polizze Nuove Annuali - Anno 2016

NA-2015		Rischio alto			Rischio basso	
	Accuracy	**Precision**	**Recall**	**F-score**	**Precision**	**Recall**
NB	56.1%	56.1%	55.6%	55.9%	56%	56.6%
DT	54.2%	58.3%	29.6%	39.3%	52.8%	78.8%
k-NN	51.8%	52.9%	33.3%	40.9%	51.3%	70.4%
SVM	52.2%	54.3%	47.1%	50.4%	53.3%	60.3%
WAODE	62.2%	64%	55.6%	59.5%	60.8%	68.8%

Tabella 7.5: Polizze Nuove Annuali - Anno 2015

NS-2015		Rischio alto			Rischio basso	
	Accuracy	**Precision**	**Recall**	**F-score**	**Precision**	**Recall**
NB	51.3%	51.3%	50.2%	50.8%	51.3%	52.4%
DT	53.5%	78.5%	9.7%	17.3%	51.9%	97.4%
k-NN	57.5%	59.4%	47.1%	52.6%	56.2%	67.9%
SVM	54.4%	55.4%	44.9%	49.6%	53.7%	63.9%
WAODE	57.7%	61.9%	40.1%	48.7%	55.7%	75.3%

Tabella 7.6: Polizze Nuove Semestrali - Anno 2015

VA-2015		Rischio alto			Rischio basso	
	Accuracy	**Precision**	**Recall**	**F-score**	**Precision**	**Recall**
NB	52.1%	52.1%	50.8%	51.4%	52.0%	53.4%
DT	55.7%	56.6%	48.6%	50%	55%	62.7%
k-NN	54.2%	56.1%	38.1%	45.3%	53.1%	70.2%
SVM	52.5%	52.5%	50%	51.2%	52.3%	54.8%
WAODE	57.1%	59%	46.7%	52.1%	55.9%	67.6%

Tabella 7.7: Polizze Vecchie Annuali - Anno 2015

VS-2015		Rischio alto			Rischio basso	
	Accuracy	**Precision**	**Recall**	**F-score**	**Precision**	**Recall**
NB	51.3%	51.3%	50.4%	50.8%	51.3%	52.3%
DT	54.8%	56%	45.4%	49.4%	54%	64.3%
k-NN	53%	55%	33.8%	41.8%	52.2%	72.4%
SVM	51.3%	51.5%	46%	48.5%	51.2%	56.7%
WAODE	56%	56.4%	52.3%	54.2%	55.6%	59.6%

Tabella 7.8: Polizze Vecchie Semestrali - Anno 2015

Come evidenziato dai dati riportati, a parità di metodi utilizzati, le performance differiscono in base anche alla polizza e al tipo di rinnovo.

7.3.2 Risultati ottenuti con dati aggregati su base annuale

Le figure seguenti mostrano le prestazioni medie ottenute dai classificatori mediamente sugli anni 2015 e 2016 nella predizione del livello di rischio più elevato (per tutte le tipologie di livello di rischio). Le barre verticali riportate in ciascun grafico indicano:

- Precisione, richiamo e F-score della classe di rischio alto del classificatore migliore tra quelli testati (Random Forest, appartenente alla classe Decision-Tree)
- Precisione, richiamo e F-score della classe di rischio alto dell'ensemble di classificatori (fino a 7 classificatori combinati insieme)
- Precisione, richiamo e F-score della classe di rischio alto ottenuta dal classificatore migliore (Random Forest) applicando il bagging.

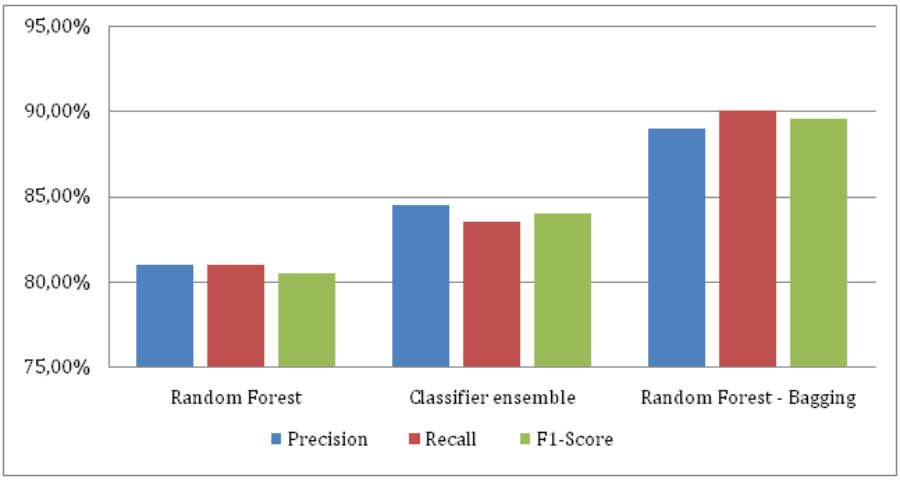

Figura 7.4: Predizione del livello di rischio alto per costo degli incidenti causati. Anni 2015-2016

I risultati mostrano come i modelli costruiti sui dati aggregati su base annuale sono in grado di predire con elevati valori di precisione e richiamo la classe di rischio alta. La tecnica Random Forest, combinata con strategie di bagging e undersampling 50-50, ha ottenuto i risultati migliori sui dati analizzati. Per tutte le tipologie di livello di rischio i valori di precisione e richiamo sono pari o superiori al 90%. I modelli predittivi generati su base annuale sono finalizzati a predire situazioni di rischio nel medio-lungo termine. Al rinnovo di una polizza annuale o alla stipula di una nuova polizza, le informazioni passate relative alla stessa polizza o a polizze simili possono consentire ad un algoritmo di classificazione di identificare in anticipo potenziali situazioni di rischio.

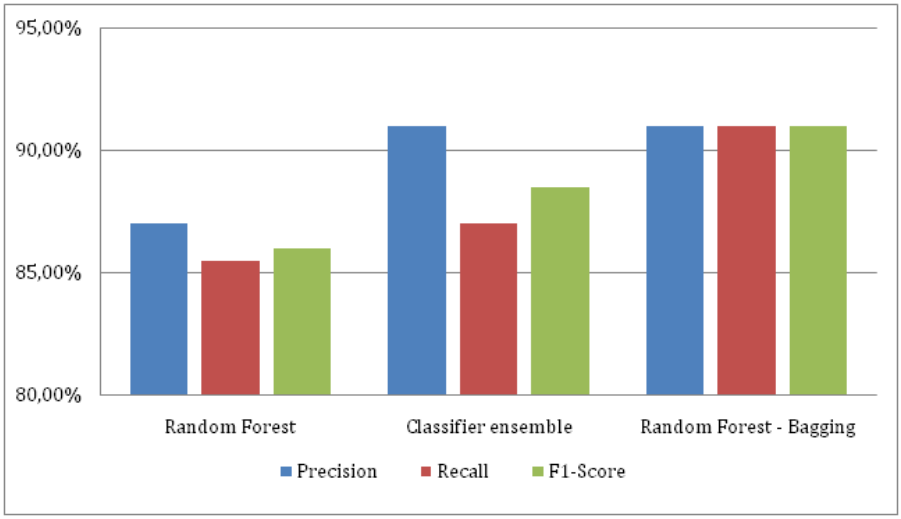

Figura 7.5: Predizione del livello di rischio alto per numero degli incidenti causati. Anni 2015-2016

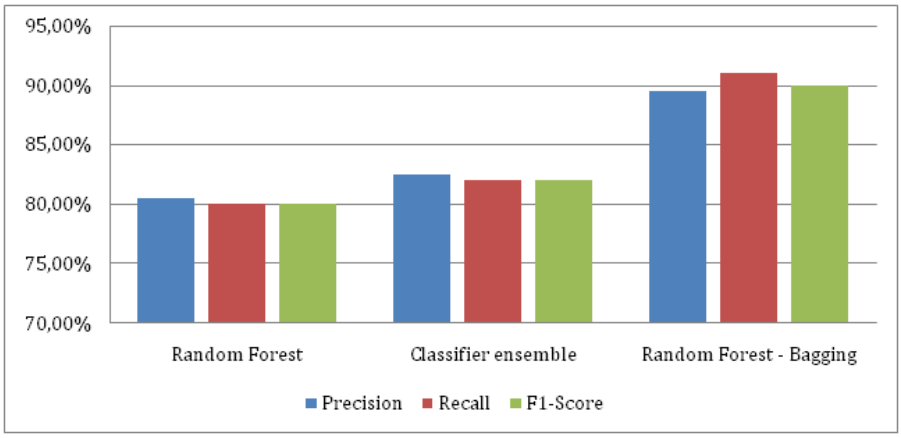

Figura 7.6: Predizione del livello di rischio alto per numero degli incidenti gravi. Anni 2015-2016

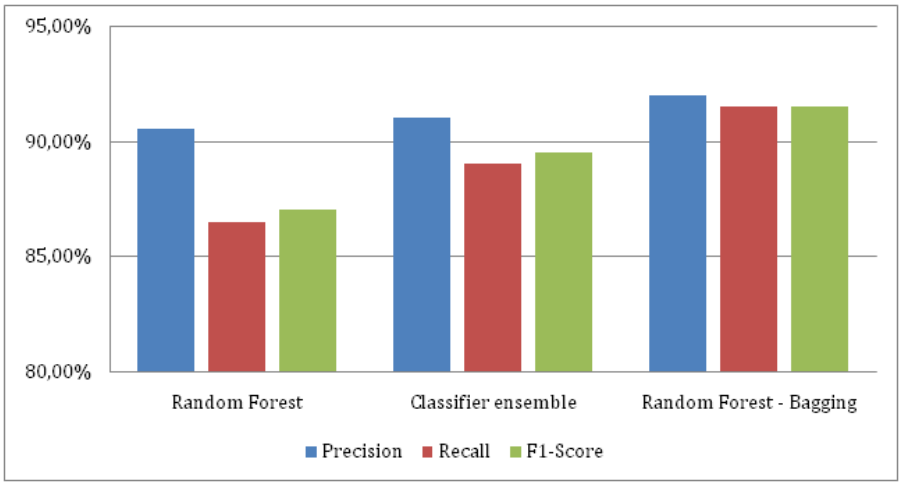

Figura 7.7: Predizione del livello di rischio alto per numero degli incidenti non gravi. Anni 2015-2016

Capitolo 8

Conclusioni e sviluppi futuri

I risultati mostrano come, su base settimanale, la correlazione tra il livello di rischio futuro e i dati recenti raccolti dalla telematica sia debole; nella maggior parte delle prove effettuate, nonostante le ottimizzazioni i classificatori non sono stati in grado di predire con una precisione superiore al 60% un livello di rischio elevato. Inoltre, i valori di richiamo ottenuti indicano come nella maggioranza dei casi non sia stato possibile prevedere un livello di rischio alto, in quanto tale evento risulta debolmente correlato con i dati raccolti dalla telematica e dall'analisi dei sinistri accaduti nel mese precedente.

Una possibile spiegazione di questi risultati può risiedere nel fatto che il database contiene dei dati che non riescono ancora a caratterizzare al meglio il comportamento di un automobilista. Un'ulteriore spiegazione delle scarse performance di classificazione deriva dalla mancanza di un numero adeguato di casi di incidente da cui il modello possa "imparare", infatti con l'aggregazione settimanale il numero di righe che indicano un incidente sono meno dello 0,2% del totale sia nell'anno 2016 che nell'anno 2015 e le tecniche usate di oversampling e undersampling non riescono a risolvere completamente il problema di sbilanciamento delle classi. Per quanto riguarda invece l'aggregazione mensile, il numero di casi di incidente su un totale di 588.457 mesi registrati, è sempre inferiore all' 1%, rendendo altrettanto difficile caratterizzare i comportamenti correlati con un incidente. Nel caso della granularità annuale, le polizze che registrano un incidente risultano essere circa il 5% totale.

Per quanto riguarda gli esperimenti effettuati sul dataset con granularità annuale, i modelli sono molto più performanti e raggiungono anche dei valori di accuracy e precisione maggiori dell'90%. Questi risultati dimostrano che a livello annuale il modello riesce a prevedere con maggiore esattezza il livello di rischio di una polizza e ciò può essere d'aiuto alla compagnia assicurativa che in fase di rinnovo di una determinata polizza può negoziare eventuali condizioni più o meno vantaggiose per l'assicurato basate sulla predizione effettuata dal modello.

8.1 Sviluppi futuri

Il dataset potrebbe integrare alcune informazioni più dettagliate sul luogo dell'incidente. Sbandate, frenate, superamenti di velocità, se non contestualizzati possono perdere il loro valore di indicatori relativi alla pericolosità di guida e all'imprudenza. Osservare il numero assoluto di eventi di guida pericolosi, ma non poterli collocare in una specifica tipologia di strada, non consente di ricercare la correlazione tra un tipo di comportamento mantenuto su un tipo di strada, e i suoi eventuali effetti (positivi o negativi). Nel dataset a nostra disposizione, non abbiamo informazioni sull'incidente: non sappiamo su che tipo di strada è accaduto, non sappiamo la fascia oraria, non sappiamo la provincia. Sicuramente integrare queste informazioni potrebbe tornare utile ai fini della predizione, poiché legherebbero in maniera molto più precisa l'incidente ai dati di percorrenza, di stile di guida e di caratteristiche di polizza. L'implementazione di eventuali integrazioni al dataset originale (per esempio i dati citati sopra, dati meteo, ecc.) potrebbe costituire un aggiornamento di sicura utilità per le future analisi da parte della compagnia assicurativa.

Bibliografia

[1] Estela Bee Dagum. *Analisi delle serie storiche*. Springer, 2001. ISBN 9788847001463.

[2] Bala Deshpande. Using rapidminer for time series forecasting in cost modeling, 2012. URL www.simafore.com/blog/bid/106430. [08/10/2018].

[3] Yung-Ching Hsu, Pai-Lung Chou, and Yung-Ming Shiu. An examination of the relationship between vehicle insurance purchase and the frequency of accidents. *Asia Pacific Management Review*, 21(4):231 – 238, 2016. ISSN 1029-3132. URL http://www.sciencedirect.com/science/article/pii/S1029313216301762.

[4] Sinan Ozdemir. *Data science*. Apogeo, 2017. ISBN 9788850318070.

[5] Johannes Paefgen, Florian Michahelles, and Thorsten Staake. Gps trajectory feature extraction for driver risk profiling. In *Proceedings of the 2011 International Workshop on Trajectory Data Mining and Analysis*, TDMA '11, pages 53–56, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0933-2. DOI 10.1145/2030080.2030091.

[6] Marco Pironti. *Economia e gestione delle imprese e dei sistemi competitivi*. libreriauniversitaria.it, 2012. ISBN 9788862922821.

[7] Alessandro Rezzani. *Big data. Architettura, tecnologie e metodi per l'utilizzo di grandi basi di dati*. Apogeo Education, 2013. ISBN 9788838789892.

[8] Norvig P. Russell S. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 2003. ISBN 9780137903955.

[9] Athanasios Theofilatos, George Yannis, Constantinos Antoniou, Antonis Chaziris, and Dimitris Sermpis. Time series and support vector machines to predict powered-two-wheeler accident risk and accident type propensity: A combined approach. *Journal of Transportation Safety & Security*, 10(5):471–490, 2018. DOI 10.1080/19439962.2017.1301611. URL https://doi.org/10.1080/19439962.2017.1301611.

[10] K. Vassiljeva, A. Tepljakov, E. Petlenkov, and E. Netsajev. Computational intelligence approach for estimation of vehicle insurance risk level. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 4073–4078, May 2017. DOI 10.1109/IJCNN.2017.7966370.

[11] A. Verma, A. Taneja, and A. Arora. Fraud detection and frequent pattern matching in insurance claims using data mining techniques. In *2017 Tenth International Conference on Contemporary Computing (IC3)*, pages 1–7, Aug 2017. DOI 10.1109/IC3.2017.8284299.