UNIVERSITA' POLITECNICA DELLE MARCHE

FACOLTA' DI INGEGNERIA

Corso di Laurea Triennale in Ingegneria Gestionale

Strumenti avanzati per l'analisi di dati guasto di macchine per trazione elettrica

Advanced tools for failure data analysis of electric traction machines

Relatore:

Prof. Maurizio Bevilacqua

Correlatore:

Ing. Sara Antomarioni

Tesi di Laurea di:

Veronica Ranieri

Anno Accademico 2019/2020

“tutto quello che oggi è una realtà,

prima era solo parte di un sogno

impossibile”

W.Blake

INDICE

INTRODUZIONE

Industry 4.0,(Subrizi, 2008), è il termine generalmente utilizzato per identificare la Quarta Rivoluzione Industriale. È fondamentale individuare sia le circostanze storiche che hanno portato all'evoluzione dei sistemi che operano, compongono e facilitano l'attività industriale, sia ciò che compone a livello tangibile questa rivoluzione.

Industry 4.0 è un nuovo paradigma produttivo nato in Germania nel 2011 che ha contribuito ad integrare, all'interno di ambienti produttivi, le tecnologie di tipo informatico ed inoltre rappresenta un tentativo di modellare un ambiente virtuale in modo tale che esso sia riferibile ad un ambiente reale.

L'espressione "Industry 4.0" è stata usata per la prima volta alla Fiera di Hannover nel 2011 in Germania. A ottobre 2012 un gruppo di lavoro dedicato all'Industria 4.0, presieduto da Siegfried Dais della multinazionale di Ingegneria ed elettronica Robert Bosch GmbH e da Henning Kagermann della Acatech (Accademia tedesca delle Scienze e dell'Ingegneria) presentò al governo federale tedesco una serie di raccomandazioni per la sua implementazione. L'8 aprile 2013, all'annuale Fiera di Hannover, fu diffuso il report finale del gruppo di lavoro.

I modelli vincenti dell'industria manifatturiera sono stati generati sulla base di esperienze concrete e di successo in diversi contesti economici come la produzione di massa (Ford) e la Lean Production (Toyota). Tutti questi modelli nascono dall'industria automobilistica che ha rappresentato negli ultimi cento anni e rappresenta ancora oggi, la sfida più complessa per qualsiasi modello organizzativo e produttivo. Infatti l'industria dell'auto è da sempre una industria di grandi volumi, elevati costi del prodotto, grande complessità e varietà dei componenti della distinta base. In poche parole è un'industria con processi, prodotti e supply chain complessi e ad alta variabilità. Di fronte alla crescente introduzione dell'elettronica nei processi produttivi e nei prodotti automobilistici, le grandi imprese dell'auto hanno fatto e sperimentato una ridefinizione complessiva del loro modello di business.

Dieter Zetsche, presidente del Board of management di Daimler e responsabile di Mercedes-Benz Cars: "Tutti i principali trend dell'industria automobilistica sono supportati dalla digitalizzazione o supportano a loro volta quest'ultima. Il nostro obiettivo è anche quello di essere la casa automobilistica leader e più innovativa al

mondo in fatto di tecnologie digitali." Markus Schäfer, membro del Divisional board Mercedes-Benz Cars, manufacturing and supply chain management di Daimler ha precisato: "In Mercedes-Benz, parliamo di 'Industria 4.0' per descrivere la digitalizzazione dell'intera catena del valore: dalla progettazione alla vendita ed assistenza, passando per la produzione (da cui ha origine il termine). Noi di Daimler non abbiamo dubbi sul fatto che la rivoluzione digitale cambierà radicalmente la nostra industria. Queste novità si applicano ai metodi con cui sviluppiamo, progettiamo e produciamo le nostre vetture. Ma non solo: riguarderanno, infatti, anche il modo in cui entriamo in contatto con i clienti risultando inoltre evidenti nei nostri prodotti".

Le dichiarazioni e i comportamenti dei manager delle altre industrie tedesche dell'auto (Wv Group e BMW) sono sostanzialmente le stesse come quelle di tutti i competitori a livello mondiale. L'industria dell'auto e soprattutto l'industria dell'auto tedesca è quindi il punto di partenza del paradigma Industry 4.0 che s'intende come la digitalizzazione dell'intera del valore. Questo è il punto di vista dell'industria dell'auto tedesca che si fonda su strutture organizzative e gestionali molto avanzate: implementazione diffusa di sistemi ERP e di Lean Production, Supply chain integrata con fornitori e clienti, automazione di fabbrica etc.

Il termine Industria 4.0 (o Industry 4.0) indica una tendenza dell'automazione industriale che integra alcune nuove tecnologie produttive per migliorare le condizioni di lavoro e aumentare la produttività e la qualità produttiva degli impianti.

Figura 1: I 9 pilastri della Industry 4.0

Nella Figura 1 sono riportate le nove tendenze tecnologiche che costituiscono gli elementi di cambiamento principali di Industry 4.0.

1. Simulazione: già in uso nei processi di progettazione, l'utilizzo di sistemi simulativi verrà esteso a tutti i processi produttivi. Questi sistemi elaboreranno i dati raccolti in tempo reale in modelli simulativi virtuali al fine di testare e ottimizzare macchine, prodotti e processi e di anticipare problemi prima che questi avvengano nella realtà;
2. Big data: raccolta e analisi di un grande numero di dati provenienti da diverse fonti a supporto dei processi decisionali;
3. Sistemi di Additive Manifacturing: tecnologie utilizzate in modo più ampio per produrre piccoli lotti di prodotti altamente customizzati, ed essendo realizzabili in più centri dislocati sul territorio, permetteranno di ridurre le distanze per il trasporto logistico dei prodotti finiti.
4. Robot autonomi: la nuova generazione di robot avrà un costo più basso e maggiori capacità rispetto a quelli attualmente in uso; inoltre saranno in grado di interagire tra loro e di apprendere da queste interazioni;
5. Strumenti di Augmented reality: sistemi che, attraverso un dispositivo mobile, aggiungono informazioni multimediali alla realtà già normalmente percepita dall'uomo. In futuro queste tecnologie verranno utilizzate per fornire informazioni in tempo reale utili per migliorare i processi lavorativi ed il decision making;

Vediamo passo per passo l'analisi sviluppata nei tre capitoli.

Il primo capitolo dell'elaborato analizza gli studi riguardanti i big data effettuati da studiosi, in ambito ferroviario, attraverso tecniche di data mining.

Il secondo capitolo tratta il data mining in linea generale, entrando poi nel dettaglio nelle tecniche di classificazione; vengono quindi approfonditi i principali metodi utilizzati per l'estrazione degli itemset frequenti, basati sull' algoritmo Apriori e sull'algoritmo FPGrowth.

Il terzo ed ultimo capitolo illustra l'analisi di un dataset, contenente dati di guasto di una locomotiva elettrica, mediante l'utilizzo delle Association Rules; inizialmente è stato necessario studiare il metodo e il software da adottare per l'analisi del dataset.

CAPITOLO 1

L'ANALISI DELLA LETTERATURA

In questo capitolo viene effettuata l'analisi della letteratura, ovvero vengono presentati gli studi, in campo ferroviario, svolti da altri ricercatori mediante l'utilizzo delle regole di associazione

La ferrovia, (*Qual è Il Modo Più Sicuro Di Viaggiare?*, 2020) è storicamente e statisticamente un mezzo di trasporto sicuro, in confronto agli altri sistemi di trasporto ed in particolare in confronto alla strada. Infatti secondo i dati dell'Istat e dell'Automobile club d'Italia (Aci) nel 2018 ci sono stati 172.344 e hanno causato 3.325 vittime tra cui 1.420 conducenti o passeggeri di autovetture, 685 motociclisti e 609 pedoni. In base all'ultimo "Rapporto annuale sulla sicurezza delle ferrovie" dell'Agenzia nazionale per la sicurezza delle ferrovie (Ansf), pubblicato nel 2019, i decessi in incidenti ferroviari nel 2018 sono stati 73. Nel 2017, invece, erano stati 55. Nel 2018 gli incidenti significativi, ossia quelli che hanno causato almeno un decesso o un ferito grave o danni significativi all'ambiente, sono stati 109, in aumento rispetto ai 104 del 2017. Più del 70% degli incidenti e più della metà delle vittime, tra cui morti e feriti gravi, sono causati dalla presenza di pedoni sui binari.

Gli incidenti ferroviari, (*UN MODELLO PER L'ANALISI DELLE CAUSE DEGLI INCIDENTI FERROVIARI Premessa*) però non sono dovuti solo dai fattori spiegati precedentemente ma anche da malfunzionamenti e guasti; in questi ultimi casi è stata utilizzata la tecnica di analisi ad Albero delle Cause (Root Cause Analysis-RCT) che consente di rappresentare le catene più o meno complesse di eventi che, a partire dal guasto del singolo componente, o del fallimento della procedura operativa, evolvono verso la realizzazione di un incidente. Per effettuare l'analisi del guasto, innanzitutto è necessario avere una conoscenza completa ed approfondita del sistema, della sua struttura, del suo funzionamento, delle sue prestazioni e dei pericoli che riguarda tutti i suoi componenti e tutte le modalità di funzionamento. Inoltre occorre evidenziare, in riferimento a ciascun componente, le cause prime che generano l'incidente e approfondire l'analisi al livello immediatamente inferiore ricercando il complesso di cause generatrici delle cause prime e ripetere il procedimento nei livelli successivi fino ad arrivare ad un livello in cui l'analisi non è più sviluppabile.

Negli ultimi decenni ci sono stati miglioramenti riguardanti la riduzione del numero di incidenti HRGC. L'analisi del database statunitense (Ghomi et al., 2017) mostra che il numero di incidenti e il numero di feriti negli HRGC sono diminuiti, ma il rapporto tra la gravità degli incidenti e il numero totale di incidenti è aumentato. La maggior parte degli studi sull'analisi della gravità degli incidenti dei conducenti sono relativi a tratti autostradali o incroci e non a HRGC. Ad incidere sulla gravità degli incidenti ci sono fattori psicologici; riguardo ciò sono stati effettuati degli studi da Saccomanno et al. (2014), applicando il modello binomiale negativo al database degli incidenti canadesi, i risultati mostrano che il numero di binari, l'angolo di incrocio, la velocità massima del treno e il numero di persone coinvolte nell'incidente sono fattori significativi. Ad aver esaminato i livelli di gravità di ogni individuo coinvolto in un incidente HRGC sono stati Miranda-Moreno et al. (2015) che ha utilizzato una struttura sistematica bayesiana ed in collaborazione con altri ha analizzato il database degli incidenti HRGC canadese in cui si sono verificati 941 incidenti tra il 1997 e il 2004. Secondo quanto riportato dagli studi è evidente che gli incidenti erano dovuti alla velocità del treno.

A differenza di quanto espresso precedentemente, dallo studio effettuato da Eluru et al. (2015), mediante lo sviluppo di un modello a classe latente per identificare i fattori di gravità degli infortuni del conducente presso HRGC e utilizzando il database della Federal Railroad Administration degli Stati Uniti, degli incidenti avvenuti dal 1997 al 2006 è emerso che a causare gli incidenti sono stati la presenza di neve o pioggia, l'età del conducente, il ruolo del veicolo nell'incidente e i movimenti degli automobilisti. Uno studio analogo è stato condotto da Hao e Daniel utilizzando invece il modello Ordered Probit (Probit Ordinato), dal quale è venuto fuori che i fattori significativi sono stati l'ora di punta, le condizioni meteorologiche, la visibilità, il tipo di veicolo stradale, il traffico medio annuo giornaliero e la velocità del treno. Gli stessi fattori sono stati rilevanti anche negli studi recenti svolti da Fan mediante il Multinomial Logit e l'Ordered Logit Model. Per individuare le cause che hanno causato incidenti è stato necessario analizzare il database per il periodo 2006-2013 considera 13.865 incidenti HRGD negli Stati Uniti e dopo la pulizia del database ne sono rimasti 10.882; la pulizia del database è necessaria per ottenere un'analisi più accurata. Per l'analisi sono stati considerati tanti fattori come la velocità del treno, l'ora dell'incidente, il tipo di veicolo, il traffico medio annuo giornaliero, l'età e il sesso del conducente, l'illuminazione e le condizioni meteorologiche, la visibilità il tipo di incidente, la posizione del veicolo e l'angolo di incrocio. Dopo la

pulizia del database è necessario sviluppare la matrice di correlazione, la quale ottiene dei risultati che mostrano un'elevata correlazione tra la velocità del treno e il traffico giornaliero medio annuo, e anche tra visibilità e condizioni meteorologiche. Per identificare i principali fattori associati alla gravità delle lesioni dei conducenti di veicoli stradali sono stati utilizzati, oltre al modello Ordered Probit, altri due metodi: CART (Classification and Regression Tree, Albero di classificazione e regressione) e le Regole di Associazione.

Poiché la gravità della lesione del conducente è una variabile discreta è necessario ricorrere al Multinomial Logit Model e Ordered Probit Model; in questo caso viene adoperato il modello Ordered Probit. Gli algoritmi basati sull'albero di decisione sono comuni applicazioni di data mining per trovare regole di previsione se la variabile dipendente è una variabile qualitativa, l'albero decisionale sarà chiamato albero di regressione. In questo studio la variabile di risposta è una variabile qualitativa quindi viene prodotto un albero di classificazione. I dati entrano come input in un nodo per essere esaminati e, in base a ciò che ottengo come risultato, i dati vengono indirizzati a uno dei due rami inferiori; inoltre in base alla risposta alla domanda presente nel nodo vengono assegnate le Regole di Associazione al ramo secondario. Quest'ultime utilizzano istruzioni if/then che evidenziano le relazioni in un database e si concentrano sulla valutazione delle caratteristiche che si verificano simultaneamente; pertanto è possibile ridurre la gravità degli incidenti modificando le condizioni. Dalle regole di associazione ottengo come output le istruzioni if/then in cui "if" include le condizioni dell'incidente e "then" contiene la conseguenza dell'incidente; in questo studio è stato utilizzato l'algoritmo Apriori che conta le condizioni per generare una nuova regola e si avvale di misure statistiche, supporto e confidenza.

Le tecniche di data mining hanno consentito di ottenere i seguenti risultati: in particolare le regole di associazione hanno mostrato che l'illuminazione non ha influenzato la gravità degli incidenti quando la velocità del treno è superiore a 42 miglia orarie (mph) e che all'aumentare della velocità del treno, la percentuale di incidenti mortali aumenta di oltre il 10%. L'algoritmo dell'albero di classificazione, inoltre ha mostrato che quando la velocità del treno è superiore a 42 mph ed è il treno a colpire i veicoli stradali si verifica una conseguenza più grave rispetto a quando un'auto colpisce un treno. Riguardo quanto affermato, l'algoritmo Apriori ha confermato l'importanza del tipo di incidente; infatti per un veicolo fermo è più probabile che non ci siano lesioni, mentre per un veicolo

stradale in movimento la probabilità di lesioni è maggiore. Allo stesso modo, se la velocità del treno è superiore a 42 mph quando colpisce un veicolo stradale, i conducenti che non superano l'incrocio hanno maggiori probabilità di non subire lesioni rispetto a coloro che cercano di muoversi e superare l'incrocio. La percentuale di mortalità è raddoppiata per i veicoli pesanti che attraversano un incrocio quando la velocità del treno supera i 42 mph, ed oltre a ciò, l'albero di classificazione ha mostrato una maggiore gravità degli incidenti mortali per i conducenti anziani quando un treno ha colpito un veicolo in movimento ad alta velocità, così come anche le regole di associazione mostrano che esiste una forte relazione tra la mortalità e i conducenti maschi anziani quando un treno ad alta velocità colpisce il loro veicolo. Significativo è anche il sesso del conducente del veicolo in quanto la percentuale di uomini di età inferiore ai 66 anni è quattro volte superiore rispetto alle donne, ma le donne subiscono gravi conseguenze. Dall'analisi dei risultati si evince che quando il treno ha una velocità inferiore a 28 mph, l'illuminazione potrebbe ridurne la gravità e questo è previsto perché all'aumentare della velocità del treno, diminuisce il tempo di reazione dei conducenti. La presenza di neve o pioggia, invece riduce la probabilità di lesioni gravi in quanto i conducenti sono più prudenti e guidano lentamente, ma ad affermare il contrario sono stati Eluru et al. (2015) e Zhang et al. (2011). Inoltre è evidente che i conducenti che si muovono su un incrocio subiscono più lesioni rispetto ai conducenti che vengono fermati sui binari, in quanto i conducenti che si sono fermati all'incrocio hanno l'opportunità di lasciare l'auto prima dell'impatto e ridurre così il rischio di lesioni. Un altro fattore significativo è il tipo di veicolo; in entrambi gli algoritmi utilizzati si deduce che i conducenti di veicoli stradali leggeri subiscono gravi lesioni rispetto a quelli che guidano veicoli pesanti, essendo che i veicoli pesanti hanno maggiore resistenza e anche maggiori dimensioni, quindi è naturale che in un incidente il conducente di un veicolo pesante possa sfuggire a gravi lesioni. I risultati degli algoritmi mostrano che quando la velocità del treno supera i 28 mph, le donne subiscono lesioni più gravi rispetto agli uomini; questo perché fisiologicamente gli uomini sono più forti. Dallo studio è emerso che durante la notte, a causa della mancanza di visibilità, si verificano lesioni gravi ed il minor traffico sulle strade porta i conducenti a viaggiare a velocità più elevate. Un aspetto positivo delle tecniche di data mining è la loro capacità di rilevare l'interazione tra i vari fattori che causavano l'incidente.

Un altro lavoro di ricerca per comprendere l'intensità dei guasti del sistema di aereo (OCS), utilizzato per trasmettere energia elettrica a tram, filobus e treni ad una certa distanza dal punto di alimentazione, è stato eseguito nel Nord-Ovest della Cina da gennaio 2016 a maggio 2018 e dai risultati ottenuti sono stati forniti suggerimenti dettagliati per guidare il funzionamento e la manutenzione di OCS. Estraendo gli itemset frequenti che soddisfano la soglia impostata nel database, vengono mostrate le regole di associazione; pertanto è stato applicato per la prima volta l'analisi delle associazioni a OCS, (Qian et al., 2019).

Il database dei guasti OCS ha le seguenti caratteristiche:

1) Scarsità, poiché l'OCS è un sistema con un'elevata affidabilità, è raro che due o più guasti si verifichino contemporaneamente su un pilastro, che porta alla maggior parte delle transazioni dei dati nel formato orizzontale che ha un solo elemento e questo crea difficoltà per l'ulteriore lavoro di estrazione.
2) Complessità, il layout ferroviario è un sistema di rete complesso in quanto ci sono differenze tra linee o sezioni.
3) Gerarchia, l'OSC è un sistema meccanico che ha tanti tipi di possibili guasti (più di 2000 guasti specifici) e alcuni di essi non sono frequenti ma possono essere considerati come un tipo di guasto e diventano frequenti nel loro insieme. Per evitare di perdere questo tipo di informazioni è opportuno analizzare le regole di associazione basate sulla struttura gerarchica dei guasti.

Per risolvere i problemi verificatisi è stato adoperato un modello di analisi gerarchica basato sull'intensità dei guasti (FHI) per il database dei guasti minerari della stazione OCS. Questo modello è composto da tre parti: metodo di partizione multidimensionale per classificare le voci di guasti e si trasforma in un formato dati orizzontale; metodo del fattore di influenza per ottenere il calcolo dell'intensità del guasto per eliminare l'influenza della diversa divisione delle voci dei guasti e la strategia di taglio per estrarre le regole di associazione a livello trasversale per la proprietà della struttura gerarchica dei guasti.

Il database dei guasti della stazione OCS memorizza tutte le informazioni sui guasti raccolte dalle apparecchiature di rilevamento o dai test manuali. Nel database dei guasti dell'OCS ferroviario, ogni episodio di guasto è un elemento nel database dei guasti contenente il codice dell'articolo, la posizione dove si è verificato l'errore e il tempo di

rilevamento. Tutti i possibili guasti specifici sono posti al livello inferiore di una struttura gerarchica a tre livelli: livello di tipo, livello di apparecchiatura e livello di attributo. La frequenza di ogni guasto di livello superiore deve essere definita ed il valore è uguale alla somma della frequenza di tutti i suoi guasti di livello inferiore. È stata eseguita un'analisi passo passo per sviluppare un modello FHI per l'estrazione di regole di associazione a livello incrociato nel database dei guasti del sistema ferroviario OCS basato sull'intensità del guasto. In generale, il database originale può essere a malapena disponibile per l'estrazione delle regole di associazione; per questo motivo è necessaria la pre-elaborazione dei dati da modellare nel database delle transazioni. Tuttavia, il metodo di partizione di base basato sull'unità più piccola non è più applicabile ai dati di guasto del sistema ferroviario OCS a causa della scarsità che causa troppe transazioni con un singolo articolo. Per raggruppare gli elementi viene proposto un metodo Multi-Dimensional Partition (MDP), ovvero un metodo di partizione basato su informazioni multidimensionali. Quando si stabilisce il database delle transazioni di guasto, bisogna assicurarsi che la partizione degli errori sia ragionevole, quindi tutte le voci in una transazione devono condividere la stessa connessione logica o fisica; ad esempio i guasti che si sono verificati in un determinato tempo o spazio possono essere inseriti in una transazione considerando che l'ambiente e la frequenza di funzionamento sono simili. Tutte le voci di guasto possono essere inserite in un sistema di coordinate rettangolari costituito dall' asse del tempo e l'asse dello spazio e ogni nodo rappresenta una transazione; in particolare nel metodo MDP, l'asse del tempo e dello spazio può essere suddiviso in intervalli finiti rispettivamente dalla scelta della scala. La scala dello spazio può essere scelta tra il livello di ufficio, il livello di linea, il livello di campo di battaglia e il livello di pilastro, nel frattempo la scala del tempo può essere scelta tra giorno, mese, trimestre e anno e questo fornisce scelte flessibili. Il database delle transazioni può essere stabilito dal database originale dei guasti con il metodo MDP che utilizza la matrice chiamata "Binary Frequency Matrix" per stabilire transazioni e memorizzare i dati. Però il metodo MDP porta a partizioni disuguali per garantire la correlazione tra le voci di guasto in una transazione e questo può interrompere i risultati dell'estrazione e interferire con la strategia di manutenzione; ad esempio un errore raro può essere scambiato per frequente quando poche transazioni hanno frequenze elevate causate solo da partizioni disuguali. Di conseguenza è necessario sostituire la frequenza con un indicatore di supporto più ragionevole e l'indicatore importante per giudicare lo stato dell'OCS è l'intensità del guasto. Quest'ultimo, in un dato intervallo di tempo e spazio, è

il rapporto tra la frequenza del guasto (F) e il tempo impostato (T) e la distanza impostata (L). la confidenza di due insiemi di elementi frequenti può essere ottenuta dalla loro intensità di errore.

La proprietà Apriori è applicabile nella struttura gerarchica ma esiste una proprietà simile nella definizione di frequenza di guasto di livello superiore che prevede due regole:

1. Se un elemento di livello superiore non è frequente, tutti gli elementi di livello secondario non sono frequenti;
2. Se un elemento di livello secondario è frequente, tutti gli elementi di livello principale sono frequenti.

Queste regole possono essere estese anche ai set di elementi; ad esempio, per qualsiasi itemset frequente, quando il nuovo itemset è ancora frequente dopo aver raggiunto un item, il vecchio itemset che aggiunge qualsiasi item di livello superiore sarà frequente. Al contrario, quando il nuovo itemset non è frequente, il vecchio itemset che aggiunge un item di livello inferiore non sarà mai frequente. Se ogni item nell'itemset A può corrispondere uno ad uno nell'itemset B ed è esso stesso un elemento di livello superiore, l'itemset A appartiene all' itemset B. Viene proposta una nuova strategia di potatura nel processo di generazione di itemset frequenti dall'itemset candidato e prevede che tutti gli itemset candidati sono disposti in ordine decrescente in base alla lunghezza totale di tutti gli elementi codificati.

Nel caso studiato è stato adottato il metodo MDP per ridurre la scarsità dei dati; la scarsità è un indicatore di pochi dati, la sua descrizione su una matrice è la percentuale di numeri diversi da zero e quando quest'ultima è inferiore o uguale al 5% si parla di matrice sparsa. Inoltre, le transazioni con un singolo item interessano solo itemset frequenti, ma hanno a malapena effetto sull'estrazione delle regole di associazione; pertanto, la proporzione di transazioni con un singolo item non può essere troppo grande. Gli indicatori scelti per valutare le prestazioni del metodo MDP sono stati la scarsità e la percentuale di transazioni con un solo item, indicato come T1%. Come già accennato in precedenza, nella creazione del database delle transazioni, il metodo MDP può generare la matrice delle transazioni in modo flessibile in base alla scelta della scala.

La Tabella 1 mostra i risultati degli indicatori sotto diverse scale:

Scarsità T1%	Giorno	Mese	Trimestre	Anno
Livello di ufficio	1.62 %	9.52 %	20.52 %	47.61 %
	35.50 %	13.79 %	0.00 %	0.00 %
Livello di linea	0.88 %	2.71 %	4.71 %	9.24 %
	58.96 %	27.61 %	17.72 %	6.25 %
Livello di campo di battaglia	0.55 %	0.77 %	1.04 %	1.65 %
	70.38 %	56.38 %	41.05 %	24.76 %
Livello di pilastro	0.24 %	0.25 %	0.25 %	0.25 %
	96.80 %	96.14 %	95.70 %	94.06 %

Tabella 1: Due indicatori di diversa scelta di scala (fonte FHI: A Fault Intensity-based Hierarchical Association Analysis Modelfor Mining Fault Database of Railway OCS (2019))

Secondo i risultati della Tabella 1, la scarsità dei dati e la percentuale di transazioni con un singolo item stanno diminuendo con l'aumento della scala nella dimensione temporale o spaziale. La scelta della scala è flessibile, ma non è che maggiore è la scala e migliori sono i risultati dell'estrazione; di conseguenza è necessario non solo ridurre la scarsità dei dati ma anche garantire che siano fornite sufficienti transazioni per l'estrazione dei dati. Confrontando i risultati nella Tabella 1 e considerando il numero di transazioni si evince che "la scala di Anno/Livello di linea" è la scelta più adatta. Dopo aver scelto la scala per dividere le voci nel database dei guasti dell'OCS ferroviario, bisogna impostare il valore del tempo e della distanza per il calcolo dell'intensità del guasto in base alla domanda obiettivo prima dell'estrazione delle regole di associazione dall'algoritmo. Di conseguenza si ottiene la matrice del fattore di aggiustamento. Per eseguire l'estrazione delle regole di associazione a livello incrociato mediante algoritmo, sono necessarie le soglie di intensità e affidabilità del guasto. Tutte le regole di associazione ottenute e gli insiemi di e gli itemset frequenti sono presenti nella Tabella 1.1 e nella Tabella 1.2:

Association rule	Confidence	Level
({'1310'}) => ({'12'})	99.7%	Cross
({'1201'}) => ({'13'})	99.7%	Cross
({'120607', '1201'}) => ({'13'})	99.6%	Cross
({'120607'}) => ({'13'})	99.4%	Cross
({'30', '13'}) => ({'12'})	96.7%	1
({'120607', '13'}) => ({'1201'})	95.5%	Cross
({'120607'}) => ({'1201'})	95.3%	Cross
({'120607'}) => ({'1201', '13'})	94.9%	Cross
({'30', '12'}) => ({'13'})	86.5%	Cross

Tabella 1.1: I risultati delle regole di associazione (fonte FHI: A Fault Intensity-based Hierarchical Association Analysis Modelfor Mining Fault Database of Railway OCS (2019))

Itemset	Fault intensity	Level	Node
({'12'})	89.50	1	A
({'13'})	63.34	1	B
({'1201'})	20.50	2	C
({'13', '1201'})	20.44	Cross	D
({'120607'})	15.34	3	E
({'13', '120607'})	15.25	Cross	F
({'1201', '120607'})	14.63	Cross	G
({'30', '12'})	13.88	1	H
({'1310'})	12.47	2	I
({'30', '13'})	12.41	2	J

Tabella 1.2: Gli itemset frequenti interessati (fonte FHI: A Fault Intensity-based Hierarchical Association Analysis Modelfor Mining Fault Database of Railway OCS (2019))

Le tabelle ci permettono di ricavare l'intensità, la frequenza dei guasti e le regole di associazione forti tra i guasti frequenti; tutto ciò ci permette di individuare rapidamente la combinazione dei guasti con una maggiore intensità di guasto e rafforzare la sua manutenzione nel futuro funzionamento in ordine decrescente di intensità del guasto. Tuttavia, la connessione tra i vari guasti non si ottiene dalla tabella ma viene costruita la rete dei guasti frequenti sulla base delle regole di associazione mostrate nella Tabella 1.1 che diventerà più complessa al diminuire della soglia. Tutti i nodi di itemset e i corrispondenti guasti coinvolti sono presenti nella Tabella 1.3.

Node	Fault Corresponding
A	Support device
B	Registration device
C	Bracket for top tube
D	Registration device & Bracket for top tube
E	Beta-split pin of bracket for strut
F	Registration device & Beta-split pin of bracket for strut
G	Bracket for top tube & Beta-split pin of bracket for strut
H	Support device & External environment impact
I	Steady arm
J	Registration device & External environment impact

Tabella 1.3: I guasti e le informazioni relative ai codici (fonte FHI: A Fault Intensity-based Hierarchical Association Analysis Modelfor Mining Fault Database of Railway OCS (2019))

Con la rete di connessione dei guasti costruita dalle regole di associazione e dalle relazioni gerarchiche, si può facilmente risalire al guasto precedente o analizzare il guasto successivo che ha un'elevata possibilità di verificarsi.

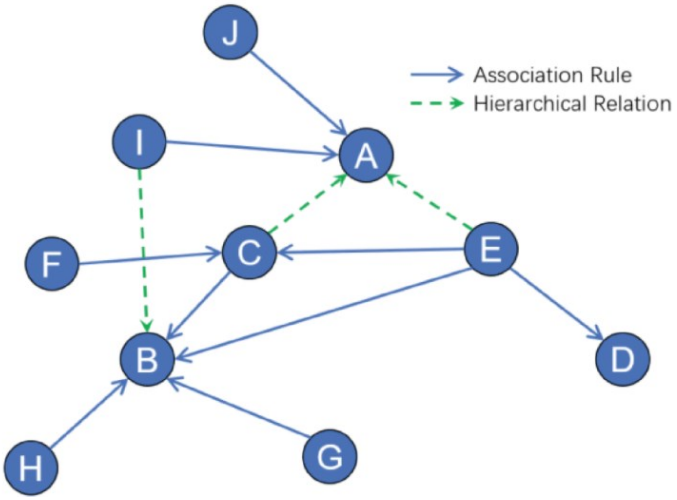

Figura 2: Rete di connessione delle regole di associazione

Per qualsiasi guasto rilevato, presente nella Figura 1, il nodo corrispondente dovrebbe essere selezionato nella rete di connessione, tutti i nodi che vengono indicati dalle regole di associazione dovrebbero essere controllati o applicare la manutenzione preventiva in ordine decrescente di fiducia. Successivamente tutti i nodi precedenti basati sulle regole di associazione dovrebbero essere controllati nello stesso ordine e poi, i nodi principali

e i nodi di livello inferiore in base alle relazioni gerarchiche dovrebbero avere la stessa operazione. Per effettuare la manutenzione preventiva si inizia dall'errore più frequente e poi vengono controllati solo i nodi precedenti in base ai due tipi di frecce fino a quando tutti i nodi non sono stati controllati o implementati. Non appena viene rilevato un guasto è necessario seguire l'ordine per poter intervenire subito. L'analisi di affidabilità, il metodo principale per studiare i dati di guasto OCS si basa sulla probabilità di accadimento di tutti i nodi terminali per trarne la probabilità di accadimento del guasto. In generale, l'albero dei guasti può contenere solo una parte dei componenti di OCS a causa della complessità, compresi i diversi tipi di guasto; quindi i risultati sono limitati ed influenzati dalla creazione dell'albero dei guasti.

Il modello FHI può stabilire in maniera flessibile il database delle transazioni dal database dei guasti dell'OCS e rivelare le connessioni interne del guasto in base alla sua intensità che ha un significato importante per la valutazione dei guasti. La rete di collegamento è costruita da regole di associazione che consentono di guidare il funzionamento e la manutenzione degli OCS ferroviari; inoltre consente di pianificare la manutenzione programmata e la riparazione di emergenza per la riparazione preventiva in anticipo. Tutto ciò rende la manutenzione dell'OCS più ragionevole ed economica.

In questo studio è stato applicato per la prima volta l'estrazione delle regole di associazione all'OCS ed esplorato un modello di analisi FHI per rivelare i collegamenti interni dei guasti dal punto di vista del sistema. Il modello di analisi utilizzato non richiede la conoscenza delle relazioni fisiche e meccaniche per stabilire l'albero dei guasti necessario per l'analisi di affidabilità. A causa della scarsità dei dati di errore dell'OCS, il metodo MDP viene proposto per ridurre l'impatto e fornire più scelte per la creazione del database delle transazioni ed inoltre viene proposto un algoritmo basato sull'intensità dei guasti per eliminare l'influenza di una divisione disuguale causata dal metodo MDP per l'estrazione delle regole di associazione cross level dal database gerarchico dei guasti. Viene condotto un caso studio sulla base del database dei guasti reali degli OCS ferroviari che verifica la validità di questo modello di analisi. In base ai risultati ottenuti dall'estrazione, è stata costruita una rete di connessione e fornito esempi per mostrare come guidare il funzionamento e la manutenzione degli OCS ferroviari.

Un'ulteriore studio sugli incidenti nelle Ferrovie Iraniane (RAI) è stato condotto da Mirabadi & Sharifian (2010), mediante l'applicazione delle regole di associazione, una

tecnica dell'estrazione dei dati. Il ricercatore P. Giudici (2003) definisce questa tecnica, chiamata anche data mining, come il processo di selezione, esplorazione e modellazione di grandi quantità di dati per scoprire le relazioni inizialmente sconosciute con l'obiettivo di ottenere risultati chiari e utili per il proprietario del database.

Oggigiorno il treno è uno dei mezzi di trasporto più utilizzato in quanto riduce l'inquinamento e permette di evitare problemi legati al traffico; pertanto, le autorità competenti dovrebbero lavorare al fine di migliorare il livello di sicurezza e ridurre i fattori che causano incidenti. Per poter sviluppare accurati sistemi di sicurezza, è necessario avere una buona conoscenza dei dati contenuti nei database relativi ai precedenti incidenti.

Molto poco si conosce riguardo l'utilità del data mining in ambito ferroviario, sebbene questa sia applicata frequentemente nell'analisi di incidenti stradali. Uno dei principali motivi per cui questo avviene è il numero limitato di incidenti che si verificano sulle reti ferroviarie rispetto a quelli stradali.

Negli ultimi anni stanno aumentando notevolmente i trasporti ferroviari e per far fronte alla domanda, le autorità competenti stanno pianificando procedure software e hardware. Queste ultime creano maggiore mobilità, dalla quale scaturiscono maggiori incidenti.

Questa indagine si basa sul modello CRISP-DM, costituito da sei fasi e le relazioni tra queste. Il risultato di ciascuna fase determina la fase da eseguire successivamente.

Figura 1.1: Fasi del modello di riferimento CRISP-DM (fonte: https://www.proglobalbusinesssolutions.com/six-steps-in-crisp-dm-the-standard-data-mining-process/ (2021))

Nella prima fase, Business Understanding (Comprensione Aziendale), viene effettuata un'analisi degli incidenti in RAI dal Dipartimento per la Sicurezza stradale.

Nella seconda fase, Data Understanding (Comprensione dei dati), un ruolo fondamentale è svolto dalla raccolta dei dati, dalla codificazione e dalla verifica qualitativa. I dati presi in considerazione per questo studio sono quelli archiviati nel database degli incidenti RAI e suddivisi in due parti: una che comprende dati dal 1996 al 2005 e un'altra contenente dati dal 2006 in poi. Questi gruppi utilizzano linguaggi software differenti e dati appartenenti a diversi ambiti di informazione. Per l'indagine sono stati scelti i dati appartenenti alla prima parte in quanto sono più numerosi, circa 6500.

Nella terza fase, Data Preparation (Preparazione dei dati), si cerca di ottenere i dati adatti per il mining ed i dati analizzati mediante le tecniche di data mining sono incompleti, in quanto privi di valori o di determinati attributi di interesse; rumorosi perché contenenti errori e incoerenti poiché contengono discrepanze.

Nella quarta fase, Modeling (Modellazione), vengono adottate le regole di associazione per rilevare le relazioni tra le variabili e per estrarre set di dati molto grandi. Gli algoritmi utilizzati, in questo caso, per l'estrazione delle regole di associazione sono Generalized Rule Induction (GRI), Apriori e CARMA; di solito l'operatore specifica il supporto minimo

desiderato e i criteri di confidenza, ma per il GRI, l'operatore specifica anche quante regole di associazione desidera siano riportate.

Nella figura 1.2, illustrata in basso, viene mostrato che i fattori a produrre più incidenti sono stati l'errore umano, il carro e il binario.

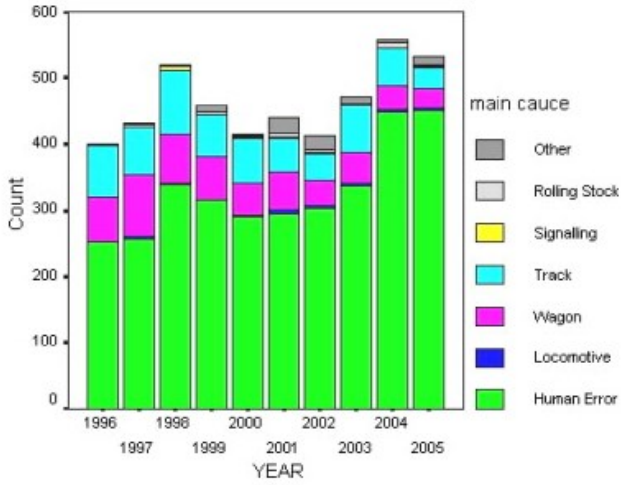

Figura 1.2: Quota di fattori di incidente dal 1996 al 2005 (fonte: Application of association rules in Iranian Railways (RAI) accident data analysis (2010)

Al fine di studiare le relazioni significative, l'esperienza dei ricercatori, insieme alle priorità degli esperti di sicurezza ferroviaria della RAI, si è concretizzata in una lista di rapporti candidati per lo studio delle regole di associazione.

La Tabella 1.4 contiene i settori esaminati considerando i fattori di infortunio:

Riferimento	Tecniche	Risultati
Chong et al. (2004)	Alberi decisionali, NN [a]	Modellando la gravità delle lesioni da incidente stradale, i tre fattori più importanti di lesioni mortali erano l'uso della cintura di sicurezza del conducente, le condizioni di scarsa visibilità della strada e il consumo di alcol da parte del conducente
Chong et al. (2005)	NN, SVM, [b] Alberi decisionali	Il documento ha sviluppato modelli che classificano accuratamente la gravità delle lesioni all'interno di cinque categorie: nessuna lesione, lesioni possibili, lesioni non invalidanti, lesioni invalidanti e lesioni mortali
Nefti e Oussalah (2004)	NN	Modello presentato prendendo come input le irregolarità nel posizionamento dei binari e predetto il rapporto di sicurezza dei binari e quindi la sicurezza del treno sui binari
Barai (2003)	Estrazione dei dati	Questo documento esamina le applicazioni delle tecniche di data mining nei problemi di ingegneria dei trasporti

Chang e Wang (2006)	Modello CART	Viene sviluppato un modello per stabilire una relazione tra la gravità della lesione e le caratteristiche del conducente / veicolo, le variabili autostradali / ambientali e le variabili degli incidenti. I risultati indicano che la variabile più importante è il tipo di veicolo
Solomon et al. (2006)	Alberi decisionali, NN, Analisi del paniere di mercato, K-Means	Il documento dimostra l'uso del data mining per valutare il miglioramento della sicurezza del traffico degli incroci controllati dal segnale a luci rosse monitorati da telecamere per ridurre gli incidenti mortali
Tesema et al. (2005)	Alberi di regressione	Questo documento applica il data mining per determinare modelli interessanti rispetto alla gravità delle lesioni sui dati sugli incidenti stradali
Sze e Wong (2007)	Regressione logistica binaria, diagnostica di regressione logistica	In questo articolo è stata rivelata una tendenza alla diminuzione del rischio di lesioni ai pedoni, controllando le influenze demografiche, dell'ambiente stradale e di altri campi. Anche le influenze del comportamento dei pedoni, della congestione del traffico e del tipo di svincolo sul rischio di lesioni ai pedoni erano soggette a variazioni temporali
Depaire et al. (2008)	Clustering	Questo documento applica il raggruppamento di classi latenti per identificare tipi di incidenti stradali omogenei. L'analisi dei cluster utilizza il tipo di veicolo come base per la segmentazione
Anderson (2009)	KDE, [c] Clustering	Questo documento studia i modelli spaziali delle lesioni da incidente stradale e applica i dati ambientali ei risultati dai modelli al fine di creare una classificazione dei punti critici degli incidenti stradali
Abugessaisa (2008)	VDM, [d] Analisi esplorativa dei dati, Clustering, SOM, [e] Classification Trees,	Questo documento scopre cluster e relazioni nel database sulla sicurezza stradale, esplora i contenuti e la struttura del set di dati e copre esplorazioni interattive basate su metodi di spazzolamento e collegamento per rilevare e riconoscere modelli interessanti nel database disponibile
Sohn e Lee (2003)	NN, albero decisionale, fusione bayesiana, insaccamento, raggruppamento	In questo documento sono stati applicati vari algoritmi per migliorare l'accuratezza dei singoli classificatori per due categorie di gravità di incidente stradale. I risultati indicano che un algoritmo di classificazione basato su cluster funziona meglio per la classificazione degli incidenti stradali in Corea
Lee et al. (2004)	Clustering	Questo documento esplora l'applicazione del data mining nelle indagini sulle situazioni di incidente. I dati utilizzati sono ottenuti dalla simulazione. La dimostrazione mostra che il data mining consente all'utente di contrassegnare l'area di impatto dell'incidente nel tempo e nello spazio

Xie et al. (2007)	NN, metodi bayesiani, binomiale negativo	Questo studio confronta tre tipi di modelli per la previsione degli incidenti automobilistici

Tabella 1.4: Fattori di infortunio (fonte: Application of association rules in Iranian Railways (RAI) accident data analysis (2010))

La Tabella 1.5, la Tabella 1.6, la Tabella 1.7, la Tabella 1.8, la Tabella 1.9 e la Tabella 1.10 contengono campioni di regole di associazione estratti dall'algoritmo GRI. Ogni tabella contiene supporto e fiducia per ogni relazione, perdite di vite umane e proprietà insieme alla rispettiva fiducia e supporto.

Job	Years of Service	Districts	Accident Type	Support	Confidence	Loss of Life	Support	Confidence	Loss of Property	Support	Confidence
Shunting man	16–20	Hormozgan	Fire accident	0.26	5	1	0.11	100	4	0.75	91.18
									5	0.99	93.33
Assistant driver	6–10	Khorasan	Fire accident	0.31	57.14	1	0.26	83.33	4	0.88	77.5
									5	1.59	88.89

Tabella 1.5: Lavoro – Anni di servizio – Distretti – Tipo di incidente (fonte: Application of association rules in Iranian Railways (RAI) accident data analysis (2010))

Job	Age	Districts	Accident Grade	Support	Confidence	Loss of Life	Support	Confidence	Loss of Property	Support	Confidence
Driver	21–30	Hormozgan	D3	0.49	90.91	4	1.92	78.16	4	0.75	91.18
									5	0.99	100

Tabella 1.6: Lavoro – Età – Aree – Grado di incidente (fonte: Application of association rules in Iranian Railways (RAI) accident data analysis (2010))

Accident Factor	Day of Week	Accident Grade	Support	Confidence	Loss of Life	Support	Confidence	Loss of Property	Support	Confidence
Human error	Saturday	D1	2.3	56.73	–	–	–	1	0.2	77.78
Human error	Sunday	I1	1.06	91.67	–	–	–	1	0.46	95.24
Human error	Saturday	D3	9.71	65.46	2	5.08	61.74	3	0.13	83.33
					5	0.42	52.63	4	1.02	76.09

Tabella 1.7: Fattore di incidente – Giorno della settimana – Grado di incidente (fonte: Application of association rules in Iranian Railways (RAI) accident data analysis (2010))

Accident Factor	Districts	Accident Type	Support	Confidence	Loss of Life	Support	Confidence	Loss of Property	Support	Confidence
Human error	Arak	Fire accident	0.46	100	3	0.15	100	–	–	–
					5	0.26	100			
Human error	Hormozgan	Fire accident	1.81	100	3	0.11	100	1	0.11	100
					4	0.75	91.18			
					5	0.99	93.33			
Human error	Esfehan	Fire accident	0.97	100	4	0.31	100	–	–	–
					5	0.71	90.62			
Human error	Tehran	Fire accident	3.89	100	4	1.99	86.67	–	–	–
					5	2.05	97.85			
Human error	South	Fire accident	0.93	100	4	0.22	90	–	–	–
					5	0.73	100			
Human error	South east	Derailment	0.77	100	2	0.07	100	–	–	–
					6	0.11	100			
Human error	Khorasan	Fire accident	2.12	100	4	0.88	77.5	1	0.26	83.3
					5	1.59	88.89			
Wagon	Arak	Collision of RS with O	0.6	77.78	–	–	–	4	0.15	100
Wagon	Hormozgan	Collision of RS with O	0.82	94.59	–	–	–	3	0.13	100
Wagon	Tehran	Collision of RS with O	1.92	73.56	–	–	–	3	0.09	100
Wagon	South	Collision of RS with O	0.57	80.77	–	–	–	3	0.07	100
Wagon	South east	Collision of RS with O	1.81	65.85	–	–	–	4	1.48	71.64
Wagon	Khorasan	Collision of RS with O	0.38	70.59	–	–	–	4	0.29	84.62
Wagon	North	Collision of RS with O	0.84	84.21	–	–	–	4	0.29	100

Tabella 1.9: Fattore di incidente – Aree – Tipo di incidente (fonte: Application of association rules in Iranian Railways (RAI) accident data analysis (2010))

Job	Accident Type	Support	Confidence
Driver	Collision of RS with O	24.42	51.36

Tabella 1.10: Lavoro – Tipo di incidente (fonte: Application of association rules in Iranian Railways (RAI) accident data analysis (2010))

Accident Factor	Accident Type	Support	Confidence
Human error	Collision of RS with O	46.29	59.9
Human error	Derailment	6.36	100
Human error	Fire accident	15.54	98.58
Wagon	Collision of RS with O	13.64	69.9

Tabella 1.11: Fattore di incidente - Tipo di incidente (fonte: Application of association rules in Iranian Railways (RAI) accident data analysis (2010))

Nella quinta fase, Evaluation (Valutazione), viene eseguita una valutazione dei risultati del modello adottato mediante dei software di data mining che richiede la supervisione di un utente qualificato. Senza il controllo umano si può ottenere un'analisi sbagliata che risulta essere peggiore di un'analisi non effettuata; pertanto i risultati del software devono essere valutati da esperti umani. Mediante il software il data miner crea i modelli e successivamente i risultati ottenuti dai modelli vengono valutati da esperti umani.

La sesta ed ultima fase, Data Presentation (Sviluppo), elabora e suggerisce alla RAI norme e regolamenti riguardanti la sicurezza, in seguito alle rilevazioni passate per evitare che si ripetano gli andamenti precedenti. Il sistema ferroviario viene visto come una catena e quando verificano gli incidenti, è come se si verificassero rotture di alcuni anelli della catena; una volta riconosciuti i punti di rottura, questi possono essere sostituiti con collegamenti più forti per far sì che, in futuro, il sistema funzioni in modo più efficiente. Tali collegamenti vengono riconosciuti con l'applicazione del data mining, scoprendo modelli ripetitivi all'interno dei dati sugli incidenti passati.

In conclusione, il lavoro di ricerca sul data mining eseguito in Iran aveva l'obiettivo di identificare le relazioni nascoste degli incidenti più comuni, le loro possibili cause e i sistemi di segnalazione con altri campi del database degli incidenti RAI. Seguendo il modello CRISP-DM sono stati applicati per l'attività di estrazione circa 6500 archivi e 38 campi di dati del database degli incidenti RAI nel corso degli anni 1996-2005 ed inoltre vengono adottate le regole di associazione per l'estrazione dei dati. Sarebbe utile l'applicazione di altre tecniche di data mining, come ad esempio la serie storica (o temporale) ottenuta dalla raccolta ordinata delle misurazioni effettuate ad intervalli regolari; i metodi di modellazione delle serie temporali presumono che la storia si ripeta,

in modo da poter prendere decisioni migliori in futuro o prevedere l'avvenimento di determinati incidenti durante l'anno in un particolare territorio o sull'intera rete.

CAPITOLO 2

INTRODUZIONE ALLE TECNICHE DI DATA MINING

In questo capitolo verranno presentate le principali tecniche di classificazione attualmente disponibili sia per dati di dimensioni non big sia per i Big Data.

2.1 Data Mining

Con il termine Data Mining si intende un insieme di tecniche di analisi e metodologie che permettono un processo automatico di estrazione di informazioni utili, chiamati pattern, da grandi quantità di dati (o dataset). Queste tecniche di estrazione automatica si basano sull'utilizzo di diversi algoritmi (*Data Mining, Cos'è? - Strumenti, Applicazioni e Rischi*)

Figura 3: Rappresentazione grafica delle fasi del Knowledge Discovery in Databases

Per poter arrivare all'estrazione di pattern è necessario seguire un processo diviso in varie fasi, chiamato Knowledge Discovery in Databases, KDD. Tali fasi sono:

1) Selezione dei dati (Data Selection): la selezione del set di dati richiede la conoscenza del dominio dal quale i dati sono presi. La rimozione dei dati non correlati tra loro dal set di dati permette una riduzione dello spazio di ricerca durante la fase data mining che si traduce in una diminuzione del tempo di analisi.
2) Pre-processamento dei dati (Data preprocessing): questa fase consiste nel pulire le informazioni, rimuovendo eventuali dati anomali (outliers) non utili all'analisi e risolvendo eventuali conflitti presenti tra i dati dopo l'integrazione delle diverse fonti.
3) Trasformazione dei dati (Data Trasformation): i dati sono trasformati e consolidati in formati adatti all'analisi delle tecniche di Data Mining. In questa fase viene ridotta la varietà dei dati preservando allo stesso tempo la qualità degli stessi.
4) Data Mining: utilizzo di alcune tecniche di Data Mining (algoritmi) per analizzare i dati e scoprire modelli interessanti o estrarre conoscenza interessante da questi dati.

5) Valutazione (Evaluation): lo step finale è la documentazione e interpretazione dei risultati raggiunti dalle fasi precedenti. Può succedere di dover tornare alle fasi precedenti per raffinare la conoscenza acquisita, o trasformare la conoscenza secondo le esigenze più richieste dall'utilizzatore.

   Infine dopo una fase di interpretazione e valutazione dei pattern estratti, si passa alla fase finale chiamata knowledge, in cui i dati finali delle analisi vengono presentati.

   Il Data Mining viene utilizzato nel settore finanziario, nel marketing e nel manufacturing. Alcuni esempi in questi campi sono:

   -l'apprendimento automatizzato: tramite le reti neurali identificano un certo pattern al cui interno sono presenti elementi con relazioni precise fra loro.

   -disposizione merce: permette di identificare i prodotti comprati assieme da un numero sufficientemente elevato di clienti.

   -direct marketing: per ridurre, ad esempio, il costo della pubblicità via posta definendo l'insieme dei clienti che, con maggiore probabilità, compreranno un nuovo prodotto di telefonia.

   -individuazione di frodi: per predire l'utilizzo fraudolento di determinate situazioni (ad esempio delle carte di credito).

   -individuazione dell'insoddisfazione del cliente: per predire clienti propensi a passare a un concorrente.

   -raggruppamento di documenti: per trovare sottogruppi di documenti che sono simili sulla base dei termini più rilevanti che in essi compaiono.

   -segmentazione del mercato: per suddividere i clienti in sottoinsiemi distinti da utilizzare come target di specifiche attività di marketing.

   Quindi le attività tipiche del Data Mining sono principalmente due:

1. Predizione: nello specifico, utilizzare alcune variabili per predire il valore incognito o futuro di altre variabili. Per la predizione dei dati si usano le tecniche di classificazione, le tecniche di regressione e la "Deviation Detection" ovvero un'analisi delle anomalie.

2. Descrizione: trovare pattern interpretabili dall'uomo che descrivano i dati. Per la descrizione dei dati si usano principalmente tecniche di Clustering e regole di associazione.

   Non esiste una tecnica migliore delle altre, ma ogni tecnica è riferita a determinati obiettivi e tipologie di dati da analizzare. Spesso i migliori risultati per trasformare i dati in informazioni si ottengono attraverso la combinazione di diverse tecniche di analisi.

   Le metodologie di Data Mining consentono di poter analizzare sia dati quantitativi, qualitativi che testuali, di elaborare un numero elevato di variabili e osservazioni, di utilizzare algoritmi ottimizzati per minimizzare il tempo di elaborazione ed infine di garantire un'interpretazione semplice del risultato.

2.2 Regole di associazione

Le regole di associazione (Guarracino, 2007), sono uno dei metodi per estrarre relazioni nascoste tra i dati. Esse descrivono correlazioni di eventi e possono essere viste come regole probabilistiche, inoltre due eventi sono correlati quando sono frequentemente osservati insieme.

Esempio: database di transazioni di vendita in un supermercato.

- {pannolino} ⇒ {birra}
- {latte, pane} ⇒ {uova}

Le regole associative hanno come scopo quello di trovare associazioni interessanti e relazioni di correlazione in grandi insiemi di transizioni. Queste regole vengono utilizzate soprattutto da grandi collezioni di dati che possono essere raccolti con facilità se esiste un concetto di "transazione".

Le regole di associazione sono una sorta di "implicazioni". La regola X⇒Y viene interpretata come: "nelle transazioni in cui compare X compare anche Y; X è detto corpo e Y è detta testa. X e Y sono itemset, collezione di uno o più elementi e si può avere anche k-itemset, ossia itemset che contiene k-elementi.

Le regole di associazione sono caratterizzate principalmente da due misure statistiche: supporto e confidenza.

- Il supporto indica la percentuale di transazioni che contengono entrambe X ed Y. È l'indicazione di quanto frequentemente l'itemset appare nel dataset.
  Se la regola è X⇒Y, supporto sarà:

$$\frac{\#\{X, Y\}}{|T|}$$

  Ad esempio se l'itemset {latte, pane} ha un supporto pari al 20%, allora esso comparirà nel dataset per il 20% delle transazioni.
- La confidenza indica, date le transazioni che contengono X, qual è la percentuale di transazioni che contengono Y. È un'indicazione di quanto spesso la regola è stata trovata vera.
  Se la regola è X⇒Y, la confidenza sarà:

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

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

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

  Esempi di regole:

  {latte, pannolino} ⇒ {birra} (s=0.4, c=0.67)

  {latte, birra} ⇒ {pannolino} (s=0.4, c=1.0)

  {pannolino, birra} ⇒ {latte} (s=0.4, c=0.67)

  {birra} ⇒ {latte, pannolino} (s=0.4, c=0.67)

  {pannolino} ⇒ {latte, birra} (s=0.4, c=0.5)

  {latte} ⇒ {pannolino, birra} (s=0.4, c=0.5)

Osservazione:

-Tutte le regole sono partizioni binarie dello stesso itemset: {latte, pannolino, birra}.

-Le regole basate sullo stesso itemset hanno sempre il medesimo supporto ma possono avere confidenze diverse.

2.2.1 Classificazione delle regole associative

- Regole booleane e quantitative:

-booleana se in tutti gli item *attr (x,v)*, *v* è un singolo valore;
-quantitativa se coinvolge attributi numerici e negli item *attr(x,v)*, *v* è un insieme di valori, tipicamente un intervallo.

- Regole mono-dimensionali e multi-dimensionali: a seconda del numero di attributi diversi coinvolti.
- Analisi di associazioni a un singolo livello o a livello multipli: a seconda che tutti gli item appartengono allo stesso livello di astrazione o no.

Uno degli algoritmi fondamentali per l'analisi di regole associative monodimensionali booleane è il principio Apriori.

Quest'ultimo si basa sul concetto di itemset frequente; la frequenza di un itemset è il numero delle istanze che lo soddisfano ed in particolare un itemset frequente è un itemset la cui frequenza relativa supera la soglia di supporto minimo. Per trovare le regole associative forti bisogna innanzitutto determinare tutti gli itemset frequenti con supporto minimo *s*; se X è un itemset frequente e X=$X_1 \sqcup X_2$, allora $X_1 \Rightarrow X_2$ è una regola di associazione che supera la soglia minima di supporto. Se la regola supera anche la soglia di confidenza minima, allora è una regola forte.

Per generare gli itemset frequenti viene utilizzato l'approccio naive, secondo il quale ogni itemset nel reticolo è candidato ad essere frequente; in particolare calcola il supporto per ogni candidato scorrendo il database e confronta ogni transazione con ogni candidato. Per la generazione di itemset frequenti vengono adottate 3 strategie:

1. Ridurre il numero dei candidati (M) utilizzando tecniche di pruning;
2. Ridurre il numero delle transazioni (N) quando il numero degli itemset è troppo elevato;
3. Ridurre il numero delle comparazioni (NM) utilizzando strutture dati efficienti per memorizzare i candidati o le transazioni.

Nella prima strategia viene adottato il principio Apriori, il quale si basa sul concetto secondo cui un itemset è frequente, allora anche tutti i suoi sotto-insiemi devono esserlo. In maniera iterativa, trovo tutti i k-itemset frequenti per k da 1 in poi; tutti i possibili (k+1) itemset frequenti sono ottenuti dall'unione di due k-itemset frequenti ed inoltre non è necessario controllare tutti i possibili sottoinsiemi di k+1 elementi per sapere se sono frequenti.

Il principio Apriori è dovuto alla seguente proprietà del supporto:

$$\forall X, Y:(X \subseteq Y) \Rightarrow s(X) \geq s(Y)$$

Quest'ultima mostra che il supporto di un itemset non eccede il supporto dei suoi sottoinsiemi ed è nota come proprietà anti-monotona del supporto.

Il principio si compone di due fasi:

-prima fase: individuazione degli itemset frequenti;

-seconda fase: generazione delle regole associative forti.

A queste due fasi può essere aggiunta una terza: la fase di valutazione dell'importanza delle regole forti individuate.

Nella prima fase, l'algoritmo Apriori individua tutti gli itemset frequenti, cioè quelli con supporto superiore alla soglia minima, partendo da quelli costituiti da un solo elemento. Successivamente, sulla base del principio Apriori, individua gli itemset frequenti composti da due elementi e itera il procedimento fino a raggiungere il numero massimo di elementi in un itemset, pari al numero di pagine del sito.

Nella seconda fase vengono individuate le regole associative forti, partendo dagli itemset frequenti estratti nella fase precedente. Per ciascun itemset frequente si costruiscono tutte le possibili regole, date dalla combinazione degli elementi dell'itemset come antecedente e conseguente. Per ogni regola viene calcolata la confidenza e vengono eliminate tutte le regole che non soddisfano la soglia minima prefissata. Il risultato è un insieme di regole forti.

Al termine di questa fase, viene calcolato l'indice di lift per ciascuna regola forte, in modo da trovare le regole maggiormente esplicative rispetto alla sola presenza del conseguente. Tali regole presentano un valore per l'indice di lift superiore all'unità.

Come già affermato, può succedere che gli algoritmi utilizzati per generare le regole associative tendono a produrne molte inutili, tra cui possiamo trovarne alcune non interessanti o meglio ovvie ed altre possono essere ridondati, ad esempio se {A,B,C} ⇒ {D} e {A,B} ⇒ {D} hanno lo stesso supporto e confidenza.

Di conseguenza vengono utilizzati misure di interesse che permettono di ordinare o eliminare le regole inutili e sino ad ora le uniche misure di interesse utilizzate sono il supporto e la confidenza. Esistono sia misure oggettive, le quali danno priorità alle regole sulla base di criteri statistici calcolati a partire dai dati, sia misure soggettive che, invece, danno priorità alle regole sulla base di criteri definiti dall'utente. Esse vengono applicate soprattutto nelle fasi di preprocessing, mining e postprocessing.

La terza strategia, ossia la riduzione del numero delle comparazioni, si basa soprattutto sul conteggio dei candidati che richiede la scansione delle transazioni per determinare il supporto degli itemset candidati; in particolare per ogni transazione si genereranno tutti i k-itemset e si procederà a incrementare il supporto dei corrispondenti itemset candidati.

2.2.2 Algoritmo Apriori e Algoritmo FPGrowth

I metodi principali utilizzati per l'estrazione degli itemset frequenti sono basati sull' algoritmo Apriori e sull'algoritmo FPGrowth. Il primo, come già spiegato nel paragrafo precedente, è basato su una struttura a livelli e si basa soprattutto sul principio Apriori. Il meccanismo con cui funziona l'algoritmo si basa sul considerare, a ogni livello, degli itemset composti da un numero sempre crescente di itemset, calcolandone il supporto e cancellando dai possibili itemset frequenti di quel livello quelli i cui subsets non sono contenuti negli itemset frequenti del livello precedente.Prendiamo ad esempio il dataset:

ID	ITEM
1	pasta, olio
2	olio, sale, pepe
3	pane, sale, pepe, origano
4	pane, pepe, origano
5	pane, olio, sale
6	pane, olio, sale, pepe

Tabella 2: Esempio di dataset utilizzato per algoritmo priori

Al primo livello si considerano tutti gli itemset possibili contenenti al loro interno un item solo e si calcola il loro supporto:

ITEMSET	SUP
Pane	5
Olio	4
Sale	4
Pepe	4
Origano	2

Tabella 2.1: Itemset candidati di livello 1

A questo punto il supporto di ogni itemset candidato viene paragonato con la soglia limite scelta per considerare l'itemset frequente e, se maggiore, l'itemset corrispondente viene inserito nell'elenco degli itemset frequenti del livello considerato. Supponendo di scegliere come soglia 1, tutti gli itemset selezionati hanno supporto maggiore e quindi tutti vengono considerati come itemset frequenti di livello 1.

Il passo successivo consiste nel considerare tutti i possibili itemset composti da due item formabili utilizzando gli item si livello 1. Una volta generati i candidati anche in questo caso ne si calcola il supporto.

ITEMSET	SUP
pane, olio	3
pane, sale	3
pane, pepe	3
pane, origano	2
olio, sale	3
olio, pepe	2
olio, origano	0
sale, pepe	3
sale, origano	1
pepe, origano	2

Tabella 2.2: Itemset candidati di livello 2

Siccome la soglia minima stabilita è pari a 1 due itemset (olio, origano) e (sale, origano) non hanno un supporto sufficiente per essere considerati frequenti e vengono quindi scartati. L'elenco degli itemset frequenti di secondo livello è dunque:

ITEMSET
pane, olio
pane, sale
pane, pepe
pane, origano
olio, sale
olio, pepe
sale, pepe
pepe, origano

Tabella 2.3: Itemset frequenti di livello 2

Il passo successivo consiste nel calcolare gli itemset di livello 3, costituiti da tutte le combinazioni possibili formate da 3 elementi considerando gli itemset di secondo livello. Anche in questo caso, dopo aver calcolato quali sono gli itemset candidati, viene scannerizzato il dataset per calcolare il loro supporto e segnarlo in una tabella simile alle precedenti.

ITEMSET	SUP
pane, olio, sale	2
pane, olio, pepe	1
pane, olio, origano	0
pane, sale, pepe	2
pane, sale, origano	0
pane, pepe, origano	2
olio, sale, pepe	2
sale, pepe, origano	0

Tabella 2.4: Itemset candidati di livello 3

In questo caso, oltre a non avere un supporto sufficiente (minore della soglia 1), gli itemset (pane, olio, origano), (pane, sale, origano) e (sale, pepe, origano) vengono scartati perché i loro subsets (olio, origano) e (sale, origano) non sono presenti negli itemset frequenti di livello 2. Anche se nella tabella mostrata è segnato il loro supporto,

questi itemset sono scartati dall'algoritmo ancora prima che esso venga calcolato, sfruttando il principio Apriori. L'itemset (sale, olio, pepe) viene invece scartato perché non ha un supporto maggiore o uguale alla soglia stabilita. L'elenco completo degli itemset frequenti di livello 3 risulta dunque:

ITEMSET
pane, olio, sale
pane, sale, pepe
pane, pepe, origano
olio, sale, pepe

Tabella 2.5: Itemset frequenti di livello 3

L'ultimo passo dell'algoritmo calcola gli itemset candidati di livello 4, esattamente come per i passi precedenti:

ITEMSET	SUP
pane, olio, sale, pepe	1

Tabella 2.6: Itemset candidati di livello 4

Anche in questo caso il subset (pane, olio, pepe) non è presente negli itemset frequenti di livello 3 e quindi l'itemset non viene considerato come frequente; questo rende l'elenco degli itemset frequenti di livello 4 vuoto.

L'algoritmo Apriori è molto utilizzato per analisi dati di questo tipo ed estrazione di regole di associazione ma la sua efficienza, tuttavia, dipende da vari fattori come il tempo di esecuzione dell'algoritmo, in quanto è correlato con il numero delle transazioni in esso presenti; dalla lunghezza media delle transazioni poiché le transazioni lunghe tendono a determinare itemset frequenti più lunghi. È fondamentale anche il numero di item presenti nel dataset e bisogna saper scegliere la soglia di supporto minimo in quanto abbassando il valore di soglia per min_supp si otterrà un maggior numero di itemset candidati e potenzialmente aumenterà la lunghezza massima degli itemset frequenti.

Il secondo metodo utilizzato è l'algoritmo FPGrowth, anch'esso utilizzato per la generazione di itemset fequenti che basa il suo funzionamento sulla costruzione di una struttura al albero che modella il dataset considerato. In particolare, gli itemset frequenti vengono estratti percorrendo l'albero, senza dover leggere più volte il database. Con l'algoritmo FPGrowth, a differenza di quello che avviene per l'algoritmo Apriori, il

dataset viene scannerizzato solamente due volte: la prima per calcolare il supporto dei singoli item, la seconda per la costruzione dell'albero rappresentante il dataset e che prende il nome di FP-tree.

Per spiegarne il funzionamento prendiamo ad esempio il seguente dataset dove le transazioni sono composte da parole all'interno delle quali le lettere corrispondono ai singoli team:

TID	ITEMSET
1	P, A, S, T, A
2	P, I, S, T, A
3	L, A, S, A, G, N, E
4	P, O, R, T, A

Tabella 2.7: Dataset di esempio per algoritmo FPGrowth

Il primo step dell'algoritmo FPGrowth è identico al primo step dell'algoritmo Apriori: si stabilisce una soglia minima di supporto, ogni item viene isolato e ne viene calcolato il supporto all'interno del dataset iniziale. Se il supporto dell'item è maggiore della soglia prefissata allora esso viene considerato dall'algoritmo, altrimenti viene scartato e non contribuisce alla costruzione dell'albero FP-tree. Nel caso in esempio il risultato di questo primo passaggio è il seguente:

ITEM	SUP
P	3
A	4
S	3
T	3

Tabella 2.8: Item dopo primo step dell'algoritmo FPGrowth

Da qui in avanti iniziano le differenze con l'algoritmo visto in precedenza.

Per prima cosa gli item vengono ordinati in base al supporto secondo un ordine crescente:

ITEM	SUP
T	3
S	3
P	3
A	4

Tabella 2.9: Item ordinati in base a supporto crescente

La stessa operazione viene effettuata sul dataset di partenza, considerando per ogni transazione solo gli item selezionati e ordinandoli in base al supporto secondo un ordine decrescente:

TID	ITEMSET
1	A, P, S, T
2	A, P, S
3	A, S
4	A, P, T

Tabella 2.10: Itemset ordinati per supporto decrescente degli item al loro interno

Una volta effettuata anche questa questa operazione inizia la costruzione vera e propria dell'albero. L'FP-tree è composto da una serie di nodi dove ogni nodo corrisponde a un item del dataset. Ogni transazione è registrata aggiungendo ogni item come figlio del nodo corrispondente all'item precedente, andando a costituire così un ramo dell'albero. Ogni item ha associato un contatore che stabilisce quante volte si passa da quel nodo nella costruzione dell'albero.

Prendendo per esempio la prima transazione filtrata con solo gli item con supporto maggiore alla soglia (A, P, S, T):

Figura 2: Costruzione dell'FP-tree dopo analisi della prima transazione del dataset

Dopo l'analisi della seconda transazione gli indici degli item vengono aumentati e l'albero diventa:

Figura 2.1: Costruzione dell'FP-tree dopo analisi della seconda transazione del dataset

Alla terza transazione, quello che considera l'itemset (A, S), il percorso effettuato precedentemente non va più bene e viene inserito in nodo esterno dopo l'item A:

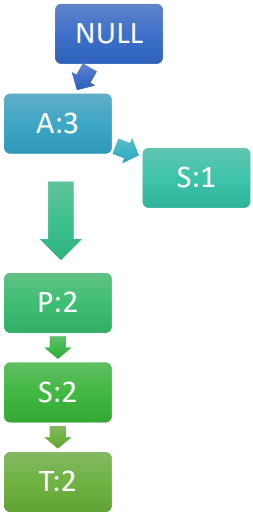

Figura 2.2: Costruzione dell'FP-tree dopo analisi della terza transazione del dataset

La stessa cosa vale anche per l'ultima transazione, dove dopo l'item P è presente subito l'item T:

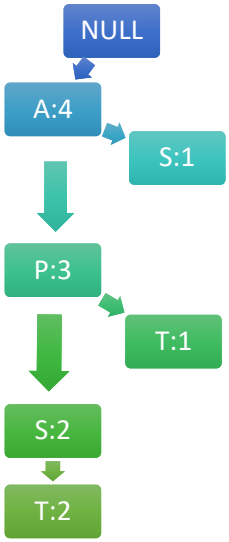

Figura 2.3: Costruzione dell'FP-tree dopo analisi della quarta transazione del dataset

Dopo la costruzione dell'albero è conclusa, per ogni item escluso quello con il supporto maggiore viene costruito il Conditional Pattern Base (CPB), ovvero l'elenco dei subset che lo precedono nell'albero.

Il risultato finale è una tabella simile alla seguente:

ITEM	CONDITIONAL PATTERN BASE (CPB)
T	(APS:1), (AP:1)
S	(AP:1), (A:1)
P	(A:3)
A	-

Tabella 2.11: Costruzione dei Conditional Pattern Base

Il passo successivo consiste nel selezionare, per ogni item, i subset comuni ai CPB determinati precedentemente e costruire quello che viene chiamato Conditional FP tree:

ITEM	CONDITIONAL PATTERN BASE (CPB)	CONDITIONAL FP TREE
T	(APS:1), (AP:1)	(AP:2)
S	(AP:1), (A:1)	(A:2)
P	(A:3)	(A:3)
A	-	-

Tabella 2.12: Costruzione dei Conditional FP tree

Infine, la generazione degli itemset frequenti finali avviene combinando i Conditional FP tree trovati con l'item base a cui fanno riferimento, creando tutti i possibili subset:

ITEM	CONDITIONAL FP TREE	FREQ ITEMSET
T	(AP:2)	(AT:2), (PT:2), (APT:2)
S	(A:2)	(AS:2)
P	(A:3)	(AP:3)
A	-	-

Tabella 2.13: Generazione degli itemset frequenti finali

L'algoritmo FPGrowth è più efficiente rispetto all'algoritmo Apriori. Il tempo impiegato a costruire l'albero dipende dal numero di item e di transazioni presenti nel dataset iniziale.

CAPITOLO 3

ANALISI DEL DATASET E DEI RISULTATI

In questo capitolo spiegherò l'analisi di un dataset, contenente dati di guasto di una locomotiva elettrica, effettuata tramite le Association Rules durante il tirocinio. Esso è stato articolato in due fasi: nella prima fase ho studiato la metodologia e il software adottato per l'analisi mentre nella seconda fase ho condotto una vera e propria analisi del dataset.

3.1 RapidMiner

RapidMiner (*RapidMiner | Best Data Science & Machine Learning Platform*, n.d.), logo in Figura 3, è un software molto completo per analizzare dati in grandi quantità.

Figura 3: Logo del software RapidMiner

RapidMiner è un software che offre procedure di data mining e machine learning che comprendono: caricamento e trasformazione dei dati (estrazione, trasformazione), preelaborazione e visualizzazione dei dati, analisi predittiva e modellazione statistica, valutazione e implementazione. RapidMiner è scritto nel linguaggio di programmazione Java e fornisce una GUI, Graphical User Interface (interfaccia grafica), per progettare ed eseguire flussi di lavoro analitici. Questi flussi di lavoro sono chiamati "Processi" in RapidMiner e sono costituiti da più "Operatori". Ogni operatore esegue una singola attività all'interno del processo e l'output di ciascun operatore costituisce l'input di quello successivo. RapidMiner fornisce schemi di apprendimento, modelli e algoritmi.

Di seguito sono elencati i principali operatori utilizzati per la creazione del processo, tratti dalla documentazione ufficiale di RapidMiner (*RapidMiner Studio Manual*).

Read Excel: Questo operatore viene utilizzato per caricare i dati dal file Excel. La tabella dei dati deve avere un formato tale che ogni riga sia un esempio e ogni colonna rappresenti un attributo. Si noti che la prima riga del foglio Excel potrebbe essere

utilizzata per i nomi degli attributi che possono essere indicati da un parametro. I valori dei dati mancanti in Excel devono essere indicati da celle vuote o da celle contenenti solo "?".

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

Numerical to Polynominal: Questo operatore viene utilizzato per modificare il tipo di attributi numerici in un tipo polinomiale. Questo operatore cambia il tipo di attributi selezionati, cioè ogni nuovo valore numerico è considerato un altro possibile valore per l'attributo polinomiale; ogni valore numerico viene semplicemente utilizzato come valore nominale del nuovo attributo. Poiché gli attributi numerici possono avere un numero grande di valori diversi anche in un intervallo ridotto, la conversione di un tale attributo numerico in forma polinomiale genererà un numero enorme di valori possibili per il nuovo attributo.

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

FP-Growth: Questo operatore calcola in modo efficiente tutti gli itemset frequenti dal dataset specificato utilizzando la struttura di dati dell'albero FP. È obbligatorio che tutti gli attributi dell'input ExampleSet siano di tipo "binominal", infatti sarà necessario il blocco aggiunto precedentemente "Nominal to Binominal". In parole semplici, gli itemset

frequenti (Frequent Pattern) sono gruppi di elementi che appaiono spesso nell'insieme dei dati. Il numero di transazioni (ovvero righe totali del dataset) è solitamente assunto come molto grande. Il problema degli itemset frequenti è quello di trovare insiemi di elementi che appaiono insieme almeno sopra una certa soglia, definita dai criteri di "supporto minimo". Il fatto di trovare itemset frequenti è spesso visto come la scoperta di "regole di associazione". È stato scelto l'algoritmo FP-Growth per questa analisi, tuttavia esistono anche molti altri algoritmi di estrazione di itemset frequenti, ad esempio l'algoritmo Apriori, ed uno dei principali vantaggi di FP-Growth rispetto ad Apriori è che utilizza solo due scansioni di dati ed è quindi spesso applicabile anche su grandi set di dati.

Questo operatore ha due modalità di lavoro di base:

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

Create Association Rule: Come si è detto precedentemente, l'operatore FP-Growth trova gli itemset frequenti e poi sono necessari operatori come "Create Association Rule" che utilizzano questi itemset frequenti per l'estrazione delle regole di associazione. Le regole di associazione sono dichiarazioni if/then che aiutano a scoprire relazioni tra dati apparentemente non correlati. Un esempio di regola dell'associazione potrebbe essere "Se un cliente acquista uova, ha l'80% di probabilità di acquistare anche latte". Una regola di associazione ha due parti; un antecedente (if) e un conseguente (then) combinato con l'antecedente.

Le regole di associazione vengono create analizzando i dati per gli itemset if/then frequenti e utilizzando i criteri di supporto e confidenza per identificare le relazioni più importanti. Il supporto è un'indicazione di quanto frequentemente gli articoli compaiono nel database mentre la confidenza indica il numero di volte in cui le affermazioni if/then sono state vere. Tali informazioni possono essere utilizzate come base per le decisioni relative ad attività di marketing quali, ad esempio, prezzi promozionali o posizionamenti di prodotti.

3.2 Costruzione del processo con RapidMiner

in questo paragrafo è descritto come il dataset è stato dato in input al software RapidMiner e come si è svolto il processo passo per passo per arrivare ad ottenere dei risultati.

In principio, il dataset è stato fatto leggere dal software RapidMiner attraverso l'operatore "ReadExcel" in formato foglio di calcolo Excel. Nella Figura 3 è rappresentato il processo intero.

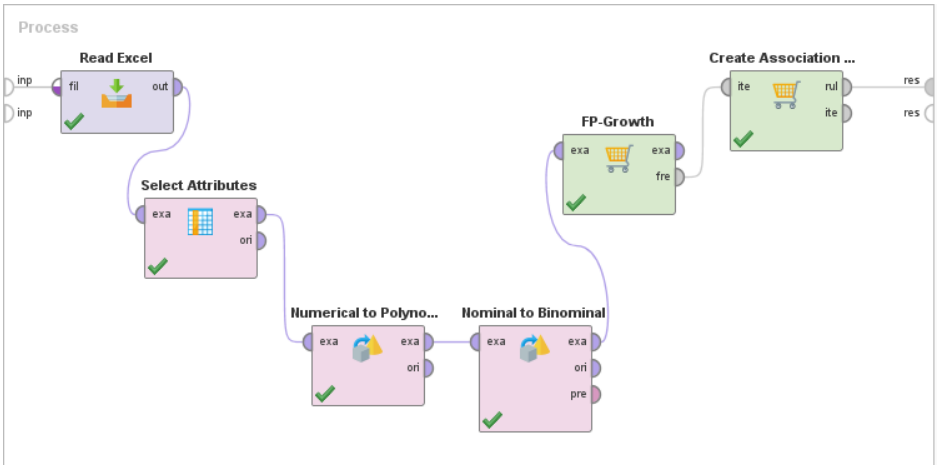

Figura 3: Processo creato con il software RapidMiner

Dopo quest'operatore è stato applicato l'operatore "Select Attributes" dove sono stati selezionati gli attributi di interesse. Successivamente sono stati aggiunti due operatori: "Numerical to Polynominal" e "Nominal to Binominal", blocchi necessari per effettuare il run con l'algoritmo FP-Growth, come menzionato nel paragrafo riguardante il processo di RapidMiner.

Infatti il passo successivo è stato aggiungere l'operatore FP-Growth, che è stato collegato all'output del processo e collegato a sua volta all'operatore "Create Association Rule". Per questi due operatori sono stati assegnati due parametri fondamentali per l'analisi: per FP-Growth (Figura 3.1) è stato necessario assegnare un supporto minimo da utilizzare per l'estrazione delle regole di associazione (Figura 3.2), e il minimo supporto utilizzato è stato l'1%.

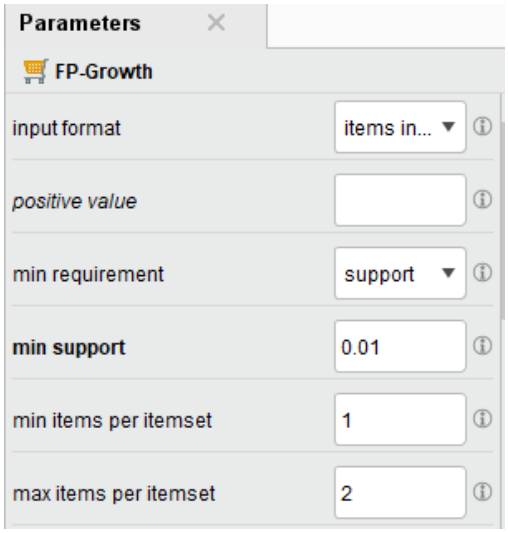

Figura 3.1: Parametri dell'operatore FP-Growth

La scelta dell'1% è stata fatta perché così l'algoritmo sarebbe stato in grado di trovare più regole e poi si sarebbero selezionate solo successivamente quelle di maggiore importanza. Invece, se si avesse scelto un minimo supporto più alto, l'algoritmo avrebbe potuto eliminare regole di qualità alta che magari sarebbero risultate utili per l'analisi.

Per quanto riguarda invece, l'ultimo operatore si doveva assegnare un livello di minima confidenza; questo livello, in questa analisi, è stato posto al 10%, come in Figura 3.2.

La spiegazione è simile a quella precedente riguardante il minimo supporto: filtrando con un livello molto basso, poiché il 10% è da considerarsi molto basso, ci permette di ottenere molte più regole rispetto ad un valore più alto, e quindi anche di non eliminare regole di associazione che potrebbero essere fondamentali per la lettura dei risultati.

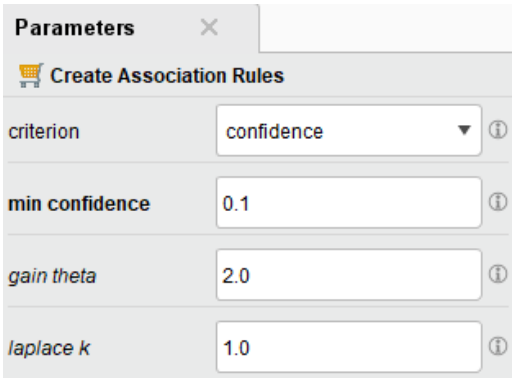

Figura 3.2: Parametri dell'operatore Create Association Rules

Dopo aver impostato questi parametri generali, si è proceduto con il run dell'algoritmo e si è aspettato che uscissero i risultati, ovvero le regole di associazione. Queste regole sono poi state inserite in un foglio di calcolo (Excel) e suddivisi per attributi.

Nel paragrafo successivo, verrà spiegata più approfonditamente questa suddivisione dei risultati.

3.3 Analisi delle regole di associazione

Questo paragrafo è dedicato alla lettura e all'interpretazione dei risultati ottenuti con il processo citato precedentemente e si sono citate le regole ritenute di fondamentale importanza; quindi sarà utile per leggere e filtrare le regole di associazione. Tra i risultati ottenuti ho preso in considerazione solo 7 attributi, in quanto li ho ritenuti più interessanti per l'analisi, con i relativi valori di supporto e confidenza:

-Process: tipo di processo in corso durante la rilevazione del guasto

-EventId: tipo di evento di guasto

-Duration (seconds): durata dell'interruzione di servizio

-Priority: priorità assegnata al guasto

-DCUState: stato della Digital Control Unit

-VCUState: stato della Vehicle Control Unit

-Configurazione di locomotiva

-ErrorCode: codice relativo all'errore

3.3.1 Regole di associazione tra VCUState e Configurazione Locomotiva

Premises	Conclusions	Support	Confidence
VCUState=4260	ConfigLoco=30	0.013926515	0.400243309
VCUState=4260	Configurazione di locomotiva=30	0.015365729	0.441605839
VCUState=4260	Configurazione di locomotiva=11	0.016804944	0.482968369
VCUState=4260	ConfigLoco=11	0.018117168	0.520681265
VCUState=4400	ConfigLoco=2	0.014349813	0.265673981
VCUState=4400	Configurazione di locomotiva=2	0.014349813	0.265673981
VCUState=4400	ConfigLoco=11	0.039663054	0.734326018
VCUState=4400	Configurazione di locomotiva=11	0.039663054	0.734326018

Tabella 3: Regole di associazione tra VCUState e Configurazione Locomotiva

Nella Tabella 3 si riportano le regole di associazione tra gli attributi VCUState e Configurazione Locomotiva; viene considerato anche l'attributo ConfigLoco in quanto coincide con Configurazione Locomotiva.

La regola VCUState=4260 → ConfigLoco=30 indica che i due eventi, Vehicle Control Unit e la configurazione di locomotiva, si verificano congiuntamente nell' 1.39% dei casi. Inoltre, se si verifica VCUState=4260, allora nel 40% dei casi si verifica anche ConfigLoco=30. Dalla Tabella 3 si evince anche che si ottiene la stessa regola ma i due eventi si verificano nell' 1.53% dei casi, e se si verifica VCUState=4260, allora nel 44% dei casi si verifica anche Configurazione di locomotiva=30.

La regola VCUState=4260 → Configurazione di locomotiva=11 indica che i due eventi si verificano insieme nell' 1.68% dei casi e oltre a ciò, se si verifica VCUState=4260, nel 48% dei casi si verifica anche Configurazione di locomotiva=11. Quanto espresso in precedenza si verifica anche per tale regola: VCUState=4260 → ConfigLoco=11; si ottiene la stessa regola ma con percentuali diverse di supporto (1.81%) e di confidenza (52%).

Un caso differente si ottiene con la regola VCUState=4400 → ConfigLoco=2; essa mostra che i due eventi si verificano assieme nell'1.43% dei casi e se si verifica VCUState=4400, allora nel 26% si verifica anche ConfigLoco=2. Con la regola VCUState=4400 →

Configurazione di locomotiva=2 si ottengono le stesse percentuali di supporto e confidenza ed in questo caso una delle due regole è ridondante, in quanto non aggiunge ulteriori informazioni.

Quanto espresso precedentemente, accade anche nella regola VCUState=4400 → ConfigLoco/Configurazione di locomotiva=11; la quale mostra mostra che i due eventi, Vehicle Control Unit e la configurazione di locomotiva, si verificano congiuntamente nel 3.97% dei casi. Oltre a quanto già detto, se si verifica VCUState=4400, allora si verifica anche ConfigLoco/Configurazione di locomotiva=11. Quindi una delle due regole possiamo tralasciarla.

3.3.2 Regole di associazione tra VCUState e DCUState

Premises	Conclusions	Support	Confidence
VCUState=4400	DCUState=40050	0.012487301	0.231191222
VCUState=4400	DCUState=40010	0.014984761	0.277429467
VCUState=4400	DCUState=40030	0.026159837	0.484326018

Tabella 3.1: Regole di associazione tra VCUState e DCUState

Nella Tabella 3.1, invece vengono riportate le regole di associazione tra gli attributi VCUState e DCUState.

La regola VCUState=4400 → DCUState=40050 indica che i due eventi, Vehicle Control Unit e Digital Control Unit, si verificano congiuntamente nell' 1.25% dei casi. Inoltre, se si verifica VCUState=4400, allora nel 23.1% dei casi si verifica anche DCUState=40050.

Dal processo in RapidMiner si è ottenuta anche la regola VCUState=4400 → DCUState=40010; la quale mostra che i due eventi si manifestano insieme nell'1.50% dei casi ed inoltre, se si verifica VCUState=4400, nel 27.7% dei casi si verifica anche DCUState=40010.

Quanto espresso precedentemente si verifica anche per tale regola VCUState=4400 → DCUState=40030 ma con percentuali diverse di supporto (2.61%) e di confidenza (48.4%).

3.3.3 Regole di associazione tra Configurazione Locomotiva e Priority

Premises	Conclusions	Support	Confidence
Configurazione di locomotiva=2	Priority=C	0.014476803	0.733905579
Configurazione di locomotiva=11	Priority=B	0.019429393	0.219617224
Configurazione di locomotiva=11	Priority=C	0.058245851	0.6583732057

Tabella 3.2: Regole di associazione tra Configurazione Locomotiva e Priority

Nella Tabella 3.2 vengono riportate le regole di associazione tra gli attributi Configurazione di locomotiva e Priority.

La regola Configurazione di locomotiva=2 → Priority=C indica che i due eventi, Configurazione di locomotiva e la priorità assegnata al guasto, si verificano congiuntamente nell' 1.44% dei casi. Inoltre, se si verifica Configurazione di locomotiva=2, allora nel 73.4% dei casi si verifica anche Priority=C.

La regola Configurazione di locomotiva=11 → Priority=B mostra che i due eventi si verificano insieme nell' 1.94% dei casi e oltre a ciò, se si verifica Configurazione di locomotiva=11, nel 21.9% dei casi si verifica anche Priority=B.

Dall'analisi si è ottenuta anche la regola Configurazione di locomotiva=11 → Priority=C; la quale indica che i due eventi si manifestano insieme nel 5.82% dei casi ed inoltre, se si verifica Configurazione di locomotiva=11, nel 65.8% dei casi si verifica anche Priority=C.

3.3.4 Regole di associazione tra Configurazione Locomotiva e DCUState

Premises	Conclusions	Support	Confidence
Configurazione di locomotiva=2	DCUState=40030	0.011556044	0.585836909
Configurazione di locomotiva=11	DCUState=40030	0.016677954	0.188516746
Configurazione di locomotiva=2	DCUState=40000	0.020487639	0.231578947

Tabella 3.3: Regole di associazione tra Configurazione Locomotiva e DCUState

Nella Tabella 3.3 vengono mostrate le regole di associazione tra gli attributi Configurazione di locomotiva e DCUState.

La regola Configurazione di locomotiva=2 → DCUState=40030 indica che i due eventi, Configurazione di locomotiva e la Digital Control Unit, si verificano congiuntamente nell' 1.15% dei casi. Inoltre, se si verifica Configurazione di locomotiva=2, allora nel 58.6% dei casi si verifica anche DCUState=40030.

La regola Configurazione di locomotiva=11 → DCUState=40030 mostra che i due eventi si verificano insieme nell' 1.67% dei casi e oltre a ciò, se si verifica Configurazione di locomotiva=11, nel 18.85% dei casi si verifica anche DCUState=40030.

Si è ottenuta anche la regola Configurazione di locomotiva=11 → DCUState=40000; la quale indica che i due eventi si manifestano insieme nel 2% dei casi ed inoltre, se si verifica Configurazione di locomotiva=11, nel 23.16% dei casi si verifica anche DCUState=40000.

Quindi si evince che la regola caratterizzata da un supporto maggiore è Configurazione di locomotiva=11 → DCUState=40000, mentre quella caratterizzata da una confidenza maggiore è Configurazione di locomotiva=2 → DCUState=40030.

3.3.5 Regole di associazione tra Configurazione Locomotiva e VCUState

Premises	Conclusions	Support	Confidence
Configurazione di locomotiva=2	VCUState=4400	0.014349813	0.727467811
Configurazione di locomotiva=11	VCUState=4260	0.016804944	0.189952153
Configurazione di locomotiva=11	VCUState=4400	0.039663054	0.448325358
Configurazione di locomotiva=30	VCUState=4260	0.015365729	0.960317460

Tabella 3.4: Regole di associazione tra Configurazione Locomotiva e VCUState

Nella Tabella 3.4 vengono indicate le regole di associazione tra gli attributi Configurazione di locomotiva e VCUState.

La regola Configurazione di locomotiva=2 → VCUState=4400 mostra che i due eventi, Configurazione di locomotiva e la Vehicle Control Unit, si verificano congiuntamente nell'

1.43% dei casi. Inoltre, se si verifica Configurazione di locomotiva=2, allora nel 72.7% dei casi si verifica anche VCUState=4400.

La regola Configurazione di locomotiva=11 → VCUState=4260 indica che i due eventi si manifestano insieme nell' 1.68% dei casi e oltre a ciò, se si verifica Configurazione di locomotiva=11, nel 18.99% dei casi si verifica anche VCUState=4260.

Dall'analisi svolta si è ottenuta la regola Configurazione di locomotiva=11 → VCUState=4400; la quale indica che i due eventi si verificano insieme nel 3.96% dei casi ed inoltre, se si verifica Configurazione di locomotiva=11, nel 44.8% dei casi si verifica anche VCUState=4400.

È emersa anche la regola Configurazione di locomotiva=30 → VCUState=4260; la quale mostra che i due eventi si manifestano insieme nell'1.53% e, se si verifica Configurazione di locomotiva=30, nel 96% dei casi si verifica anche VCUState=4260.

Quindi, da quanto analizzato, si evince che la regola caratterizzata da un supporto maggiore è Configurazione di locomotiva=11 → VCUState=4400, mentre quella caratterizzata da una confidenza maggiore è Configurazione di locomotiva=30 → VCUState=4260.

3.3.6 Regole di associazione tra Priority e Description

Premises	Conclusions	Support	Confidence
Priority=B	Description=Eccezione Richiesta apertura IR	0.010963427	0.173360107
Priority=B	Description=Sonde Temperatura Motore 3	0.025397900	0.401606425
Priority=C	Description=Frenatura pneumatica in atto	0.114290551	0.125844791
Priority=C	Description=Sforzo di trazione reso>0	0.181510328	0.199860172
Priority=C	Description=Sforzo di trazione reso<0	0.261725364	0.288184572

Tabella 3.5: Regole di associazione tra Priority e Description

Nella Tabella 3.5 vengono mostrate le regole di associazione tra gli attributi Priority e Description.

La regola Priority=B → Description=Eccezione Richiesta apertura IR mostra che i due eventi, Priorità assegnata al guasto e la descrizione del guasto, si verificano insieme nell'

1.09% dei casi. Inoltre, se si verifica Priority=B, allora nel 17.3% dei casi si verifica anche Description=Eccezione Richiesta apertura IR.

La regola Priority=B → Description=Sonde Temperatura Motore 3 indica che i due eventi si verificano congiuntamente nell' 2.53% dei casi e oltre a ciò, se si verifica Priority=B, nel 40.16% dei casi si verifica anche Description= Sonde Temperatura Motore 3.

Dal processo eseguivo si è ottenuta la regola Priority=C→ Description=Frenatura pneumatica in atto; la quale indica che i due eventi si manifestano insieme nel 11.43% dei casi ed in più se si manifesta Priority=C, nel 12.58% dei casi si verifica anche Description=Frenatura pneumatica in atto.

La regola Priority=C → Description=Sforzo di trazione reso>0 mostra che i due eventi si verificano congiuntamente nell'1.81% dei casi e, se si verifica Priority=C, nel 19.98% dei casi si verifica anche Description=Sforzo di trazione reso>0.

Un'ultima regola riguardante i due attributi sopra citati è la seguente: Priority=C → Description=Sforzo di trazione reso<0; i due eventi avvengono insieme nel 26.17% ed inoltre, se si manifesta Priority=C, nel 28.82% si manifesta anche Description=Sforzo di trazione reso<0.

Quindi è evidente che la regola caratterizzata da un supporto maggiore è Priority=C → Description=Sforzo di trazione reso<0, mentre quella caratterizzata da una confidenza maggiore è Priority=B→ Description=Sonde Temperatura Motore 3.

3.3.7 Regole di associazione tra EventId e Description

Premises	Conclusions	Support	Confidence
EventId=101	Description=Eccezione Richiesta apertura IR	0.010963427	1.0
EventId=111	Description=Nuovo pittogramma attivo	0.023408398	1.0
EventId=111	Description=Nuovo pittogramma attivo	0.023408398	1.0
EventId=113	Description=Velocità maggiore di 5 km/h	0.082162207	1.0
EventId=114	Description=Sforzo di trazione reso>0	0.181510328	0.999300862
EventId=115	Description=Sforzo di trazione reso<0	0.261725364	1.0

EventId=116	Description=Frenatura pneumatica in atto	0.114290551	1.0
EventId=117	Description=Fren. Pneum. Da Rubinetto	0.074796816	1.0
EventId=122	Description=Sblocco impulsi	0.072384016	1.0
EventId=323	Description=RSV03:Richiesta taglio trazione	0.023493057	1.0
EventId=324	Description=RSV04:Richiesta frenatura elet.	0.021080257	1.0
EventId=357	Description=Sonde Temperatura Motore 3	0.025397900	1.0

Tabella 3.6: Regole di associazione tra EventId e Description

Nella Tabella 3.6 vengono riportate le regole di associazione tra gli attributi EventId e Description, rispettivamente il tipo di evento di guasto e la sua descrizione. Osservando la tabella notiamo che per tutte le regole abbiamo lo stesso valore di confidanza, ossia 1.0; questo valore indica che se si verifica EventId, allora nel 100% dei casi si verifica anche Description. Ciò che varia è il valore di supporto, ossia la % dei casi in cui i due eventi si verifichino insieme.

La regola EventId=115→ Description=Sforzo di trazione reso<0 presenta un valore di supporto maggiore mentre la regola con valore di supporto minore è EventId=101→ Description=Eccezione Richiesta apertura IR.

3.3.8 Regole di associazione tra EventId e Duration

Premises	Conclusions	Support	Confidence
EventId=116	Duration(seconds)=3	0.011556044	0.101111111
EventId=116	Duration(seconds)=2	0.012571960	0.11
EventId=116	Duration(seconds)=1	0.013587876	0.118888888
EventId=117	Duration(seconds)=0	0.011175076	0.149405772
EventId=117	Duration(seconds)=1	0.012952929	0.173174872

Tabella 3.7: Regole di associazione tra EventId e Duration

Nella Tabella 3.7 vengono riportate le regole di associazione tra gli attributi EventId e Duration.

La regola EventId=116→Duration(seconds)=3 indica che i due eventi, tipo di evento di guasto e la durata dell'interruzione del servizio, avvengono congiuntamente nell' 1.15%

dei casi. In più, se si verifica EventId=116, allora nel 10% dei casi si verifica anche Duration(seconds)=3.

Dalla regola EventId=116→Duration(seconds)=2 si evince i due eventi si verificano insieme nell' 1.26% dei casi, e se si manifesta EventId=116, allora nell'11% dei casi si manifesta anche Duration(seconds)=2.

Dal processo in RapidMiner si ottiene anche la regola EventId=116→Duration(seconds)=1, la quale mostra che i due eventi si manifestano insieme nell' 1.36% dei casi e oltre a ciò, se si verifica EventId=116, nel 11.9% dei casi si verifica anche Duration(seconds)=1. Quanto espresso in precedenza si verifica anche per tale regola: EventId=117→Duration(seconds)=0 ma con percentuali diverse di supporto (1.11%) e di confidenza (14.9%).

Un'ulteriore regola ottenuta è EventId=117→Duration(seconds)=1; essa mostra che i due eventi si verificano assieme nell'1.29% dei casi e se si verifica EventId=117, allora nel 17.3% dei casi si verifica anche Duration(seconds)=1.

In conclusione è evidente che la regola caratterizzata da un supporto e una confidenza maggiore è l'ultima analizzata : EventId=117→Duration(seconds)=1.

3.3.9 Regole di associazione tra Process e DCUState

Premises	Conclusions	Support	Confidence
Process=DCU1	DCUState=40030	0.016000677	0.377245508
Process=FLG2	DCUState=40000	0.020318320	0.25
Process=FLG2	DCUState=40010	0.012317981	0.1515625
Process=FLG2	DCUState=40030	0.012825939	0.1578125
Process=FLG2	DCUState=40050	0.010413139	0.128125

Tabella 3.8: Regole di associazione tra Process e DCUState

Nella Tabella 3.8 vengono riportate le regole di associazione tra gli attributi Process e DCUState, rispettivamente il tipo di processo in corso durante la rilevazione del guasto e lo stato della Digital Control Unit.

La regola Process=DCU1→DCUState=40030 indica che i due eventi, avvengono congiuntamente nell' 1.60% dei casi ed in più, se si verifica Process=DCU1, allora nel 37.7% dei casi si verifica anche DCUState=40030.

Dalla regola Process=FLG2→DCUState=40000 si evince i due eventi si verificano insieme nell' 2.03% dei casi, e se si manifesta Process=FLG2, allora nell'25% dei casi si manifesta anche DCUState=40000.

Nella Tabella 3.8 viene mostrata anche la regola Process=FLG2→DCUState=40010, la quale mostra che i due eventi si manifestano insieme nell' 1.23% dei casi e oltre a ciò, se si verifica Process=FLG2, nel 15.15% dei casi si verifica anche DCUState=40010. Quanto espresso in precedenza si verifica anche per tale regola: Process=FLG2→DCUState=40030 ma con percentuali diverse di supporto (1.28%) e di confidenza (15.78%).

Un'ultima regola, riguardante gli eventi sopra citati, è Process=FLG2→DCUState=40050; essa mostra che i due eventi si verificano assieme nell'1.04% dei casi e se si verifica Process=FLG2, allora nel 12.8% dei casi si verifica anche DCUState=40050.

In conclusione è evidente che la regola caratterizzata da un supporto maggiore è Process=FLG2→DCUState=40000 e quella caratterizzata da una confidenza maggiore è Process=DCU1→DCUState=40030.

3.3.10 Regole di associazione tra Process e ErrorCode

Premises	Conclusions	Support	Confidence
Process=DCU1	ErrorCode=3932517	0.025397900	0.598802395
Process=FLG2	ErrorCode=3014979	0.023493057	0.2890625
Process=FLG2	ErrorCode=3014980	0.021080257	0.259375
Process=VCUMON	ErrorCode=2949234	0.181510328	0.207530732
Process=VCUMON	ErrorCode=2949235	0.261725364	0.299244990
Process=VCUMON	ErrorCode=2949236	0.114290551	0.130674668

Tabella 3.9: Regole di associazione tra Process ed ErrorCode

Nella Tabella 3.9 vengono mostrate le regole di associazione tra gli attributi Process ed ErrorCode, ovvero tra il tipo di processo in corso durante la rilevazikone del guasto ed il codice errore verificato.

La regola con un valore di supporto più basso è Process=FLG2→ErrorCode=3014980, mentre quella con un valore di supporto più alto è Process=VCUMON→ErrorCode=2949235.

Invece la regola caratterizzata da un valore di confidenza più bassa è Process=VCUMON→ErrorCode=2949236 e quella caratterizzata da un valore di confidenza più alta è Process=DCU1→ErrorCode=3932517.

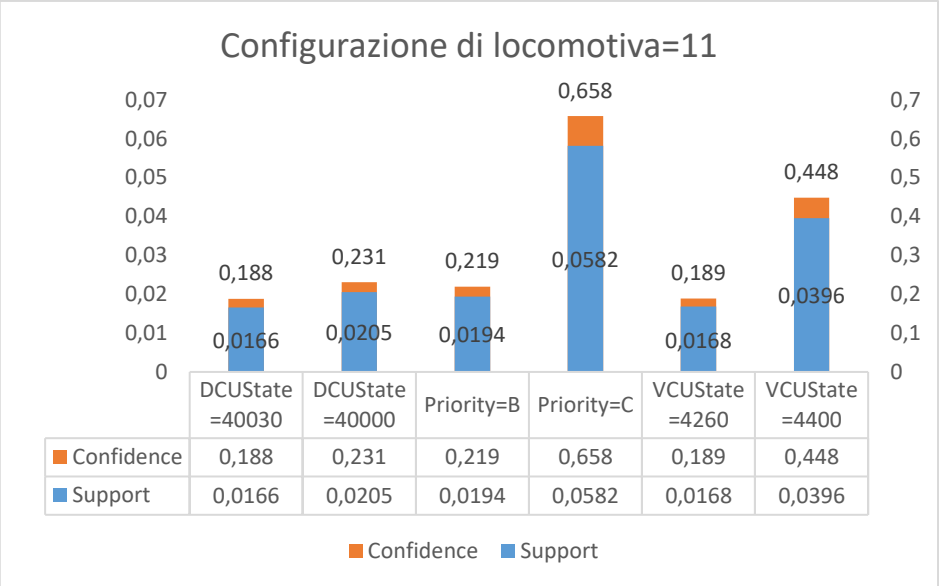

Grafico 3: Grafico relativo alla regola aventi come premessa Configurazione di locomotiva=11

Tutti i grafici contengono sull'asse delle ascisse la conclusione della regola di associazione, mentre sull'asse delle ordinate il supporto e la confidenza.

In particolare, le regole rappresentate nel Grafico 3 hanno come premessa l'attributo Configurazione di locomotiva=11 ed è evidente che la regola avente supporto e confidenza minore è quella avente come conclusione DCUState=40030, mentre quella avente un valore maggiore di supporto e confidenza è quella avente come conclusione Priority=C.

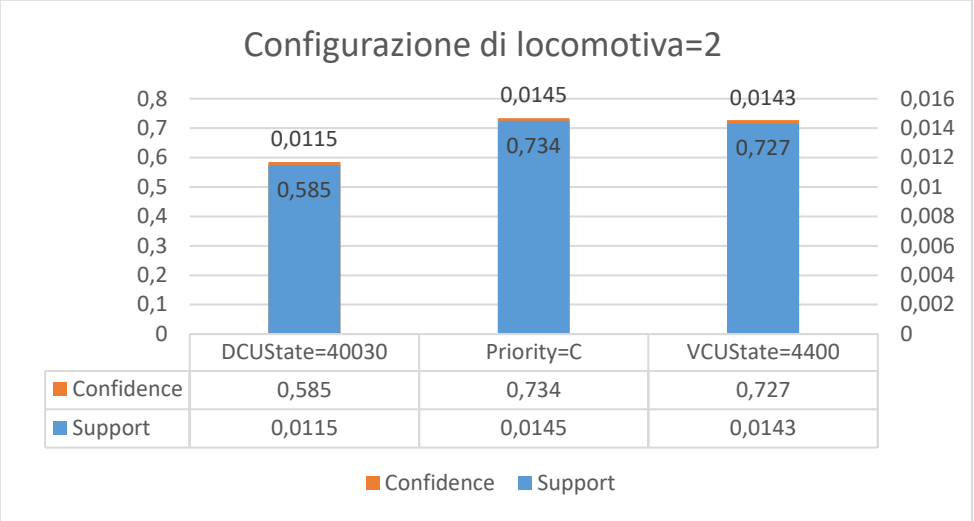

Grafico 3.1: Grafico relativo alla regola aventi come premessa Configurazione di locomotiva=2

Nel Grafico 3. 1 vengono mostrate le regole aventi come premessa l'attributo Configurazione di locomotiva=2 e la regola avente un valore di supporto e confidenza minore è quella mostrata nell'estrema sinistra riguardante DCUState=40030 come conclusione. Invece, la regola avente un supporto e una confidenza maggiore è quella posizionata al centro avente come conclusione Priority=C.

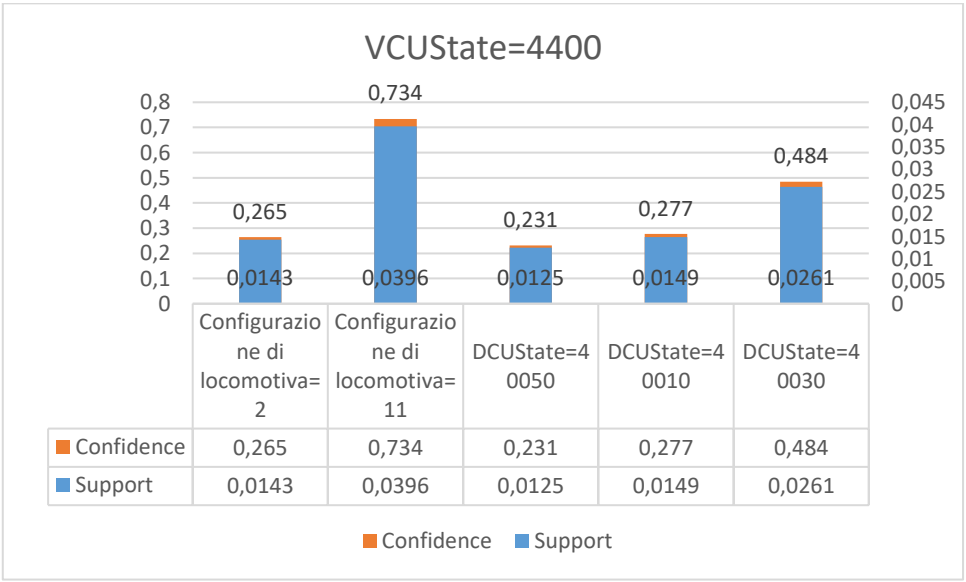

Grafico 3.2: Grafico relativo alla regola aventi come premessa VCUState=4400

Il Grafico 3.2 mostra le regole aventi come premessa l'attributo VCUState=4400. La regola avente come conclusione DCUState=40050 si è rilevata quella con un valore di

supporto e confidenza minore, mentre quella con un supporto e una confidenza maggiore è quella avente come conclusione Configurazione di locomotiva=11.

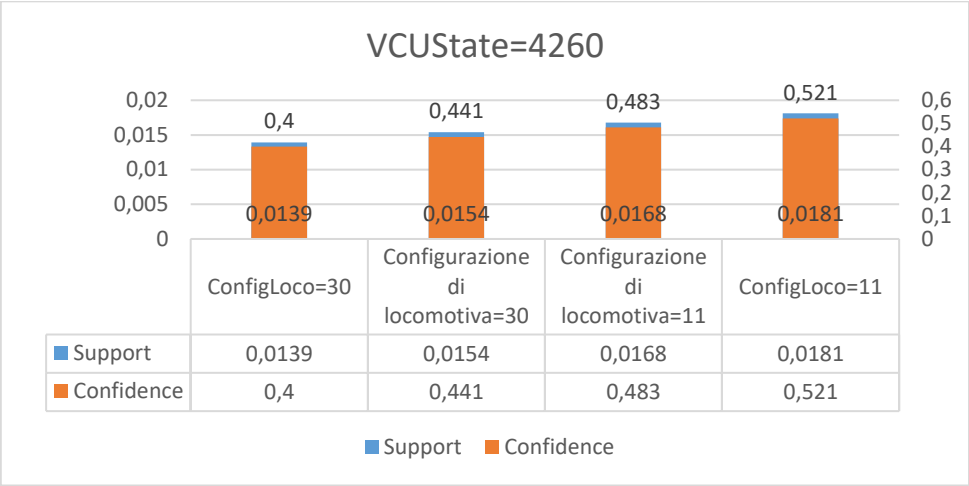

Grafico 3.3: Grafico relativo alla regola aventi come premessa VCUState=4260

Nel Grafico 3.3 sono rappresentate le regole aventi come premessa l'attributo VCUState=4260 e sono mostrate in ordine crescente in base al valore di supporto e confidenza. Quindi le regole all'estrema sinistra sono quelle con valori di supporto e confidenza minori (conclusione: ConfigLoco=30) mentre quelle all'estrema destra hanno valori di supporto e confidenza maggiori (conclusione: ConfigLoco=11).

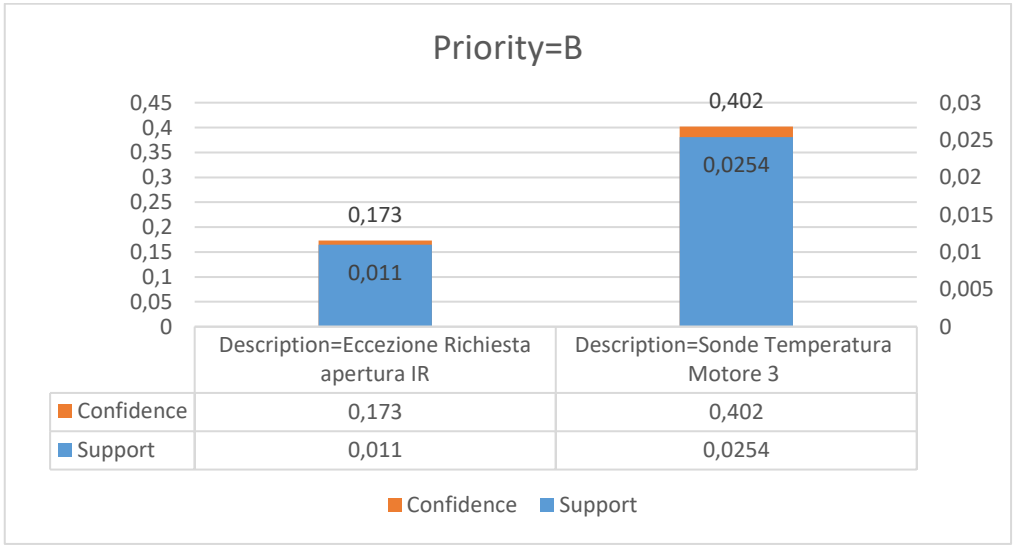

Grafico 3.4: Grafico relativo alla regola aventi come premessa Priority=B

Nel Grafico 3.4 sono illustrate le regole aventi come premessa l'evento Priority=B e l'evento Description= Eccezione Richiesta apertura IR è caratterizzato da un supporto e

una confidenza minore, mentre l'evento Description=Sonde Temperatura Motore 3 presenta un valore maggiore di supporto e confidenza.

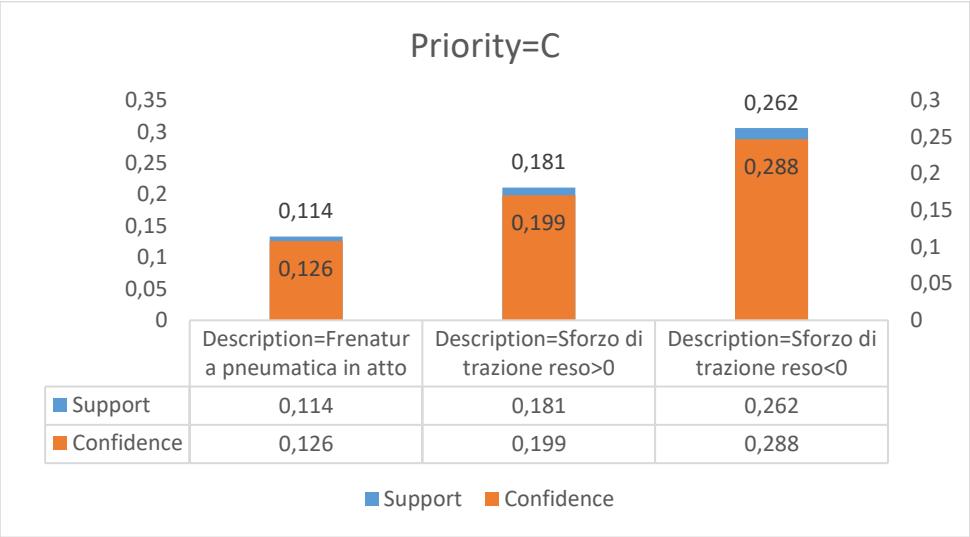

Grafico 3.5: Grafico relativo alla regola aventi come premessa Priority=C

Il Grafico 3.5 mostra le regole aventi come premessa l'evento Priority=C e l'evento presente in estrema sinistra è caratterizzato da un supporto e confidenza minore, mentre quello presente in estrema destra è caratterizzato da valori di supporto e confidenza maggiori.

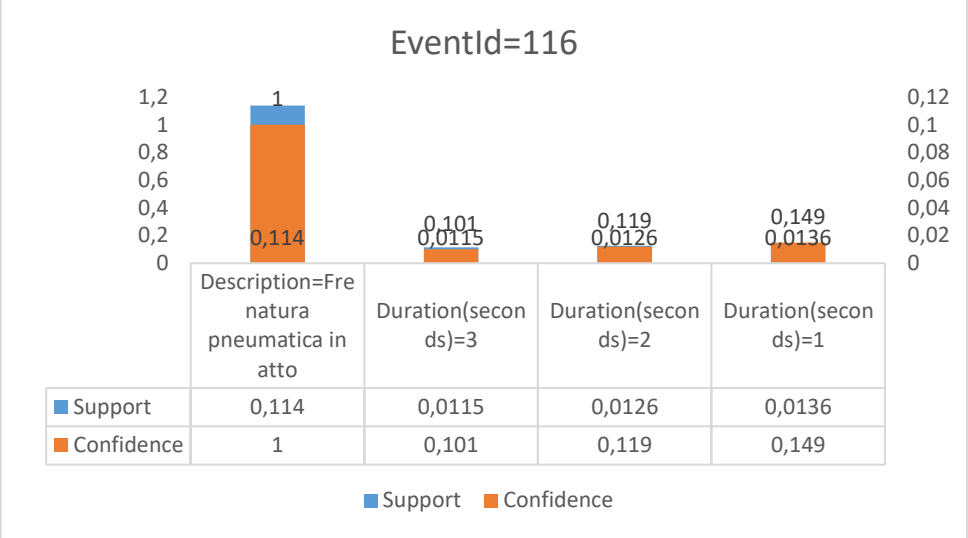

Grafico 3.6: Grafico relativo alla regola aventi come premessa EventId=116

Nel Grafico 3.6 vengono riportante le regole aventi come premessa l'evento EventId=116; da esso si evince che la regola caratterizzata da un supporto e una confidenza minore è quella avente in conclusione l'evento Duration(seconds)=3, mentre

quella caratterizzati da valori di supporto e confidenza maggiori hanno come conclusione l'evento Description=Frenatura pneumatica in atto.

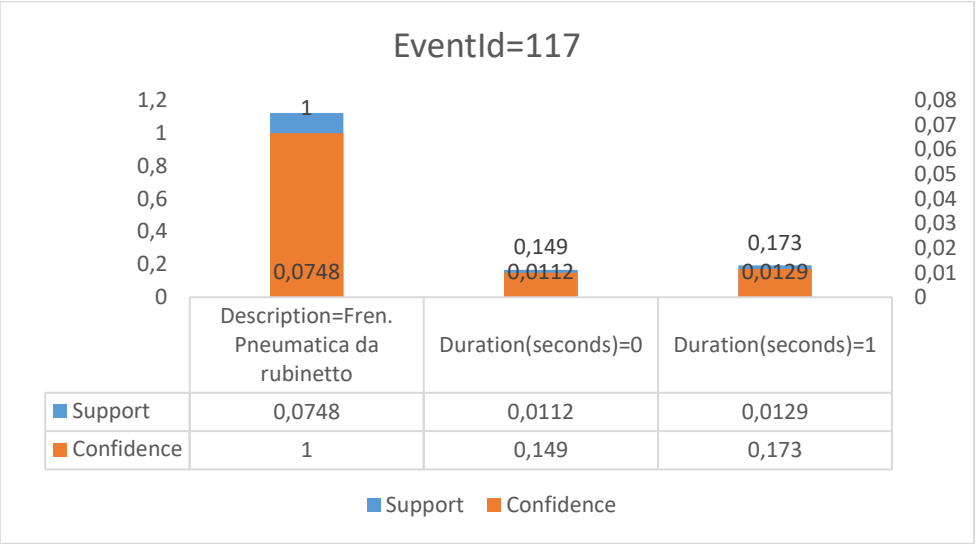

Grafico 3.7: Grafico relativo alla regola aventi come premessa EventId=117

Il Grafico 3.7 mostra le regole aventi come premessa l'attributo EventId=117; nell'estrema sinistra è presente la regola caratterizzata da un supporto e una confidenza maggiore avente in conclusione l'evento Description= Fren. Pneumatica da rubinetto. Al centro, invece, è rappresentata la regola avente in conclusione l'evento Duration(seconds)=0, caratterizzata da valori di supporto e confidenza minori.

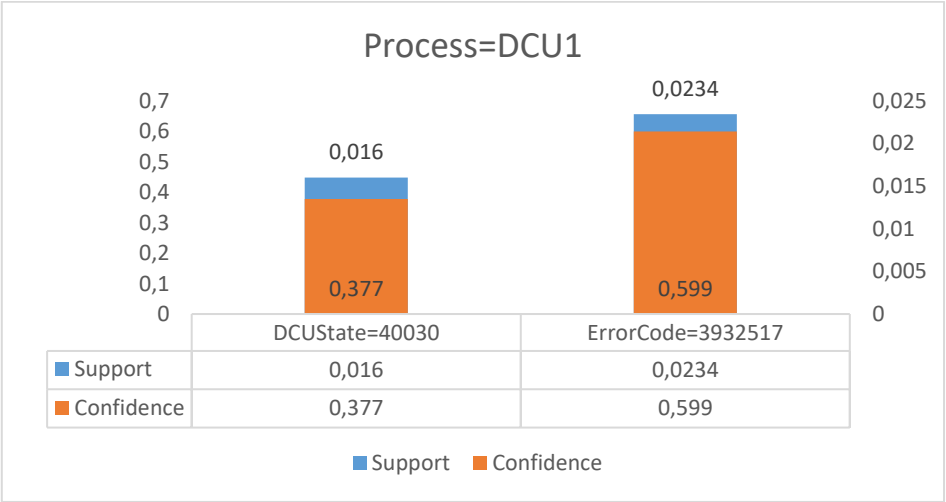

Grafico 3.8: Grafico relativo alla regola aventi come premessa Process=DCU1

Il Grafico 3.8 presenta le regole aventi l'evento Process=DCU1 come premessa, mentre in conclusione abbiamo a destra l'evento caratterizzato da valori di supporto e confidenza

più bassi (DCUState=40030) e a sinistra l'evento caratterizzato da valori più elevati (ErrorCode=3932517).

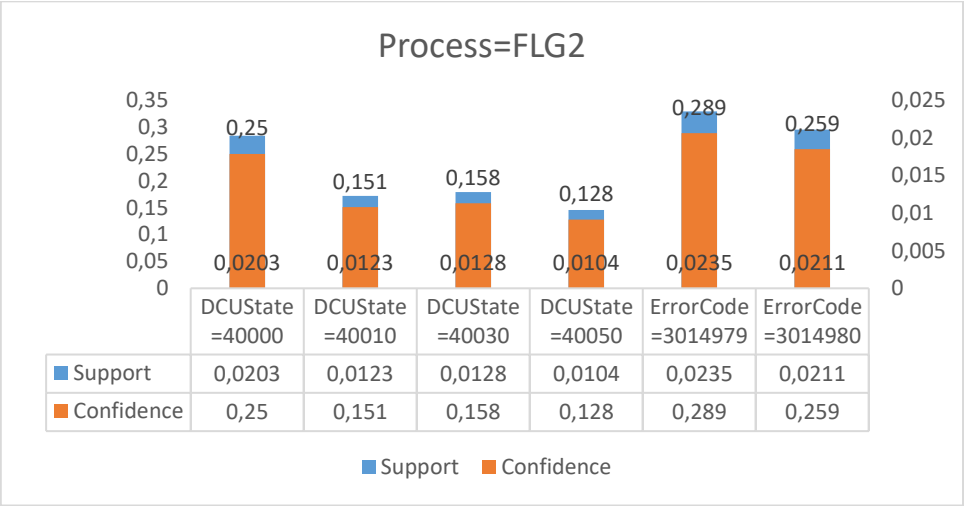

Grafico 3.9: Grafico relativo alla regola aventi come premessa Process=FLG2

Nel Grafico 3.9 vengono riportate le regole aventi come premessa l'attributo Process=FLG2. In conclusione abbiamo vari eventi; tra essi quello caratterizzato da valori di supporto e confidenza più bassi è DCUState=40050, mentre quello caratterizzato da valori più elevato di supporto e confidenza è ErrorCode=3014979.

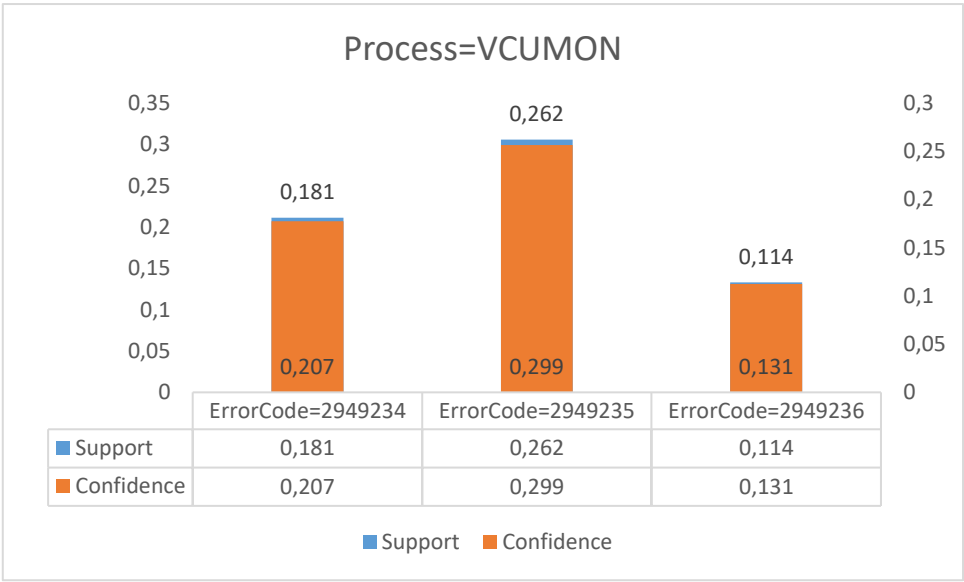

Grafico 3.10: Grafico relativo alla regola aventi come premessa Process=VCUMON

Il Grafico 3.10 mostra le regole che hanno nella premessa l'evento Process=VCUMON. L'evento, in conclusione, avente valori minori di supporto e confidenza è ErrorCode=2949236, mentre quello avente valori maggiori è ErrorCode=2949235.

CAPITOLO 4

CONCLUSIONI

L'obiettivo di questa tesi è stato l'analisi di dati relativi ai guasti di macchine per trazione elettrica in ambito ferroviario con lo scopo di trasformarli in informazione. Si è valutato di utilizzare l'algoritmo FP-Growth attraverso l'uso del software RapidMiner e di creare le regole di associazione. Dal lavoro di analisi si sono ottenuti vari attributi ma quelli dominanti e più significativi sono DCUState, VCUState, Configurazione di locomotiva, Priority, EventId, Duration, Process ed ErrorCode; quindi ho analizzato le regole ottenute tra loro.

Dalle analisi effettuate con questo algoritmo è emerso che la regola che presenta maggiore probabilità in cui se si verifica un evento, se ne verifichi anche un altro è quella riguardante gli eventi EventId e Description e abbiamo due regole caratterizzate da più probabilità che due eventi che si verifichino insieme, ossia Process=VCUMON→ErrorCode=2949235 e Priority=C→Description= Sforzo di trazione reso<0. Invece, la regola avente una minore probabilità che se si verifica un evento, si verifica anche l'altro è quella riguardante gli eventi EventId=116 e Duration(seconds)=3 e gli eventi per cui si ha meno probabilità che si verifichino insieme sono Process=FLG2 e DCUState=40050. Quindi, da tutto ciò, si è dimostrato che si possono estrarre informazioni utili dai risultati ottenuti.

L'analisi, in futuro, potrebbe essere svolta utilizzando nuovi algoritmi e magari anche nuovi software per notare se i risultati ottenuti convergono; in quanto l'analisi svolta, spiegata nel terzo capitolo, è stata esplorativa quindi non si è certi che sia il metodo più adatto per lo studio di dati di guasto riguardanti locomotive a trazione elettrica.

Per concludere, i risultati ottenuti e quelli ottenibili tramite altri algoritmi, potrebbero portare in futuro ad una vera e propria raccolta, schematizzazione di questi dati, aggiornati il più frequente possibile in modo da informare, in tempo reale, tutti coloro che usufruiscono dei mezzi di trasporto trattati in questa tesi.

BIBLIOGRAFIA

Ghomi, H., Fu, L., Bagheri, M., & Miranda-Moreno, L. F. (2017). Identifying vehicle driver injury severity factors at highway-railway grade crossings using data mining algorithms. *2017 4th International Conference on Transportation Information and Safety, ICTIS 2017 - Proceedings*, 1054–1059. https://doi.org/10.1109/ICTIS.2017.8047900

Guarracino, M. (2007). *Regole associative*.

Mirabadi, A., & Sharifian, S. (2010). Application of association rules in Iranian Railways (RAI) accident data analysis. *Safety Science*, *48*(10), 1427–1435. https://doi.org/10.1016/j.ssci.2010.06.006

Qian, K., Yu, L., & Liu, Y. (2019). FHI: A Fault Intensity-based Hierarchical Association Analysis Model for Mining Fault Database of Railway OCS. *Proceedings of IEEE 14th International Conference on Intelligent Systems and Knowledge Engineering, ISKE 2019*, 616–621. https://doi.org/10.1109/ISKE47853.2019.9170406

RapidMiner Studio Manual. (n.d.).

SITOGRAFIA

1. *Qual è il modo più sicuro di viaggiare?* (n.d.). Retrieved January 28, 2021, from https://www.agi.it/fact-checking/incidenti_treni_aerei_navi_italia-7023150/news/2020-02-07/
2. *Data mining, cos'è? - Strumenti, applicazioni e rischi*. (n.d.). Retrieved December 29, 2020, from https://www.intelligenzaartificiale.it/data-mining/
3. *RapidMiner | Best Data Science & Machine Learning Platform*. (n.d.). Retrieved January 26, 2021, from https://rapidminer.com/
4. Subrizi, C. (2008). *Introduzione a Duchamp*. http://www.loc.gov/catdir/toc/casalini07/08624038.pdf
5. *UN MODELLO PER L'ANALISI DELLE CAUSE DEGLI INCIDENTI FERROVIARI Premessa*. (n.d.). Retrieved January 28, 2021, from https://webcache.googleusercontent.com/search?q=cache:AJtsZRjJekUJ:https://www.ansf.gov.it/documents/20142/177538/AA.VV.Unmodelloperlanalisidellecausedegliincidentiferroviari.pdf/ccb70f3b-2a50-2381-e6dc-81eb6192975b%3Fdownload%3Dtrue+&cd=2&hl=it&ct=clnk&gl=it

RINGRAZIAMENTI

Giunta alla fine di questo percorso, ci tengo a ringraziare le persone che mi sono state vicine in questo percorso.

Un sincero ringraziamento va al Professore Maurizio Bevilacqua che è sempre stato disponibile e pronto ad aiutarmi ma un grazie immenso va alla mia correlatrice Ing. Sara Antomarioni che con la sua gentilezza e incredibile disponibilità è sempre stata in grado di ascoltarmi e consigliarmi nel modo giusto.

Un grazie va soprattutto a me stessa, per non essermi mai arresa nei momenti di difficoltà, per essermi posta degli obiettivi ed esser riuscita a raggiungerli, per la mia volontà e determinazione che mi hanno permesso di raggiungere questo traguardo che sembrava irraggiungibile.

Un grazie ai miei genitori che mi hanno dato la possibilità di intraprendere questo percorso e di essermi sempre stati vicini. Spero che i loro sacrifici siano stati, oggi, ripagati e che siano orgogliosi di me.

In particolare a mia sorella Giulia per avermi sempre sostenuta, dato consigli nei momenti in cui ne avevo bisogno, per aver sopportato i miei sbalzi d'umore, per avermi rimproverata ogni volta che facevo qualcosa di sbagliato, per avermi ascoltata, per le nostre litigate.

Un grazie di cuore alle mie amiche di sempre, Chiara, Anna, Simona e Francesca, con le quali ho condiviso anni di studio, passeggiate, feste, vacanze, sorrisi e lacrime. Per qualsiasi cosa sapevo di poter contare su di loro. Grazie per aver sempre creduto in me, anche quando ero io stessa a non crederci. Grazie per essere state sempre al mio fianco.

Un grazie a Mariolina, la mia amica di liceo, che nonostante non ci vediamo spesso, ci siamo sempre sentite e supportate a vicenda, soprattutto nell'ultimo periodo.

Un grazie speciale alla mia coinquilina Francesca, nonché mamma e sorella universitaria. Mi ha sostenuta nell'affrontare ogni difficoltà, mi ha consigliato nelle scelte difficili, mi ha asciugato le lacrime nei momenti di sconforto. Grazie per i momenti passati insieme, per avermi sempre aspettata per il pranzo, per le serate sedute sulla poltrona a vedere la tv o sul mio letto per confidarci e sfogarci, per i passaggi in macchina, per avermi aiutata

con la tesi. Non finirò mai di ringraziarti. Mi mancherà tutto di noi, ma sono certa che non ci separeremo e che ci saremo sempre l'una per l'altra.

Ringrazio le mie coinquiline, Francesca e Marta, per le risate, per le avventure con Ciccio, per le serate passate insieme, per avermi fatto trascorrere questi tre anni nel migliore dei modi.

Un grazie a tutti i miei compagni di università, Chiara, Sara, Maria Vittoria, Martina, Davide e Alfiero per i momenti passati insieme, per i pomeriggi trascorsi in aula studio, per avermi sempre aiutata quando ne ho avuto bisogno, per i sabati sera in pizzeria e al duomo. Avete reso questo percorso indimenticabile.

Un grazie ai miei cugini Alessia e Nicola, Marco e Floriana, Mariangela per essere sempre stati orgogliosi di me e per avermi fatto sentire la loro "Dottoressa" anche quando questa avventura era appena iniziata.

Ai miei nonni e al mio angelo custode, che oggi non possono essere qui con me, ma che spero che mi guardino da lassù e che siano orgogliosi di me.

Non possono mancare i miei due cuori, Enzo ed Emma, che mi hanno sempre tenuto compagnia con le videochiamate quando ero a Fermo; mi hanno dato forza con i loro sorrisi, abbracci e carezze. Mi riempivano il cuore con i loro "TI VOGLIO BENE", "MI SEI MANCATA". Senza la loro spensieratezza e vivacità questo splendido traguardo sarebbe stato diverso.

Un grazie va a Martina, per avermi dato supporto quando ne ho avuto bisogno, per avermi regalato momenti di felicità, soprattutto durante il periodo di quarantena rendendo tutto più serenamente, e ad Umberto per il suo modo di sdrammatizzare sempre i miei momenti di ansia e di nervosismo con le sue fantastiche "battute".

Un ultimo grazie voglio dirlo a Peppino e Filomena, che mi hanno considerata come se fossi una loro nipote, grazie per i caffè pomeridiani, grazie per avermi sempre sostenuta.