UNIVERSITÀ POLITECNICA DELLE MARCHE

Facoltà di Ingegneria
Corso di Laurea in Ingegneria Meccanica
Dipartimento di Ingegneria Industriale e Scienze Matematiche

LE ASSOCIATION RULES NELL'ANALISI DEI DATI DI GUASTO FERROVIARIO

THE ASSOCIATION RULES IN THE ANALYSIS OF RAILWAY FAILURE DATA

RELATORE:
Prof. Ing. Maurizio Bevilacqua

TESI DI LAUREA DI:
Giuseppe Memmo

CORRELATORE:
Dott. Ing. Sara Antomarioni, PhD

Anno Accademico 2020-2021

A mamma e papà, la mia forza

Sommario

Indice delle Figure

Indice delle Tabelle

1 Introduzione

La disponibilità di nuove tecnologie e di considerevoli quantitativi di informazioni sono alcuni tra i potenziali elementi che possono fare la differenza nella rivoluzione delle *strategie manutentive* nel ventunesimo secolo. Attraverso lo sviluppo e l'integrazione di sensori intelligenti e connessi, il trasporto ferroviario sta diventando più puntuale, più funzionale e in grado di garantire standard di manutenibilità sempre più elevati e a basso costo, superando le carenze che lo avevano caratterizzato nei decenni precedenti.

A tal proposito, la *Manutenzione Predittiva* si sta rivelando una strategia chiave per massimizzare l'efficienza e l'efficacia dell'intero sistema ferroviario, in quanto permette di intuire in anticipo la necessità di eseguire interventi di manutenzione, prevedendo l'insorgere di problemi. Alla base ci sono le logiche per la diagnosi automatica, anche dette *Algoritmi Diagnostici*, sviluppate e validate dal team di tecnici della casa madre, grazie alle quali è stato possibile replicare la conoscenza inerente alla diagnosi dei guasti in un sistema informatico centrale.

Il lavoro di tesi prende in esame la flotta di veicoli adibiti al trasporto regionale, focalizzandosi sui *Diagnostic Data Record* delle locomotive E464: attraverso il *mining* delle *Regole di Associazione* si cerca, in via sperimentale, di estrarre dai dati a disposizione relazioni interessanti, con l'obiettivo di fornire un supporto valido alla realizzazione di nuovi algoritmi diagnostici per la manutenzione predittiva. Molte delle informazioni presenti sui dati, infatti, non sono direttamente evidenti, in quanto le analisi effettuate con metodi tradizionali, utilizzando un approccio *knowledge-based*, possono richiedere diverse settimane per mostrare risultati utili e di fatto lasciano una larga parte di dati non analizzati.

L'indagine condotta si basa su un approccio *data-driven*: si analizzano i dati senza il supporto delle esperienze pregresse acquisite dai progettisti e manutentori, con l'obiettivo di identificare delle relazioni tra elementi apparentemente indipendenti e disgiunti, estraendo modelli causali e schemi ricorrenti precedentemente sconosciuti.

2 La mobilità e il trasporto ferroviario

La mobilità è una componente irrinunciabile della società moderna, che risponde a una crescente propensione al movimento non solo per i piccoli spostamenti quotidiani ma anche per quelli a lungo raggio.

Al giorno d'oggi, il sistema delle infrastrutture e il trasporto pubblico locale in particolare, pone tra i suoi obiettivi la migliore qualità della vita e la riduzione dell'impatto ambientale: in tal senso, la circolazione ferroviaria ne costituisce un enorme contributo grazie alla sua evoluzione itinerante, che la rende un mezzo attuale, sostenibile e con ulteriori margini di sviluppo.

La nascita della ferrovia coincide con l'invenzione della locomotiva a vapore e si contestualizza nell'Inghilterra degli inizi del *XIX secolo,* in un clima di rivoluzione e nuove scoperte tecnologiche: la possibilità di attraversare spazi sempre più vasti e in tempi brevi favorì lo spostamento di persone e fu presa a modello anche in altri Paesi, compresa la nostra penisola.

In Italia il primo brevissimo tratto fu inaugurato nel *1839* nel Regno delle Due Sicilie e collegava Napoli-Portici: fu un evento pioneristico, ma sostanzialmente simbolico sia per le ridotte dimensioni del percorso sia per l'assenza di finalità economiche (Viola 2009).

Figura 2-1: Il tratto ferroviario Napoli-Portici: pittura di Salvatore Fergola (1799-1877) - Museo Nazionale di San Martino, Napoli

Solo verso la metà degli anni *Quaranta* dell'*Ottocento*, il peso economico e politico della rete si fece più evidente e furono delineate le prime linee della piccola mappa che collegava alcune città del nord.

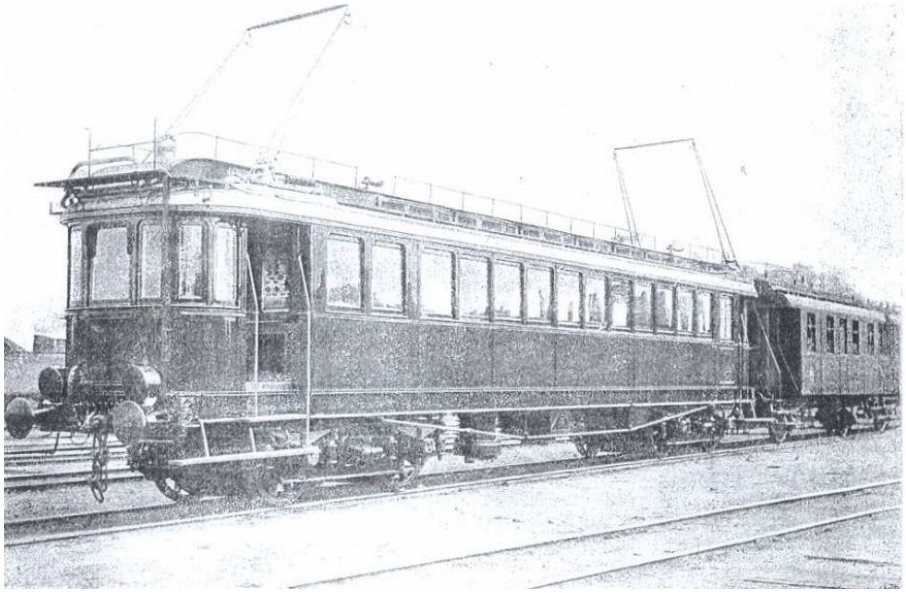

Figura 2-2: Elettromotrice di prima classe della Ferrovia della Valtellina del 1902 (Guidi Buffarini 2007)

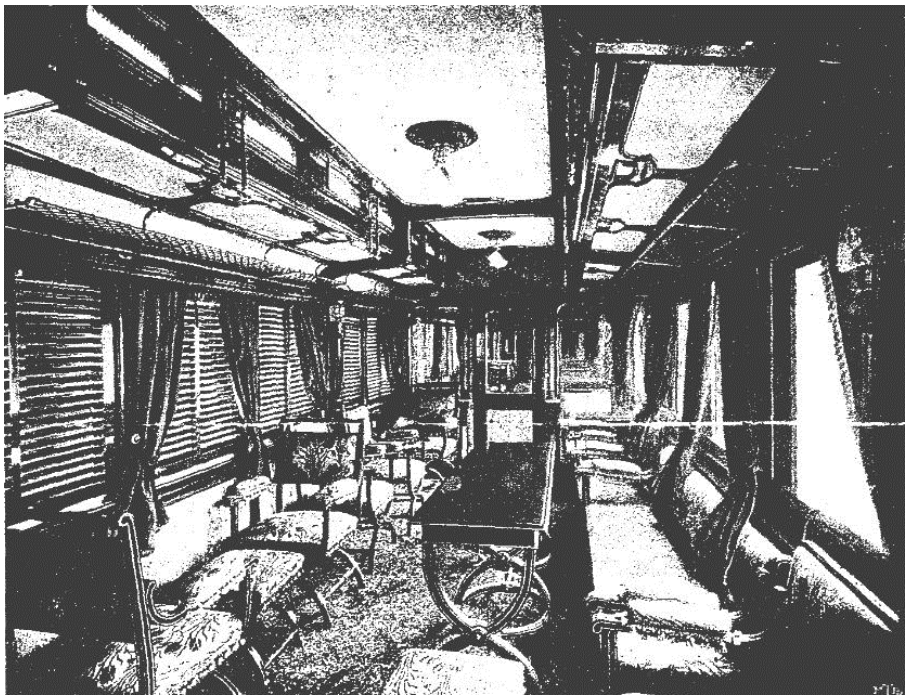

Figura 2-3: Arredamento dell'elettromotrice di prima classe della Ferrovia della Valtellina del 1902 (Guidi Buffarini 2007)

Nell'arco di un quarantennio, la geografia ferroviaria alpina fu completamente rivoluzionata, inserendo l'Italia nella rete infrastrutturale europea: studi e ricerche in questo campo contribuirono all'ulteriore maturazione del ceto tecnico italiano dapprima rispetto ai tratti di montagna e poi verso altri tipi di percorsi.

Se, inizialmente, i tracciati si irradiavano dalle città verso le zone limitrofe, solo in un secondo momento si svilupparono dei collegamenti tra le varie singole reti fino a portare, nei primi del *Novecento,* all'attuale configurazione europea.

Con l'avvento del *XX secolo*, il trasporto ferroviario, che fino ad allora era dominante su tutti fronti, conservò il suo primato nel settore delle merci, mentre nell'ambito della mobilità di persone fu gradualmente surclassato dalla crescente concorrenza di mezzi alternativi, come quello stradale: l'abbondante disponibilità di energia di origine fossile e il continuo progresso dei motori a scoppio incentivò lo spostamento su strada contro costi di infrastruttura e di carico maggiori appannaggio della ferrovia.

Per questo motivo, gli obiettivi principali da perseguire in prima battuta furono: rafforzare ed estendere l'Alta velocità ferroviaria nazionale e potenziare la rete regionale per far tornare l'impresa competitiva su tutti i fronti.

Queste tematiche sono sempre attuali e si arricchiscono di input e problematiche contemporanee: i progetti futuri continuano a focalizzarsi sullo sviluppo sostenibile e sull'attrattività del Paese, mettendo al centro la soddisfazione delle esigenze dei viaggiatori e della logistica e aumentando la connettività del sistema Italia.

2.1 L'evoluzione del sistema ferroviario in Italia

Nel *1905*, con la *Legge del 22 aprile n.137*, lo Stato acquisì la proprietà e l'esercizio della maggior parte delle linee ferroviarie nazionali, fino a quel momento prerogativa di imprese private. Così, il 1° luglio dello stesso anno nacque l'*Azienda Unitaria delle Ferrovie dello Stato*: per la prima volta si parla non solo di un unico operatore nella gestione, ma anche di sperimentare il criterio dell'unificazione, dai fabbricati di stazione allo stile delle carrozze.

Da questa fusione furono escluse le linee locali, che continuarono a essere attive sulla base di concessioni: sono le attuali reti regionali, chiamate *ferrovie concesse* in quanto affidate alle Regioni.

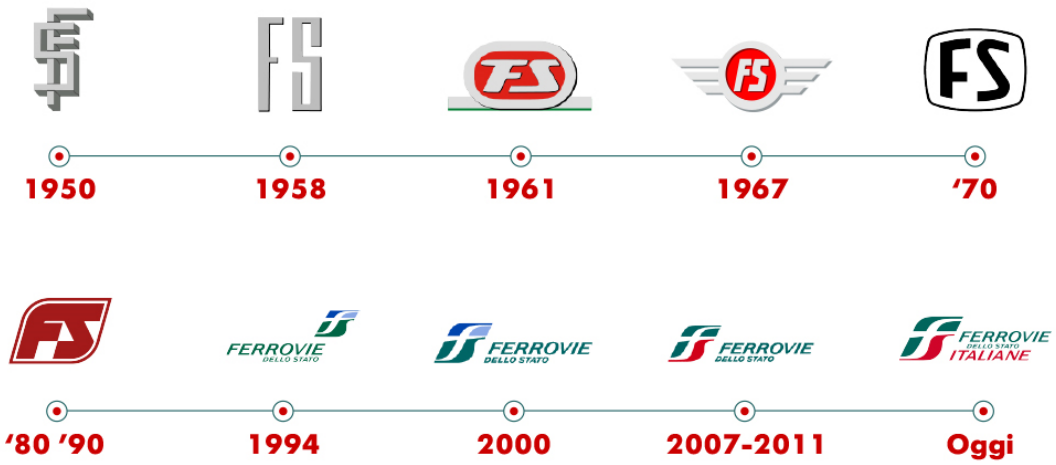

Figura 2-4: Evoluzione del logo istituzionale

Il periodo tra il *1986* e il *1992* fu caratterizzato da profonde trasformazioni strutturali e organizzative, che implicarono una riduzione del personale a meno della metà, la creazione di nuove divisioni e società controllate e infine riassegnazioni di dipendenti e mezzi: nel *1986 l'Azienda Unitaria delle Ferrovie dello Stato* si trasforma in *Ente Ferrovie dello Stato*, un ente pubblico ed economico, e nel *1992* diviene *Ferrovie dello Stato – Società di trasporti e servizi per azioni*, con un unico azionista quale il *Ministero dell'economia e delle finanze* e nuove società come

T.A.V. Spa, *ITALFERR-SIS-TAV Spa* e *METROPOLIS Spa* (Cambini e Buzzo Margari 2005).

Con questa configurazione, l'assetto della rete ferroviaria nazionale inizia ad allinearsi con quanto previsto dall'ordinamento europeo, caratterizzato dalla separazione tra gestione dell'impianto e svolgimento del servizio.

In tal senso, il *15 dicembre 2000*, l'impresa si trasformò in *Ferrovie dello Stato Holding S.r.l.*, e il processo di separazione si concluse nel *2001* con la creazione della *Rete Ferroviaria italiana – RFI*, società titolare della concessione sessantennale della rete, e di *Trenitalia*, società che effettua il trasporto e affidataria dei contratti di servizio pubblico viaggiatori e merci.

In ogni caso, il complesso delle strutture occorrenti per l'espletamento del servizio ferroviario permane sussidiato dallo Stato, in quanto *RFI* non ha il vincolo di coprire integralmente i costi, non avendo quindi responsabilità economica, e, come reso noto dal *Ministero delle infrastrutture e della mobilità sostenibile*: "Il Ministero svolge numerose attività riguardanti la rete ferroviaria con l'obiettivo di renderla più efficiente e funzionale. È importante precisare che, dal punto di vista operativo, il Ministero non realizza direttamente le infrastrutture né effettua la loro manutenzione, ma affida questi compiti al *Gruppo Ferrovie dello Stato Italiane – RFI S.p.A.*. Attraverso un contratto di programma, il Ministero indica le attività da realizzare e controlla l'utilizzo degli investimenti pubblici destinati ai progetti per la manutenzione e la realizzazione delle ferrovie" (Ministero s.d.).

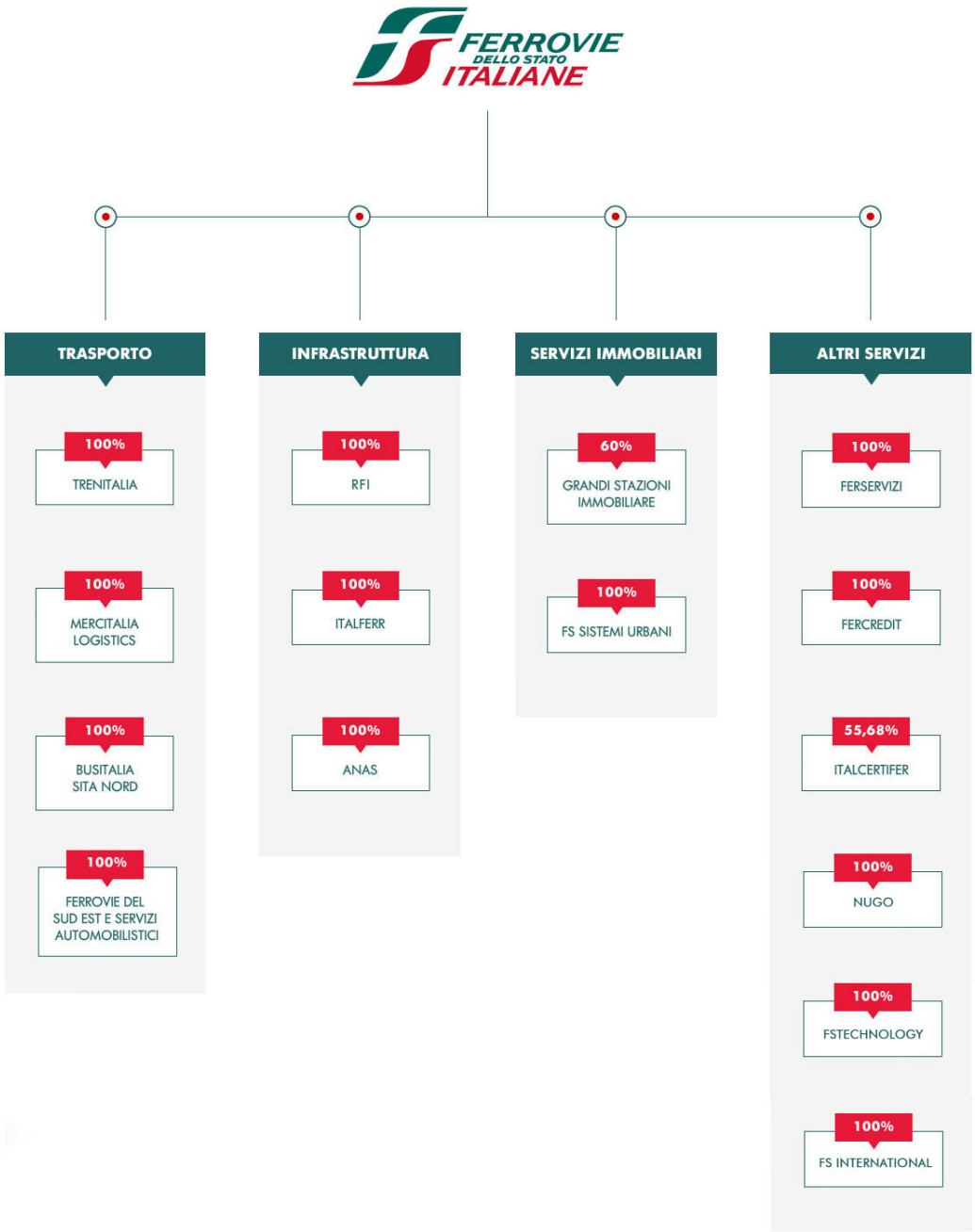

Figura 2-5: Il gruppo Ferrovie dello Stato italiane

2.2 Il trasporto ferroviario passeggeri

La mobilità di viaggiatori si suddivide in due insiemi in funzione della richiesta: il *trasporto locale*, connesso al contesto lavorativo del posto o legato alle esigenze dei pendolari, e quello a *lunga percorrenza*, in cui la circolazione su rotaia ha come sostituti il trasporto individuale o collettivo su gomma e quello aereo.

In particolare, per il primo segmento vi è una forte dipendenza della domanda dalle condizioni al contorno e dalla politica del posto: il tipo di urbanizzazione (ad esempio la concentrazione della popolazione lungo la rete); l'estensione del pendolarismo; l'ampiezza e la cogenza delle misure restrittive al traffico privato (ad esempio la presenza di zone a traffico limitato o l'uso di targhe alterne) e la disponibilità di servizi accessori come i parcheggi di interscambio possono influire sui tempi complessivi del percorso, sulla puntualità e più in generale sulla qualità del servizio offerto.

Poiché il viaggio in treno richiede quasi sempre l'utilizzo di altre infrastrutture per completare lo spostamento, i tempi di attesa per lo scambio intermodale possono scoraggiare l'utilizzo di questo mezzo per brevi distanze, anche se si usufruirebbe di velocità relativamente più elevate.

Per tutti questi aspetti, è opportuno introdurre nell'analisi economica il concetto di *costo generalizzato*, che include, oltre alla componente monetaria, anche quella relativa al tempo di trasporto per ogni modalità di spostamento (Bentivogli e Panicara 2012).

Nell'economia dei trasporti, il costo generalizzato è la somma dei costi monetari e non percepiti dall'utente per effettuare un trasferimento: nel primo caso si intendono le tariffe per i mezzi pubblici o il costo del carburante, della sosta e del pedaggio autostradale per chi usa il proprio veicolo; nel secondo caso si fa riferimento ai disagi e ai costi opportunità supposti proporzionali al tempo speso nello spostamento.

Per monetizzarli, si utilizza un ulteriore parametro, detto *valore del tempo*, che in quanto variabile soggettiva viene in genere relazionata a grandezze

socioeconomiche statistiche della persona (reddito, età, sesso, occupazione) e a elementi prettamente trasportistici, come lo scopo del viaggio.

Il monitoraggio del trasporto ferroviario passeggeri ha portato in evidenza:

- notevoli oscillazioni del flusso nel corso della giornata, della settimana e nei diversi mesi dell'anno;
- un movimento pendolare di tipo radiale, cioè verso e dal centro urbano più importante e concentrato nelle ore di punta;
- l'alta variabilità della domanda nell'arco della giornata e l'elevato investimento minimo necessario per supplire all'offerta tendono a determinare capacità in eccesso o affollamento di passeggeri;
- una domanda sempre più individualizzata a causa della minore uniformità degli orari di lavoro e del tempo libero che porta alla scelta del trasporto su auto privata, tendenza incentivata dalla riduzione dei costi di acquisto del mezzo e dalla dislocazione dei centri commerciali nelle aree extraurbane.

In risposta a questo inquadramento territoriale, le politiche pubbliche cercano di incentivare gli spostamenti su rotaia perseguendo due principali obiettivi.

Il primo è garantire il diritto alla mobilità di tutti i cittadini, orientando verso il treno anche gli abitanti più abbienti e quindi meno sensibili al costo del viaggio, motivandoli attraverso un miglioramento qualitativo del servizio.

Il secondo è disincentivare l'uso dell'automobile con l'introduzione o l'aumento di imposte e tasse sul carburante o vincoli alla circolazione per raggiungere un minore grado di inquinamento, congestione, incidentalità e costi di usura delle strade (Bentivogli e Panicara 2012).

Oltre a ciò, è importante considerare il fattore di impatto ambientale: la lotta al cambiamento climatico è oggi al centro dell'agenda politica europea, che prevede un piano di investimenti per la realizzazione di un programma

strategico in grado di rispondere all'emergenza climatica e, al tempo stesso, di rilanciare l'economia e agevolare la vita dei cittadini.

All'interno di questo scenario, la rete del ferro rappresenta la spina dorsale di un sistema incentrato sul trasporto pubblico con linee prioritarie e un servizio efficiente, integrato a un insieme di percorsi ciclabili e pedonali (Legambiente 2019).

2.2.1 La rete regionale

Il trasporto ferroviario regionale è un sottoinsieme del *Trasporto Pubblico Locale – TPL*, che regolamenta viaggi di piccola distanza, a livello territoriale, principalmente per motivi di lavoro o studio.

Questo tipo di mobilità risponde alla necessità di offrire un servizio pubblico, senza imposizione di vincoli, cioè rispettando delle frequenze e delle tempistiche prestabilite, con copertura di percorsi rapidi e a prezzi sostenibili.

La concessione è responsabilità delle Regioni, che stipulano dei *Contratti di Servizio* con le imprese ferroviarie: vengono definite le quantità, i costi e gli standard di qualità dei servizi erogati; inoltre, in caso di asimmetrie informative, gli accordi prevedono anche gli incentivi necessari per accrescere l'efficienza e un'offerta migliore.

Nei servizi ferroviari viaggiatori, l'output è misurato in *passeggeri-km* (somma dei prodotti del numero di passeggeri trasportati per le relative percorrenze espresse in chilometri) mentre il servizio fornito dalle imprese si concretizza in una determinata disponibilità di *treni-km* (sommatoria dei prodotti del numero dei treni, inteso come numero dei posti, per le relative percorrenze espresse in chilometri).

Segue che i costi variabili sono legati più al numero di treni utilizzati, ai posti disponibili, ai chilometri complessivamente percorsi e ai tempi di percorrenza

che ai passeggeri trasportati, i quali incidono principalmente sui ricavi (Bentivogli e Panicara 2012).

Segue che l'amministrazione regionale stabilisce un corrispettivo economico per l'erogazione di un quantitativo di *treni/km* e richiede che i proventi del traffico concorrano a coprire almeno il *35%* dei costi operativi, mentre il rimanente *65%* dovrebbe essere assicurato dal corrispettivo pagato dalle Regioni, le quali possono imporre determinati indici di qualità relativi a pulizia, comfort, informazione e puntualità delle corse, prevedendo delle sanzioni nel caso di mancato rispetto di questi ultimi (Cambini e Buzzo Margari 2005).

Ogni giorno circa *5.700.000* persone prendono il treno per spostarsi di città in città e sono *2.894* le vetture in servizio gestite da diversi concessionari come Trenitalia, che dirige il trasporto su rotaia in esclusiva nel Lazio, nelle Marche, in Molise, In Sicilia e in Valle D'Aosta (Legambiente 2019).

Nelle altre regioni, a Trenitalia si affiancano una ventina di gestori di proprietà prevalentemente locale e provinciale, in alcuni casi a essa consorziati rafforzandone il potere di mercato (Bentivogli e Panicara 2012).

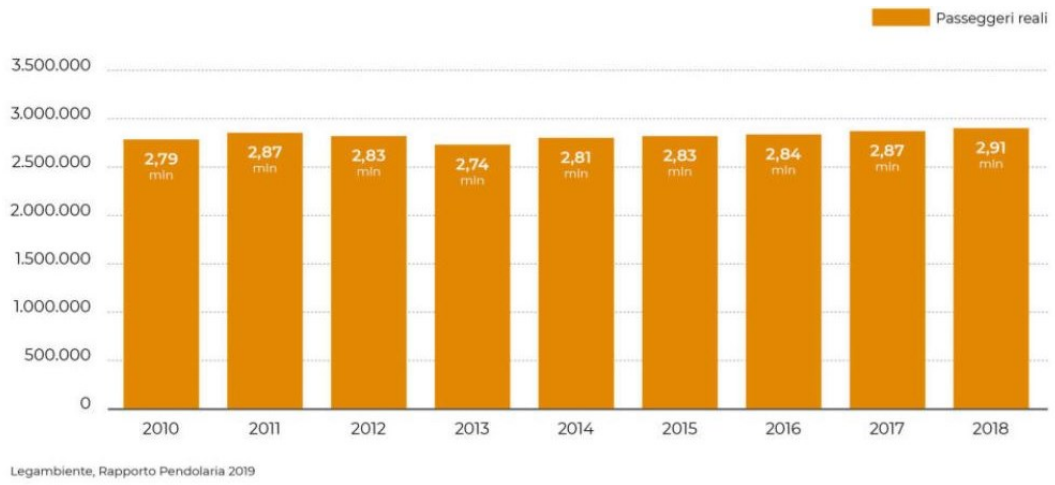

Figura 2-6: Passeggeri al giorno del trasporto regionale ferroviario

Tabella 2-1: Andamento del numero di passeggeri e dell'offerta di trasporto pubblico locale nel periodo 2014 - 2018

Comune	Andamento passeggeri	Andamento offerta TPL
Ancona	⇔	⇔
Bari	↑	↑
Bergamo	↑	⇔
Bologna	↑	↑
Brescia	↑	⇔
Cagliari	↑	↑
Catania	↓	↓
Firenze	↑	⇔
Genova	↑	↑
Livorno	↓	↓
Messina	↑	↑
Milano	↑	↑
Modena	↑	⇔
Napoli	↓	↓
Padova	⇔	⇔
Palermo	⇔	⇔
Perugia	↓	↑
Roma	↓	↓
Torino	↑	↓
Trieste	↑	⇔
Venezia	↑	↓
Verona	↑	⇔

A favore del trasporto ferroviario regionale si avvalgono: la concorrenzialità rispetto a quello privato; la diffusione e la capillarità sul territorio; la convenienza economica; la sicurezza e la velocità commerciale.

Di contro, in aree a domanda debole come quartieri periferici di città con difficile accesso o piccoli paesi di vallata, si può incorrere in una carenza o addirittura in una assenza del servizio e, da un punto di vista più individuale e

soggettivo, può venire a mancare una piena autonomia nella pianificazione dello spostamento (Bentivogli e Panicara 2012).

La soluzione è univoca: per qualsiasi tipo di offerta e in qualsiasi contesto, da nord a sud, laddove si investe in un miglioramento della mobilità su rotaia e la si rende comoda, puntuale e competitiva nei confronti dell'auto il successo è garantito.

2.3 L'infrastruttura

L'input principale del servizio di trasporto è l'infrastruttura: la strada ferrata è caratterizzata da un basso consumo del suolo e da un minore impatto ambientale rispetto ad altri impianti, ma gli alti costi che ne derivano fanno sì che la sostenibilità economica sia maggiore in aree densamente popolate o con ampia concentrazione dell'attività produttiva (Bentivogli e Panicara 2012).

I principali elementi che compongono la linea sono:

- il corpo stradale, l'insieme delle opere fuori terra che predispongono la rete e sono di supporto alla sovrastruttura ferroviaria (rilevati e trincee); delle opere minori (muri, tombini e canalette idrauliche); delle opere di protezione e di confine e delle opere accessorie (sentieri pedonali e canalette portacavi);
- la sovrastruttura ferroviaria, costituita da rotaie, traverse, attacchi, ballast e strati di supporto.

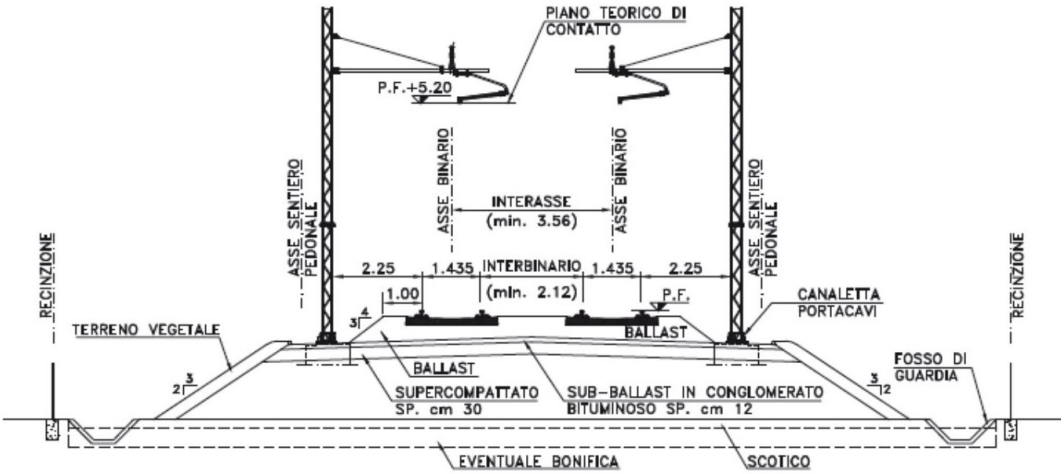

Figura 2-7: Rappresentazione schematica dell'infrastruttura ferroviaria (Giunta 2018)

Le linee ferroviarie, inoltre, possono essere a semplice, a doppio e a più binari e la capacità si misura in termini di numero di treni nell'unità di tempo.

Questo dato dipende da numerosi fattori:

- la distribuzione della velocità dei treni che la percorrono;

- l'impianto tecnologico;
- la lunghezza dei treni.

Le locomotive che la percorrono sono raggruppate e collegate tra loro a formare un convoglio, che può comprendere uno o più veicoli motori e uno o più veicoli rimorchiati.

La lunghezza totale del convoglio non è standard, ma è limitata dai vincoli imposti dalle caratteristiche della linea, ad esempio la lunghezza dei binari nelle stazioni, e dalla potenza del sistema.

Attualmente, il fulcro del trasporto regionale è il veicolo ferroviario appartenente al gruppo delle *E464*, una famiglia di locomotive elettriche e leggere, ideate per l'utilizzo su treni navetta a corto e medio raggio.

2.3.1 La locomotiva E464

Le più diffuse locomotive attualmente in uso dalle Ferrovie dello Stato italiane sono le *E464*, progettate tra il *1994* e il *1996* e in uso dal *2000* a oggi.

Furono concepite negli anni *Ottanta* come eredi delle *E646* e le *E424*, affidabili ma ormai obsolete quanto a tecnologia e prestazioni poiché risalenti agli anni *Cinquanta*.

Ne esistono *688* esemplari, costruiti dalla *Bombardier Transportation Italy* nello stabilimento di Vado Ligure, e costituiscono la forza motrice di gran parte dei treni navetta per brevi e medie distanze. Nel *2013* Trenitalia ha commissionato all'azienda la costruzione di altre macchine, per un totale di *300 milioni di euro*, da smistare tra Liguria, Lazio e Veneto.

Hanno una struttura modulare ideata per adattarsi ai diversi profili di missione: le E464 rappresentano i primi veicoli in Italia dotati di un sistema di accoppiamento automatico in grado di unire o separare rapidamente due diverse motrici per creare treni composti da due convogli distinti.

Grazie al sistema *Train Communication Network (TCN)* a 18 poli, possono essere unite fino a quattro locomotive per situazioni emergenziali o trasporti molto lunghi e pesanti: una volta combinate, le macchine possono generare in totale una potenza di *14.000 kW*, superiore a quella di un *TGV-Eurostar*.

La loro flessibilità risiede anche nella possibilità di poter essere collegate a locomotive più moderne, come le *E402A*, e a una grande varietà di vetture del tipo semipilota, una carrozza passeggeri dotata di una cabina per il comando della motrice.

Sono costituite da una cassa con struttura in lamiera d'acciaio e il tetto in alluminio, materiali che la rendono molto leggera, con una massa in servizio di *72 tonnellate*; la lunghezza del singolo vagone è di *15,75 metri* e la potenza oraria di *3.500 kW*, con una velocità massima omologata di *160 chilometri orari*.

Figura 2-8: Interno di una vettura in fase di allestimento

Il vagone di testa è dotato di una facciata posteriore piatta, in cui si trova una porta intercomunicante di servizio per l'accesso alle carrozze, e una anteriore aerodinamica con cabina di guida.

La visibilità sull'anteriore è assicurata da un finestrino fisso dotato di tergicristalli posizionato sulla testata e da uno laterale di tipo semiaperto a scorrimento. È inoltre presente la botola di accesso all'imperiale, la parte superiore della scocca del rotabile ferroviario.

Il retro, oltre all'intercomunicante, presenta un finestrino di fronte al quale c'è un ridotto posto di comando per poter eseguire movimenti in retromarcia con una velocità limitata al massimo a *30 chilometri orari*.

Un corridoio percorre tutto il veicolo al quale si accede per mezzo di porte scorrevoli lungo la fiancata e che permettono di accedere anche alla sala macchine. La caratteristica di avere una apertura intercomunicante con le carrozze trainate rende il prototipo omogeneo in tutta la sua lunghezza e dà la possibilità al capotreno di accedere anche alla cabina di guida, in modo da avere, oltre a lui, un conduttore unico in luogo di due macchinisti.

Figura 2-9: A sinistra vista anteriore e a destra vista posteriore della locomotiva E464

Figura 2-10: Viste laterali della locomotiva E464

La cassa poggia su due carrelli a due assi mediante sospensioni primarie e secondarie realizzate con molle a elica, invece del consueto perno.

Figura 2-11: Particolare del carrello

L'elettronica di trazione è stata semplificata, con l'adozione di un solo convertitore a due inverter per l'alimentazione di quattro motori asincroni

trifase. Questo sistema, detto a *schema incrociato*, è notevolmente più leggero della vecchia architettura a due circuiti indipendenti e altrettanto sicuro: ha la possibilità che, in caso di guasto, la motrice non perda la forza di trazione di due o più motori, come sarebbe avvenuto negli impianti tradizionali, ma sia in grado di continuare a muoversi e a raggiungere la prima stazione utile per recuperi di emergenza anche su linee in pendenza o con treni lunghi e pesanti.

Infatti, se normalmente l'alimentazione è data da una rete a *3.000 V* a corrente continua, il funzionamento è assicurato, a prestazioni e velocità ridotte, anche da una rete a *1.500 V* a corrente continua: anche se in caso di avaria la funzione dell'inverter è ridotta del *50%*, con tutti i motori attivi si raggiunge ugualmente il massimo sforzo di trazione.

Due convertitori ausiliari alimentano: il primo il sistema di raffreddamento dei motori, costituito da una miscela di acqua deionizzata e glicole etilenico con funzione antigelo; il secondo motocompressori, motoconvettore, carica delle batterie, climatizzazione della cabina e reostato di frenatura, che in caso di frenatura elettrica a recupero dissipa l'energia prodotta sotto forma di calore.

La batteria e i caricabatteria sono alloggiati in un vano posto inferiormente, tra i due carrelli sotto la cassa.

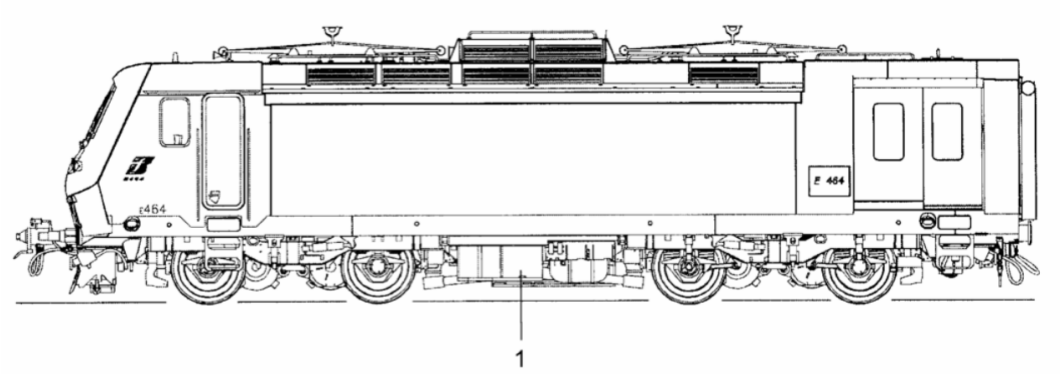

Figura 2-12: Vista laterale e posizione del vano batteria (punto 1)

Il controllo del veicolo presuppone la conoscenza delle varie componenti e delle rispettive funzioni:

- il *motore*, per generare un movimento rotatorio a partire dall'energia elettrica;
- l'*alternatore*, per generare energia elettrica usando un movimento rotatorio;
- *tristori* (*SCR*), *GTO* (*Gate Turn Off*) e *IGBT* (*Insulated Gate Bipolar Transistor*), dispositivi di controllo che fanno passare o meno una corrente;
- il *chopper*, convertitore statico che converte una tensione di ingresso continua in una tensione di uscita continua di diverso valore;
- il *trasformatore*, per cambiare la tensione di una corrente alternata;
- il *raddrizzatore*, per trasformare una corrente alternata in una corrente continua;
- l'*inverter*, per trasformare una corrente continua in una corrente alternata;
- il *controller*, per regolare la frequenza della corrente alternata generata.

Attraverso il computer di bordo è possibile monitorare lo stato della macchina e azionare i vari dispositivi, per mano di macchinisti con un altissimo grado di specializzazione.

Nella Tabella 2-2 sono riportate le principali caratteristiche della motrice E464.

Tabella 2-2: Sintesi dei parametri della locomotiva E464

Descrizione	Parametri
Tipo di alimentazione	Elettrica
Costruttore	Bombardier Transortation Italy
Periodo di costruzione	1999 – 2015
Profilo di impronta	UIC 500 - 1
Rodiggio	Bo' Bo'
Scartamento	1435 mm
Diametro delle ruote (nuove/consumate)	1100 mm/1010 mm
Passo (tra i carrelli)	7540 mm
Passo (tra gli assi di ogni carrello)	2650 mm
Lunghezza	15750 mm
Larghezza	3106 mm
Altezza	4282 mm
Massa totale	72 t
Massa dell'asse	18 t
Tensione della batteria	24 V DC (valore nominale)
Sistemi elettrici	3000 V DC
Motori di trazione	Trifase sincroni
Trasmissione	27/55 e 26/44 rapporto di trasmissione
Velocità massima	160 km/h
Potenza in uscita	3,5 MV
Sforzo di trazione	200 kN (avviamento) – 85 kN (potenza frenante)
Potenza massima di trazione alle ruote	3500 KW
Potenza di frenata reostatica	2350 KW
Rapporto di trasmissione	1/5
Tensione di linea	3 KV CC, 1,5 KV CC
Tensione di alimentazione ausiliaria	450 V a 60 HZ fissa e variabile
Operatori	FS Trenitalia, TiLo, Ferrovie Emilia-Romagna, Trenord
Numero in classe	728

3 Il sistema di Telediagnostica

Il sistema manutentivo dei trasporti, in particolare quello ferroviario, rappresenta un passaggio cruciale per potenziare il servizio e soddisfare le aspettative dei suoi fruitori, in termini di affidabilità e disponibilità dei veicoli: è infatti un processo volto al mantenimento e al miglioramento della qualità dei beni e al tempo stesso al contenimento dei costi attraverso l'ottimizzazione della gestione tecnico-economica.

L'approccio classico si basa su due principi basilari e affidabili: è *correttivo*, finalizzato alla riparazione di componenti guasti al termine del servizio, e *preventivo*, in quanto organizza operazioni su determinati elementi con scadenze fisse temporali o chilometriche.

Questa pratica ha portato, nel tempo, a evidenziare diversi inconvenienti: si subiscono gli effetti dei danni che solo successivamente saranno riparati; il legame tra usura dei componenti e calendario manutentivo non è efficace e le scadenze raggruppano attività non omogenee per problemi logistici (Masini 2015).

Storicamente i sintomi delle rotture venivano rilevati tramite annotazione sui libri di bordo e in seguito anche con segnalazioni in sala operativa, ma la ricerca degli errori era complicata e il materiale di ricambio e di riparazione richiesto non era compatibile con quello disponibile (Masini 2015).

La maggiore esigenza del buon funzionamento globale dei veicoli, in termini di meno guasti, più sicurezza e costi di gestione ridotti, ha dato un notevole impulso alla ricerca di sistemi più efficienti di rilevazione e trasmissione di dati: si passa da una *manutenzione preventiva*, programmata a priori, a quella *predittiva*, dinamica e basata sullo stato dei componenti.

La tecnologia è lo strumento che rende possibile questa rivoluzione e che è alla base del nuovo approccio della *telediagnostica*: a monte una mole di dati e informazioni da analizzare, a valle un intervento materiale e tempestivo.

A livello teorico, in uno scenario come quello attuale, l'impiego della telediagnostica ha un impatto pienamente risolutivo, in quanto è in grado di:

- limitare, con la prevenzione, i casi di guasto in esercizio poiché le informazioni fornite in tempo reale dal sistema di bordo, grazie all'elaborazione delle grandezze fisiche acquisite da sensori dedicati e/o grazie a logiche di correlazione tra i segnali diagnostici disponibili, consentono di desumere il potenziale stato di degrado nonché il residuo di vita utile degli elementi monitorati e di predire la possibile insorgenza di guasti prima del loro accadimento in opera. A livello economico, il beneficio che ne risulta è che la riparazione o sostituzione del componente può avvenire solo quando è realmente necessario, fino al limite di vita utile;
- ottimizzare i processi di manutenzione predittiva/correttiva, riducendo il tempo medio di ripristino (*Mean Time To Reapair – MTTR*), infatti, dal momento che i manutentori sono in grado di controllare con continuità e in tempo reale lo status della flotta, la pianificazione viene migliorata e l'approvvigionamento di parti di ricambio può essere effettuato in funzione della priorità e dei veicoli operativi. In questo modo, la disponibilità della flotta aumenta e i costi di gestione diminuiscono significativamente (Pucci , Mattera e Berlincioni 2016).

3.1 Lo stato dell'arte

La *manutenzione predittiva ferroviaria* può essere applicata utilizzando due approcci differenti:

1. *Knowledge-based*: si basa sia sulle conoscenze acquisite da progettisti e manutentori nell'esercizio delle loro rispettive funzioni, sia sull'utilizzo di specifiche analisi in termini di affidabilità, disponibilità, manutenibilità e sicurezza. Grazie a tali studi e alle esperienze pregresse acquisite, è possibile identificare a priori i comportamenti dei sottosistemi del treno a fronte di guasti incipienti. Effettuando i campionamenti dei valori con frequenze opportune, note a priori le soglie di malfunzionamento, vengono inviati degli allarmi quando i valori soglia sono superati.
2. *Data-driven*: grazie alla diffusione della digitalizzazione degli apparati, l'ingegneria di manutenzione dispone di una mole crescente di dati eterogenei e multisorgente, che però risiedono in archivi distinti, che li rendono poco fruibili ai fini delle analisi comparative. Per superare questa problematica, si ricorrere all'utilizzo di tecniche di *data mining*, che permettono di identificare delle relazioni tra dati apparentemente indipendenti e disgiunti, estraendo modelli causali e schemi ricorrenti precedentemente sconosciuti, e utilizzabili adesso per predire malfunzionamenti e cali di prestazioni.

La volontà di ottenere risultati a lungo termine, di ridurre i tempi connessi alla logistica, di ottimizzare le scorte e minimizzare gli sprechi e i guasti casuali, ha richiesto un ulteriore sviluppo nell'ambito della diagnostica che si è concretizzato attraverso la *Condition Based Maintenance* (*CBM*), un programma di manutenzione che si compone di tre fasi fondamentali:

I. acquisizione dei dati – raccolta di informazioni eseguita tramite sistemi di sensori di varia natura a seconda dei parametri interessati;

II. elaborazione dei dati – gestione e analisi dei dati raccolti con l'ausilio di tecniche di intelligenza artificiale;
III. fase decisionale – definizione delle politiche di manutenzione più efficaci.

La *CBM* è una strategia nella quale un componente o un equipaggiamento viene sottoposto a manutenzione solo quando c'è l'evidenza oggettiva del guasto incipiente, a seguito dell'analisi e della valutazione dei dati raccolti, in modo che le opportune attività di intervento avvengano in tempi consoni.

Alla luce della *CBM*, le nuove metodologie rielaborano radicalmente il ciclo di vita manutentivo del treno, riducendo la quota correttiva e investendo sul controllo continuo dei sottoinsiemi per studiare e seguire l'evoluzione delle loro performance e del loro degrado prima che intervenga una totale perdita di funzionalità (Castoldi e Perlini 2016).

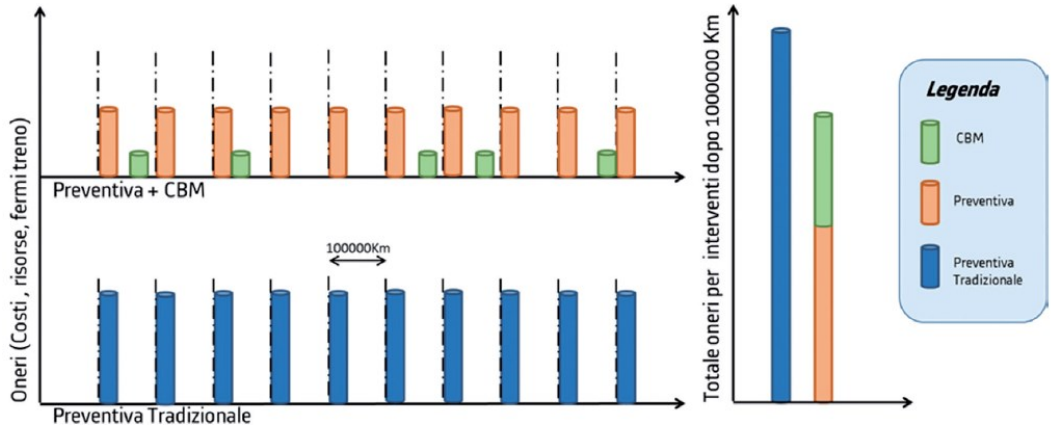

Figura 3-1: Confronto qualitativo tra manutenzione tradizione e manutenzione con CBM

Per implementare la funzione *CBM*, in aggiunta ai classici dati di avaria ed eventi anomali, a ogni sottosistema è richiesto di fornire le grandezze fisiche da esso misurate, tutti gli stati e i dati di contesto. A bordo del treno un dispositivo di comando dedicato (centralina) fungerà da collettore, eseguirà algoritmi diagnostici e invierà i dati a terra a una apposita *Centralina di Telediagnostica*.

La creazione e gestione degli algoritmi diagnostici dei sistemi a terra e dell'unità di bordo utilizzati per la generazione di avvisi manutentivi viene effettuata

attraverso l'utilizzo delle seguenti funzioni: *gestione AD*, *associazione AD/veicoli*, *gestione AD/veicoli* e *simulatore AD*.

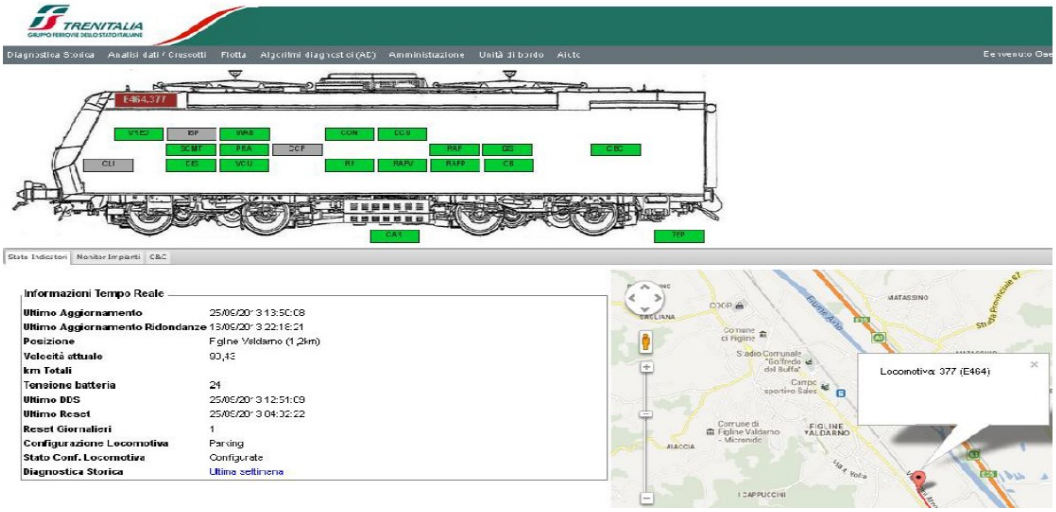

Figura 3-2: Visualizzazione dello stato dei sottoinsiemi di un veicolo in funzione dei suoi indicatori e della posizione corrente

In Figura 3-3 sono riportati alcuni avvisi manutentivi generati dagli algoritmi diagnostici.

Figura 3-3: Avvisi manutentivi generati dagli algoritmi diagnostici

La colorazione dell'avviso è funzione dello stato di avanzamento e assume il seguente significato:

- bianco, quando è semplicemente inviato al portale;
- arancione, quando è stato inoltrato al sistema informatico di Trenitalia RSMS (Rolling Stocks Management System), che supporta le attività di

pianificazione e gestione della manutenzione corrente, ciclica e revamping, ma non è ancora stato preso in carico dall'officina;

- verde, quando l'officina ha risolto l'anomalia.

Tali algoritmi sono monitorati costantemente e, sulla base dell'analisi continua del funzionamento dei componenti sotto osservazione, vengono affinate le soglie e gli algoritmi stessi, in un loop di miglioramento per iterazioni successive.

Le caratteristiche ambientali, di esercizio o di contesto possono richiedere un aggiornamento dei parametri della *CBM*, come soglie di intervento o logiche di diagnostica.

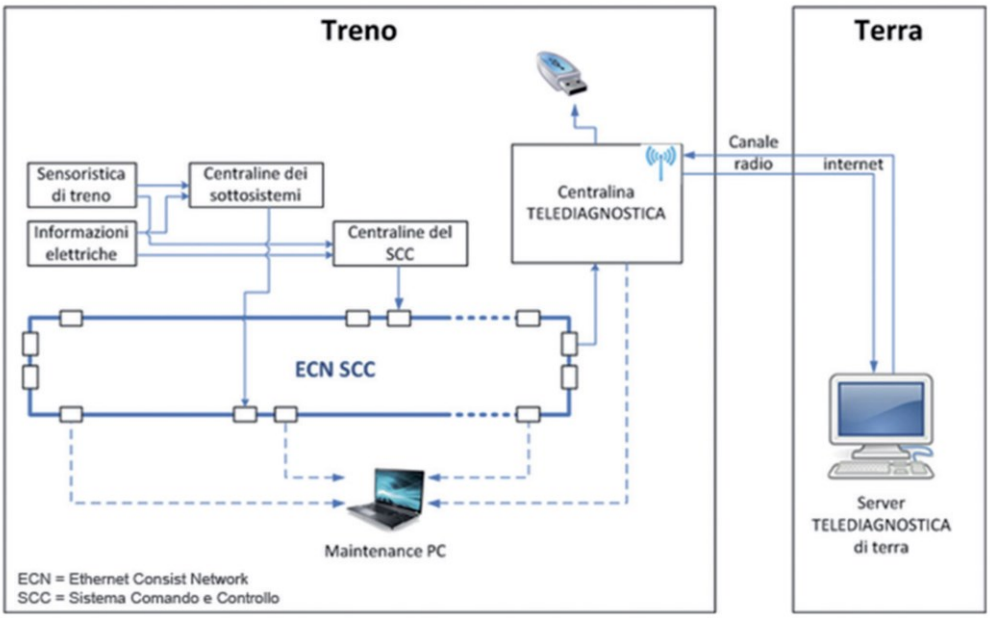

Figura 3-4: Confronto qualitativo tra manutenzione tradizionale e manutenzione con CBM

Questo approccio è caratterizzato da una crescita esponenziale dei dati trasmessi dai treni, ai quali si sommano le informazioni raccolte a terra e dei sistemi gestionali, rendendo di fatto inutilizzabili i vecchi database (E*xcel*, *DB SQL – Database Structured Query Language* e altri).

Gli operatori devono dotarsi di tecnologie e processi adeguati a recepire e trarre vantaggio dalla mole di dati che, in seconda battuta, dovrà restituire dei risultati che non rimangano fini a se stessi ma che si traducano in veri e propri interventi.

Ciò si esplica attraverso la *Control Room*: un luogo fisico ma anche un team di esperti ingegneri specializzati sul sistema treno e sull'analisi e interpretazione dei dati, i quali si focalizzano verso una prospettiva centralizzata ad alto livello piuttosto che su una prospettiva locale.

Infatti, la gestione centralizzata permette una visione d'insieme sui processi di manutenzione, sui dati diagnostici e sul comportamento dei rotabili, che è significativa per perseguire i seguenti obiettivi:

- industrializzare i processi definendo regole condivise e procedure ripetibili;
- anticipare il guasto, emettendo azioni correttive prima che si verifichi una perdita di funzionalità nel servizio, nel comfort o nelle prestazioni del treno;
- capitalizzare il ritorno di esperienza;
- fornire informazioni di supporto alle attività amministrative e di gestione;
- individuare delle correlazioni tra dati non omogenei.

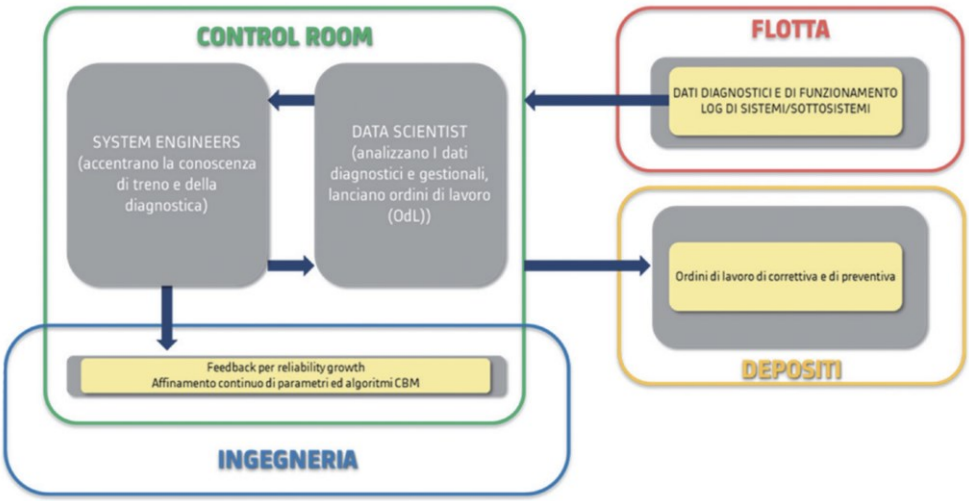

Figura 3-5: La Control Room: il fulcro della manutenzione

L'obiettivo della *Control Room* (*CR*) è quello di creare un legame tra l'analisi dei dati delle flotte dei treni e gli algoritmi e gli indicatori di *CBM*, già validati durante le prove di tipo e installati nella centralina di Telediagnostica di ogni rotabile (Castoldi e Perlini 2016).

3.2 Prevenzione guasti, telediagnostica e controllo continuo dello stato di un rotabile

La telediagnostica è un sistema di gestione delle informazioni rilevate da flotte ferroviarie progettato e realizzato per supportare le attività di manutenzione correttiva e predittiva.

È stato sviluppato dall'azienda *Bombardier Transportation Italy,* secondo specifiche di *Trenitalia,* al fine di aumentare la disponibilità di locomotive *E464* ed *E405*: è formato da un apparato di bordo, installato sule vetture, per la raccolta e l'invio a terra dei dati inerenti alle avarie e da un sistema di terra che storicizza ed elabora tali informazioni, presentandole agli utenti remoti del sistema.

La principale funzione del sistema di Telediagnostica è la generazione automatica di avvisi manutentivi attraverso l'elaborazione degli elementi forniti dai sistemi di controllo a bordo.

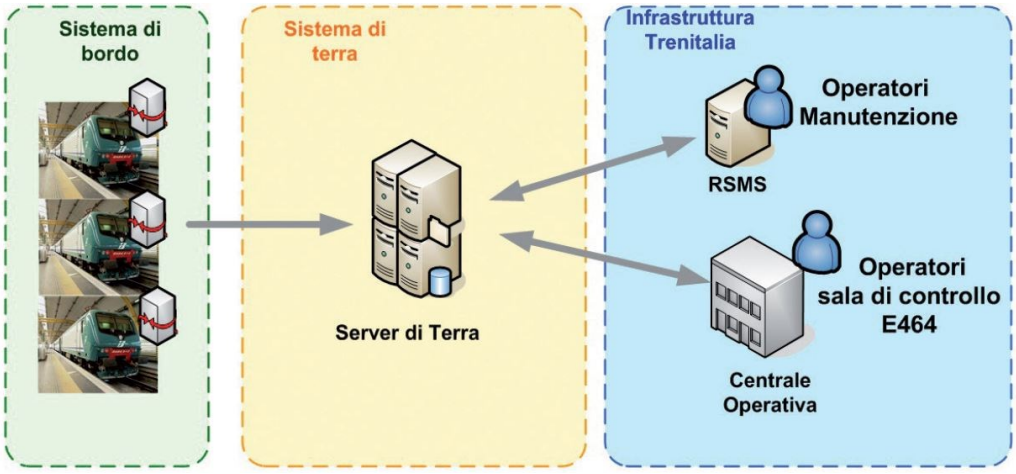

Figura 3-6: Rappresentazione schematica del sistema di diagnostica

L'elaborazione delle informazioni ricevute dai rotabili è affidata ad un sistema di intelligenza artificiale che a sua volta si basa su logiche complesse per la diagnosi automatica che sono definite, verificate e validate da un team di esperti manutentori e progettisti. Queste, dette *Algoritmi Diagnostici (AD)*, hanno permesso di replicare la conoscenza inerente alla diagnosi dei guasti in un sistema informatico centrale guasto (Agnoli e Del Gobbo 2016).

Le funzionalità del sistema di Telediagnostica sono associabili a tre principali processi operativi: esercizio, manutenzione e ingegneria di flotte di veicoli ferroviari. In particolare, per ognuno di questi ambiti, rende disponibile le seguenti funzioni principali:

- *Funzioni per l'esercizio:* comprendono il monitoraggio della flotta tramite opportuni cruscotti che forniscono i parametri di esercizio e lo stato dei sistemi del veicolo e anche la remotizzazione del banco di manovra a supporto del personale di macchina nelle situazioni critiche (depannage);
- *Funzioni per la manutenzione:* come la generazione automatica degli ordinativi di lavoro tramite Algoritmi Diagnostici che permettono di rilevare automaticamente i guasti (diagnosticabili) indicando cosa sostituire e/o l'attività manutentiva da realizzare. Tramite un ambiente grafico di configurazione è inoltre possibile eseguire indagini specifiche sulla ricerca di guasti di maggiore complessità senza necessità di presenziare il veicolo o svolgere corse prova dedicate.
- *Funzioni per l'ingegneria:* tramite un sistema di configurazione remota è possibile integrare le informazioni diagnostiche del veicolo aggiungendo nuovi eventi diagnostici e contatori di funzionamento senza la necessità di modificare la versione del software di controllo/diagnosi del rotabile. Tramite l'utilizzo dei contatori di funzionamento è inoltre possibile verificare l'effettivo fattore di utilizzo dei sottosistemi e dei componenti del rotabile. Ciò è funzionale all'introduzione di tecniche di manutenzione CBM (Agnoli e Del Gobbo 2016).

Un avviso manutentivo è caratterizzato dalle seguenti caratteristiche:

- Codifica *UIC* (*Union Internationale des chemins de fer - Unione Internazionale delle Ferrovie*) del rotabile che ne specifica il rodiggio, l'insieme delle parti comprese fra le rotaie e la sospensione elastica;
- codice algoritmo e descrizione sintetica;
- data e ora di rilevazione del malfunzionamento;

- descrizione dell'anomalia rilevata, in quanto un algoritmo può emettere più tipi di anomalia;
- ipotesi di uno o più componenti dell'albero di prodotto che possono aver prodotto l'anomalia;
- attività manutentiva consigliata per risolvere il guasto (Agnoli e Del Gobbo 2016).

Ogni avviso generato automaticamente dal sistema produce un beneficio in termini di ore risparmiate per effettuare una particolare attività manutentiva e si ha conseguentemente un abbattimento dei tempi di attraversamento.

Segue che, la sensibile diminuzione del tempo di ricerca dei guasti e la predisposizione del materiale necessario per la riparazione concorro alla riduzione dei guasti bloccanti in linea ottimizzando le attività di manutenzione, azzerando i tempi di attesa dovuti alle operazioni di carico/scarico dal magazzino e integrandosi con la pianificazione esistente in virtù del continuo monitoraggio degli indicatori di salute e di vita dei componenti.

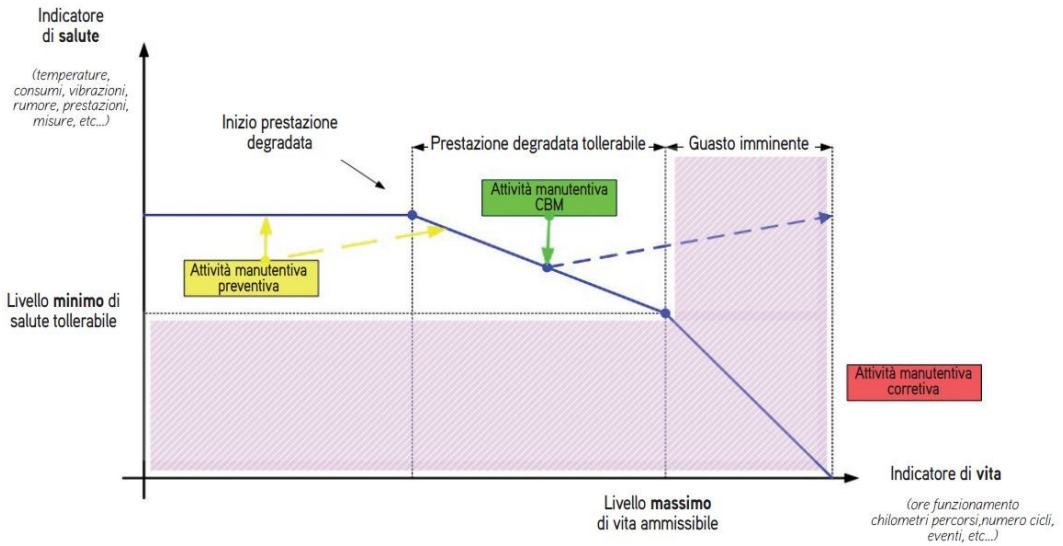

Figura 3-7: Rappresentazione grafica della logica della manutenzione alla luce della telediagnostica

4 Data Mining

Il costante aumento della quantità di dati prodotta quotidianamente e l'elevata crescita della capacità di elaborazione dei calcolatori fanno sì che, al giorno d'oggi, la vera difficoltà non sia reperire input quanto piuttosto estrarne informazioni utili, poiché la mole di partenza si presenta in forma eterogenea, ridondante e non strutturata.

Da questa forte esigenza, nasce e si sviluppa il *Data Mining*: una tecnica in grado di sopperire alle mancanze dei sistemi di analisi tradizionali, inapplicabili alle masse di dati grezzi, e che contribuisce a classificare e segmentare le informazioni oltre che a formulare ipotesi.

Il *Data Mining* non è un processo recente, ma affonda le sue radici nella statistica e nell'informatica applicata contestualmente alla consapevolezza che una così grande gestione avrebbe richiesto moltissima manodopera:

- *1960*: inizia la raccolta dei dati su dispositivi informatici e nascono i primi *database* con modelli gerarchici o relazionali.
- *1970*: si afferma il modello relazionale come paradigma di rappresentazione e strutturazione dei dati di un insieme di archivi e si sviluppano i primi *Data Base Management System (DBMS) relazionali*, un sistema di software progettato per consentire la creazione, la manipolazione e l'interrogazione delle banche dati e che snellisce la correlazione tra varie collezioni di informazioni.
- *1980*: con l'avvento dei *Data Warehouse*, una aggregazione di database indipendenti dai sistemi operativi di elaborazione dati e in cui vengono raccolti, compressi e archiviati dati storici provenienti da fonti diverse ed eterogenee, e del *Machine Learning*, un sottoinsieme dell'intelligenza artificiale che si occupa di creare sistemi che apprendono o migliorano le performance contestualmente ai dati che utilizzano, si parla di *Data Mining* in chiave moderna, inteso come quel processo che interseca le

competenze del *Machine Learning*, della statistica e della gestione dei *Data Warehouse*;

– *2011*: il *Data Mining* raggiunge lo stadio oggi conosciuto permettendo di costruire modelli predittivi e/o offrendo un affidabile supporto decisionale.

Affinché questo metodo risulti valido, deve include a monte diversi passaggi: inquadrare il problema, comprendere e preparare i dati, adottare le tecniche giuste per individuare dei pattern, interpretare i risultati e creare processi per distribuire i modelli (Vijay Kotu 2014).

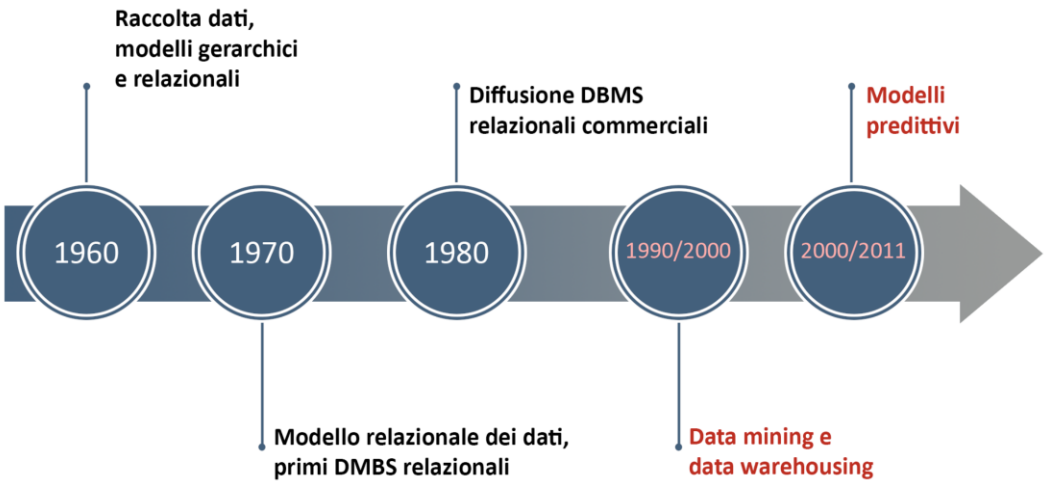

Figura 4-1: Linea del tempo: sviluppo del Data Mining

4.1 Il Data Mining e i suoi Obiettivi

Il *Data Mining* (*DM*) è l'insieme delle attività e degli strumenti necessari a ricavare informazioni da un sistema di dati non organizzato (es. banche dati, *Data Warehouse*, ecc.). Il suo scopo è quindi l'estrazione di informazione implicita, cioè utile e non precedentemente nota dagli elementi a disposizione, attraverso metodi automatici o semi-automatici: nello specifico, si procede analizzando una grande quantità di dati per individuarne le occorrenze (*pattern*) e cogliere le interrelazioni tra fenomeni, in modo da esprimerle come regole (De Agostini 2020).

Il *pattern* è una rappresentazione sintetica di un insieme: esprime in genere un modello ricorrente nei dati, ma può anche rilevare un modello eccezionale. In ogni caso deve essere:

- valido sui dati con un certo grado di confidenza;
- comprensibile, da un punto di vista semantico e sintattico, affinché l'utente possa interpretarlo;
- precedentemente sconosciuto e potenzialmente utile, affinché l'utente possa intraprendere determinate azioni di conseguenza.

L'individuazione di pattern è una fase cruciale del processo in quanto sta a significare che i dati sono correlati, che hanno una relazione e che sono prevedibili: la presenza di uno schema dà l'idea di quando o dove qualcosa accadrà prima che effettivamente accada.

Al contrario, la loro mancanza indica la vera casualità e di conseguenza l'impossibilità di utilizzare con criterio tutti gli input rilevati.

Un'abbondante letteratura è stata dedicata a questa ricerca e sono stati fatti enormi progressi, che vanno da algoritmi efficienti e scalabili per il *frequent itemset mining* nei *transaction database* a numerose frontiere di ricerca, come il *sequential pattern mining*, lo *structured pattern mining*, il *correlation mining*, la *classificazione associativa* e il *frequent pattern-based clustering*, così come le loro ampie applicazioni.

In sintesi, il DM non crea nuova informazione ma rende esplicita quella presente nel set iniziale attraverso un *data miner*, un programma che ne esplora il contenuto con l'obiettivo di estrarre regolarità o pattern.

Può però accadere che vengano identificate regole inutili o false o che si debba far fronte all'assenza o all'inesattezza di dati: le regole inutili sono quelle derivate da fenomeni che, pur presentendosi contemporaneamente, non sono legati da nessi causali; le regole false sono in generale la conseguenza della presenza di elementi spuri, probabilmente rilevati in condizioni eccezionali. È anche possibile che l'insieme delle informazioni disponibili sia insufficiente, o viziato da interferenze, per estrarre regole.

Un data miner gestisce un insieme di notizie che sono in genere organizzate in una *tabella T* con *m righe* (*record* o *example)* e *n colonne (attributi)*, dove gli elementi possono essere numerici oppure no. Il problema del data mining può essere allora formulato come segue: *data* T*, trovare una funzione classificatore* (f) *tale che "per molti esempi"* t*,* $T[t,n] = f(T[t,1],\ldots,T[t,n-1])$.

4.1.1 Metodi Data Mining

I dati di origine e i contesti applicativi possono essere estremamente diversi, ma le finalità principali per i quali si utilizza il Data Mining sono la descrizione e la previsione.

Il *metodo descrittivo* si focalizza sull'individuazione di correlazioni, tabulazioni incrociate o frequenze per trovare una regolarità negli elementi a disposizione e per rilevare dei modelli, da utilizzare nel reporting e nel monitoraggio. In questo modo, si approfondisce la conoscenza di ciò che avviene all'interno dei dati e quindi del mondo che rispecchiano, ma le azioni del processo non possono essere automatizzate.

Il *metodo predittivo* ha come obiettivo quello di prevedere i risultati futuri e non il comportamento attuale: si utilizzano alcune variabili dei dati analizzati per

predire valori sconosciuti e futuri di altre variabili di interesse. Si cerca così di realizzare l'automazione del processo decisionale, creando un modello in grado di dare una previsione o stimare un valore.

Sebbene i confini tra predizione e descrizione non siano netti (alcuni dei modelli predittivi possono essere anche descrittivi e viceversa), la distinzione è utile per comprendere la finalità generale da perseguire nella conoscenza.

Tali obiettivi possono essere perseguiti usando diverse attività, di seguito elencate (Daniel T. Larose 2015):

- Descrizione;
- Stima;
- Predizione;
- Classificazione;
- Raggruppamento (clustering);
- Associazioni.

4.1.1.1 Descrizione di concetti

In questa attività si sceglie un criterio per descrivere modelli e tendenze che si trovano all'interno di un set di informazioni. La descrizione può essere espressa in termini di caratterizzazione o di confronto:

- la *caratterizzazione di concetti* consiste nel riassumere le caratteristiche generali di un insieme di dati, ad esempio descrivere i clienti che hanno speso più di 1000€ presso il negozio X nell'ultimo anno. Un possibile risultato può essere un profilo di utente dai 40 ai 50 anni, occupato e non sposato;
- il *confronto tra concetti* fornisce una descrizione che confronta due o più insiemi di dati, ad esempio caratterizzare i clienti che comprano regolarmente al negozio X contrapposti a quelli che fanno acquisti di rado.

4.1.1.2 Stima

Nella stima, si approssima il valore di una *variabile target numerica* utilizzando un insieme di variabili predittive numeriche e/o categoriali.

Il software analizza un database iniziale e costruisce un modello utilizzando record "completi", cioè che forniscono il valore della variabile target e dei predittori. Sulla base di questa collezione di record, detta *training set*, il programma riesce a individuare un modello per cui, per nuove osservazioni, vengono effettuate stime del valore della variabile target, tenendo conto dei valori dei predittori.

Ad esempio, è possibile stimare la lettura della pressione arteriosa sistolica di un paziente, in base all'età, al sesso, all'indice di massa corporea e ai livelli di sodio nel sangue del paziente: la relazione tra la pressione arteriosa sistolica e le variabili predittive nel training set potrebbe fornire un modello di stima applicabile a nuovi casi.

4.1.1.3 Predizione

È possibile predire il valore di una *variabile a valori continui* sulla base di valori di altre variabili, assumendo un modello di dipendenza lineare/non lineare: l'obiettivo è quello di ottenere un dato continuo (numero reale).

La predizione è simile alla stima e alla classificazione, a eccezione del fatto che i risultati si trovano nel futuro. Esempi di attività di previsione includono:

- prevedere il prezzo di un'azione tra tre mesi;
- prevedere l'aumento percentuale dei decessi nel prossimo anno in caso di aumento del limite di velocità;
- pronosticare il vincitore della Serie A sulla base di un confronto delle statistiche della squadra;
- predire la velocità del vento in funzione della temperatura, umidità e pressione atmosferica.

Qualsiasi metodo e tecnica utilizzati per la classificazione e la stima può essere utilizzato anche, in circostanze appropriate, per la previsione (ad esempio regressione lineare, alberi decisionali, reti neurali, ecc...).

4.1.1.4 Classificazione

La classificazione è il processo che consiste, dato un training set in cui ogni riga è composta da un insieme di attributi di cui uno esprime la *classe di appartenenza* del singolo record, nel trovare un modello che sia in grado di predire il valore di una determinata classe, tra quelle già definite in partenza, su dati sconosciuti.

Tale processo si articola in tre passaggi:

- Fase I: si assume che ogni istanza in input faccia parte di una tra un numero predefinito di classi diverse e grazie a un attributo classificatore si può inquadrare a quale classe appartenga uno specifico record;
- Fase II: si stima l'accuratezza del modello con un data set suddiviso in training set, per costruire il modello, e in test set per validarlo;
- Fase III: di fronte a nuove istanze di cui sono noti tutti gli attributi eccetto quello classificatore, si usa il modello per inquadrare il record di classe ignota.

Questa attività è simile alla stima, tranne per il fatto che la variabile di destinazione è di categoria anziché numerica.

Ad esempio, una possibile variabile target è la fascia di reddito, che si suppone essere suddivisa in tre categorie: *reddito alto, reddito medio e reddito basso.* Il modello di data mining esamina un ampio set di record, ognuno contenente informazioni sulla variabile di destinazione, e un insieme di variabili di input o predittori, in modo che l'algoritmo si alleni a riconoscere quali combinazioni di variabili sono associate a una certa fascia di reddito.

Ne può risultare che le donne sopra ai 60 anni abbiano una fascia di reddito alta, segue che a una professoressa di 63 anni è associata una fascia alta.

4.1.1.5 Clustering

Un *cluster* è una raccolta di record simili tra loro e dissimili dai record di altri cluster: si tratta di un'analisi di raggruppamento che genera automaticamente classi interessanti senza consultare o tener conto di nessuna informazione nota. Questo metodo differisce dai precedenti in quanto l'obiettivo non è classificare, stimare o prevedere il valore di una variabile di destinazione, ma è quello di cercare di segmentare l'intero set di dati in sottogruppi o cluster relativamente omogenei, in cui la somiglianza dei record all'interno del cluster è massimizzata e la somiglianza con i record al di fuori di questo cluster è ridotta al minimo.

Esempi di attività di clustering possono essere:

- marketing mirato di un prodotto di nicchia per una piccola impresa che non dispone di un ampio budget per il marketing;
- come strumento di riduzione delle dimensioni quando il set di dati ha centinaia di attributi.

Il clustering viene spesso eseguito come fase preliminare in un processo di data mining: i raggruppamenti risultanti sono utilizzati come ulteriori input in una tecnica diversa a valle, come le reti neurali.

4.1.1.6 Associazioni

Le regole di associazione furono introdotte cercare regolarità tra i prodotti all'interno delle transazioni registrate nelle vendite dei supermercati (Agrawal, Imielinski e Swami 1993). Per esempio, la regola $\{\text{cipolle}, \text{patate}\} \Rightarrow \{\text{hamburger}\}$ individuata nell'analisi degli scontrini di un supermercato indica che il se il cliente compra insieme cipolle e patate è probabile che acquisti anche della carne per hamburger. Tale informazione può essere utilizzata come base per le decisioni riguardanti le attività di marketing, come ad esempio le offerte promozionali o il posizionamento dei prodotti negli scaffali. Le regole di associazione sono anche usate in molte altre aree, quali il Web mining, la scoperta di anomalie e la bioinformatica.

4.2 Le Regole di Associazione

Il problema della scoperta di regole di associazione non è di recente discussione, ma è stato identificato negli anni Novanta ed è rappresentato come segue. Consideriamo l'insieme di n attributi binari (oggetti o item) $I = \{i_1, i_2, \dots, i_n\}$ e l'insieme di transazioni (database) $D = \{t_1, t_2, \dots, t_m\}$. Ciascuna transazione appartenente a D possiede un codice identificativo (*ID*) e contiene un sottoinsieme degli oggetti contenuti in I. Una regola è definita come un'implicazione nella forma $X \Rightarrow Y$ dove $X, Y \subseteq I$ e $X \cap Y = \emptyset$. L'insieme di oggetti, noti anche come *itemsets,* X e Y vengono chiamati rispettivamente *antecedente* e *conseguente* della regola (Agrawal, Imielinski e Swami 1993).

Passando da livello teorico a quello pratico, si può fare riferimento a un esempio riguardante i prodotti venduti in un supermercato.

L'insieme è $I = \{pomodori, pane, olio, vino, miele\}$ e il database contenente gli oggetti è rappresentato nella Tabella 4-1, dove 1 indica la presenza di un oggetto in una transazione e 0 l'assenza. Una possibile regola di associazione potrebbe essere: $\{olio, pane\} \Rightarrow \{vino\}$, che indica che se il cliente acquista pane e olio, comprerà anche il vino.

Tabella 4-1: Esempio di base di dati con 5 oggetti e 5 transazioni

ID	Pomodori	Pane	Olio	Vino	Miele
1	1	1	0	0	0
2	0	0	1	0	0
3	0	0	0	1	1
4	1	1	1	0	0
5	0	1	0	0	0

Si osserva che l'esempio sopradescritto utilizza un set di dati estremamente limitato: di norma, in un'applicazione reale una regola necessita di un supporto di diverse centinaia di transazioni perché sia considerata statisticamente significativa e il database deve contenere migliaia/milioni di transazioni, a seconda del tipo di applicazione.

4.2.1 Proprietà delle associazioni

Per selezionare regole interessanti dall'insieme di tutte quelle possibili, vengono utilizzati dei vincoli su varie misure, cioè soglie minime da raggiungere affinché l'associazione sia potenzialmente significativa. I più noti sono *supporto* e *confidenza*.

Siano X, Y insiemi di elementi (itemsets), $X \Rightarrow Y$ una regola di associazione e T un insieme di transazioni di un assegnato database.

4.2.1.1 Supporto

Il *supporto* indica la frequenza con cui l'insieme di elementi appare nel set di dati.

Il supporto di X rispetto a T è definito come la proporzione di transazioni nel data set iniziale che contiene l'insieme di elementi X. Denotando una transazione con (i, t), dove i è l'identificatore univoco della transazione e t è il suo insieme di elementi, il supporto può essere definito come:

$$\text{supp}(X) = \frac{|(i,t) \in T : X \subseteq t|}{|T|}$$

Nell'esempio precedente, il set di articoli $X = \{olio, miele\}$ ha un supporto di $\frac{1}{5} = 0.2$ poiché si verifica nel 20% di tutte le transazioni (1 su 5).

L'argomento di supp() è un insieme di precondizioni, che diventa più restrittivo man mano che cresce (anziché più inclusivo).

Inoltre, anche il set di elementi $Y = \{pomodori, pane, olio\}$ ha un supporto di $\frac{1}{5} = 0.2$ in quanto appare anch'esso nel 20% di tutte le transazioni.

4.2.1.2 Confidenza

La *confidenza* è un'indicazione della frequenza con cui la regola è stata trovata vera, cioè misura la frequenza con cui gli elementi in Y vengono visualizzati nelle transazioni che contengono X.

Il valore di confidenza di una regola $X \Rightarrow \mathrm{Y}$ rispetto a un insieme T, è la proporzione delle transazioni contenenti X che contiene anche Y e viene calcolato come il numero di scontrini contenenti X e Y diviso per il numero di quelli contenenti X:

$$\mathrm{conf}(X) = \frac{supp(X \cup Y)}{supp(X)}$$

Ad esempio, la regola $\{olio, pane\} \Rightarrow \{pomodori\}$ ha una confidenza di $\frac{0.2}{0.2} = 1.0$ nel database, il che significa che per il 100% delle transazioni contenenti olio e pane la regola è corretta (il 100% delle volte che un cliente compra olio e pane, si acquistano anche i pomodori).

4.2.1.3 Lift

Il *lift* di una regola è espresso nella forma:

$$\mathrm{lift}(X \Rightarrow \mathrm{Y}) = \frac{\mathrm{supp}(X \cup Y)}{\mathrm{supp}(Y) \times \mathrm{supp}(Y)}$$

o il rapporto tra il supporto osservato e quello atteso se X e Y fossero indipendenti.

Ad esempio, la regola $\{pomodori, pane\} \Rightarrow \{olio\}$ ha un lift di $\frac{0.2}{0.4 \times 0.4} = 1.25$.

Affinché una regola sia effettivamente forte, cioè per la quale supporto e confidenza superino i valori di soglia, il lift deve essere maggiore di 1.

Se il lift è uguale a 1, la probabilità di occorrenza dell'antecedente e quella del conseguente sarebbero indipendenti l'una dall'altra: quando due eventi sono indipendenti l'uno dall'altro, non è possibile scrivere una regola che li coinvolga.

Se il lift è > 1, le due occorrenze dipendono l'una dall'altra e le regole sono potenzialmente utili per prevedere il conseguente in set di dati futuri.

Se il lift è < 1, gli elementi si sostituiscono l'uno con l'altro: ciò significa che la presenza di un elemento ha un effetto negativo sulla presenza di un altro e viceversa.

4.2.1.4 Conviction

La *conviction* di una regola è definita come:

$$\mathrm{conv}(X \Rightarrow \mathrm{Y}) = \frac{1 - \mathrm{supp}(Y)}{1 - \mathrm{conf}(X \Rightarrow \mathrm{Y})}$$

Questo parametro è sensibile alla direzione della regola, vale a dire che $\mathrm{conv}(X \Rightarrow \mathrm{Y})$ non è uguale a $\mathrm{conv}(Y \Rightarrow \mathrm{X})$. La convinzione si ispira modo alla definizione logica di implicazione e cerca di misurare il grado di implicazione di una regola.

Ad esempio, la regola $\{pomodori, pane\} \Rightarrow \{olio\}$ ha una conviction di $\frac{1-0.4}{1-0.5} = 1.2$ e può essere interpretata come il rapporto tra la frequenza attesa che X si verifichi senza Y (cioè la frequenza che la regola esprima una previsione errata, se X e Y fossero indipendenti), diviso la frequenza osservata che X si verifichi senza Y. In questo esempio, il valore di 1.2 mostra che la regola $\{pomodori, pane\} \Rightarrow \{olio\}$ sarebbe errata il 20% più spesso (1,2 volte più spesso) se l'associazione tra X e Y fosse puramente casuale.

Più la conviction è alta, meno le regole sono casuali e variegate, infatti, come il lift, i valori lontani da 1 indicano risultati interessanti mentre se è pari a 1 indica che X e Y sono indipendenti.

4.2.2 Applicazione del modello

Il processo che precede l'estrazione vera e propria delle regole è solitamente suddiviso in due fasi distinte e i valori di soglia per il supporto e la confidenza sono definiti caso per caso:

1. si individuano nel database tutti gli itemset che hanno una certa frequenza e che soddisfano il valore minimo di supporto;
2. agli itemset frequenti viene applicato un vincolo di confidenza minimo per formare regole.

Mentre il secondo passaggio è relativamente più semplice, il primo richiede maggiore accortezza e attenzione.

Trovare tutti gli oggetti frequenti implica la ricerca dell'insieme dei possibili gruppi di elementi, che è l'*insieme delle parti* di I e ha dimensione $2^n - 1$, escluso l'insieme vuoto che non è un insieme di elementi valido. Sebbene la dimensione dell'insieme delle parti cresca esponenzialmente al crescere del numero n degli elementi in I, è possibile una ricerca efficiente utilizzando la chiusura verso il basso del supporto. Tale proprietà, chiamata anche anti-monotonicità, garantisce che, per un dato itemset frequente, anche tutti i suoi sottoinsiemi siano frequenti e che quindi nessun itemset non frequente possa essere un sottoinsieme di itemset frequenti.

4.2.3 Algoritmi di Ricerca

In letteratura si annoverano diversi algoritmi di ricerca: tra i più utilizzati vi sono *Apriori* e *FP-Growth*, che utilizzano la proprietà di chiusura del dataset iniziale.

4.2.3.1 Algoritmo Apriori

L'algoritmo Apriori è costituito da una struttura a livelli e si basa sul presupposto teorico che, se un itemset è frequente, allora anche tutti i suoi sottoinsiemi subset sono frequenti. Si considerano, a ogni livello, un numero sempre crescente di oggetti, calcolandone il supporto e cancellando, dai possibili itemset frequenti di quel livello, quelli i cui subset non sono contenuti negli itemset frequenti del livello precedente.

Tabella 4-2: Esempio di dataset utilizzato per l'algoritmo Apriori

Transazioni	Item
1	{pane, olio}
2	{olio, sale, pepe}
3	{pane, sale, pepe, origano}
4	{pane, pepe, origano}
5	{pane, olio, sale}
6	{pane, olio sale, pepe}

Al primo livello, vi sono tutti gli itemset possibili, contenenti al loro interno un oggetto, e se ne calcola il supporto:

Tabella 4-3: Itemset candidati per il livello 1

Itemset	Supporto
{pane}	5
{olio}	4
{sale}	4
{pepe}	4
{origano}	2

Il supporto di ogni itemset candidato viene confrontato con la soglia limite scelta specificatamente per il caso in esame e, se maggiore, questo viene inserito nell'elenco degli itemset frequenti del livello considerato.

Supponendo che la soglia sia 1, tutti gli item selezionati hanno supporto maggiore e tutti vengono considerati come itemset frequenti del livello 1.

Nel passo successivo, si considerano tutti i possibili itemset composti da due item a partire da quelli di livello 1: una volta generati i candidati, se ne calcola il supporto e si confronta con la soglia minima.

Tabella 4-4: Itemset candidati per il livello 2

Itemset	Supporto
{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

I due itemset che non hanno un supporto tale da essere considerati frequenti vengono quindi scartati. L'elenco degli itemset frequenti di secondo livello è:

Tabella 4-5: Itemset frequenti di livello 2

Itemset
{pane, olio}
{pane, sale}
{pane, pepe}
{pane, origano}
{olio, sale}
{olio, pepe}
{sale, pepe}
{pepe, origano}

Si procede con il livello 3, costituito da tutte le combinazioni possibili formate tra 3 elementi, a partire dagli itemset frequenti del livello 2.

Tabella 4-6: Itemset candidati per il livello 3

Itemset	Supporto
{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

In questo step, oltre a non avere un supporto sufficiente in quanto minore di 1, gli itemset $\{pane, olio, origano\}$,$\{pane, sale, origano\}$ e $\{sale, pepe, origano\}$ vengono eliminati perché i loro subset $\{olio, origano\}$ e $\{sale, origano\}$ non sono presenti tra gli itemset frequenti di livello 2.

Anche se compaiono nella Tabella 4-6, l'algoritmo li scarta ancora prima di calcolarne il supporto, applicando il principio Apriori.

L'elenco dei sottoinsiemi frequenti del terzo livello è:

Tabella 4-7: Itemset frequenti del livello 3

Itemset
{pane, olio, sale}
{pane, sale, pepe}
{pane, pepe, origano}
{olio, sale, pepe}

Nell'ultimo step si procede analogamente ai precedenti, ottenendo:

Tabella 4-8: Itemset candidati di livello 4

Itemset	Supporto
{pane, olio, sale, pepe}	1

Dal momento che il subset $\{pane, olio, pepe\}$ non è presente nei gruppi frequenti del livello 3, l'itemset riportato nella Tabella 4-8 non è considerato come frequente e questo rende l'insieme di livello 4 vuoto.

L'algoritmo ha funzionato, ma la sua efficienza dipende dal numero di item presente nel dataset iniziale, che deve essere scannerizzato di volta in volta, richiedendo un tempo per le varie sottofasi non predicibile (Vijay Kotu 2014).

4.2.3.2 Algoritmo FP-Growth

L'algoritmo *Frequent Pattern (FP)-Growth* fornisce un modo alternativo di calcolare un itemset frequente comprimendo i record delle transazioni e usando una speciale struttura di dati a grafo chiamata *FP-Tree*.

Un FP-Tree può essere pensato come una trasformazione dell'insieme di dati in formato grafico: a differenza dell'algoritmo Apriori, prima si individuano e si schematizzano i dati in un albero compresso e poi lo si utilizza per generare insiemi di elementi frequenti. L'efficienza dell'algoritmo FP-Growth dipende quindi da quanta compressione può essere raggiunta nel generare l'albero FP (Vijay Kotu 2014).

4.2.3.2.1 Generazione dell'albero FP

Si consideri l'insieme di dati mostrato nella Tabella 4-9, che contiene sei transazioni di quattro item, quali $News, Finance, Sports, Entertaiment$.

Tabella 4-9: Dataset di esempio per algoritmo FP-Growth

Transazioni	Items
1	{News, Finance}
2	{News, Finance}
3	{News, Finance, Sports}
4	{Sports}

Transazioni	Items
5	{News, Finance, Sports}
6	{News, Entertaintment}

Preliminarmente si deve stabilire una soglia minima di supporto che l'algoritmo userà come discriminante: se il supporto dell'item è maggiore di quello della soglia prefissata allora verrà considerato, in caso contrario verrà scartato e non contribuirà alla costruzione dell'FP-tree.

L'obiettivo è quello di trasformare la lista di transazioni in una mappa ad albero conservando tutte le informazioni e rappresentando i percorsi frequenti, per questo l'algoritmo esegue una dopo l'altra le seguenti operazioni:

1. il primo passo è quello di ordinare tutte le voci di ogni transazione in ordine decrescente di frequenza (o di numero di supporti). Ad esempio, $News$ è l'elemento più frequente e $Sports$ è l'elemento meno frequente nella transazione, in base ai dati della Tabella 4-9. In particolare, la terza transazione $\{Sports, News, Finance\}$ deve essere riorganizzata in $\{News, Finance, Sports\}$, questo aiuterà a semplificare la mappatura dei percorsi frequenti nei passi successivi;
2. una volta che gli elementi all'interno di una singola transazione sono stati riorganizzati, è possibile ora mappare la transazione nell'albero FP: a partire da un nodo nullo, la prima transazione $\{News, Finance\}$, può essere rappresentata dalla Figura 4-2, dove il numero tra parentesi accanto al nome dell'elemento è il numero di transazioni che seguono il percorso e quindi quante volte si passa per quel nodo;

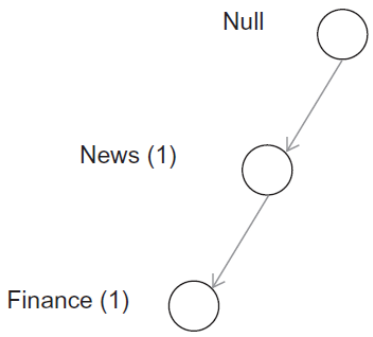

Figura 4-2: FP-tree: transazione 1

3. dal momento che la seconda transazione $\{News, Finance\}$ è uguale alla prima, il percorso sarà lo stesso e si incrementano semplicemente i numeri;
4. la terza transazione contiene $\{News, Finance, Sports\}$: l'albero è ora esteso a $Sports$ e il conteggio del percorso dell'elemento è incrementato, vedi Figura 4-3;

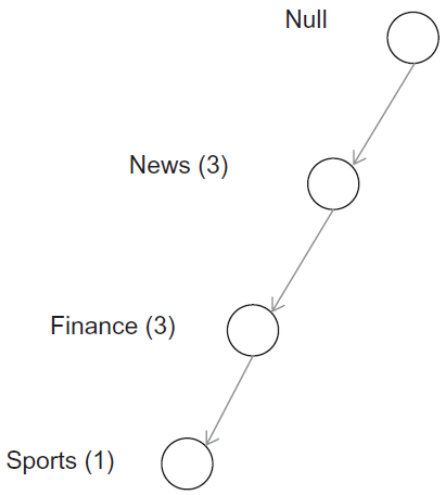

Figura 4-3: FP-tree: transazione 1,2 e 3

5. la quarta transazione contiene solo l'elemento $Sports$: poiché non è preceduto da $News$ e $Finance$ è necessario creare un nuovo percorso dall'elemento nullo e annotare il conteggio degli elementi. Tuttavia, poiché entrambi i nodi indicano lo stesso elemento, dovrebbero essere collegati da una linea tratteggiata;
6. questo processo continua fino alla scansione di tutte le transazioni: in questo modo tutti i record delle transazioni possono ora essere rappresentati da un compatto FP-Tree, come mostra la Figura 4-4.

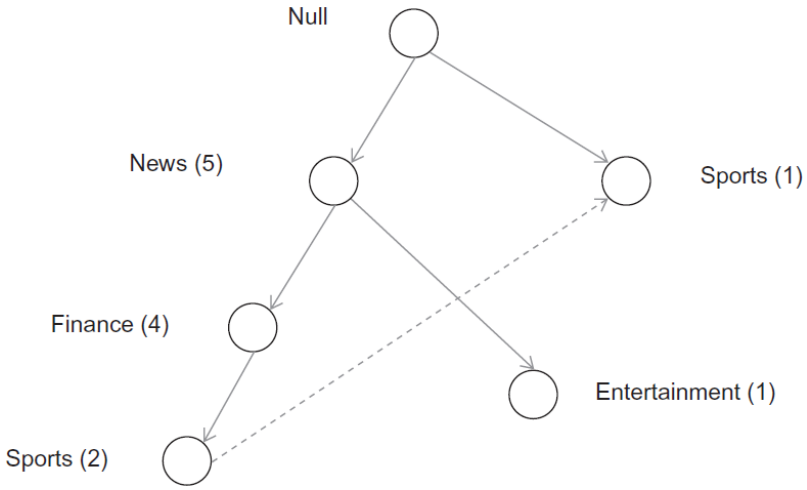

Figura 4-4: FP-tree: transazioni 1 - 6

La compattezza dell'FP-Tree dipende dalla frequenza con cui un percorso si ripropone all'interno di un dato insieme di transazioni: essendo l'obiettivo chiave quello di identificare i percorsi comuni, gli insiemi di dati utilizzati in questo tipo di analisi contengono molti percorsi frequenti.

Nel caso peggiore, infatti, si avrebbero tutte transazioni da cui risulterebbero percorsi unici, con la conseguenza che la stessa generazione di regole sarebbe meno significativa per l'analisi di associazione.

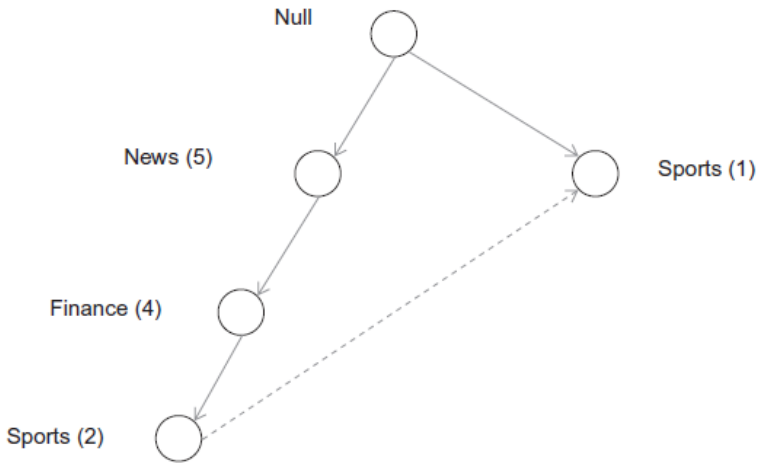

Figura 4-5: FP-tree compatto

Per raggruppare in un diagramma l'insieme degli elementi frequenti, l'algoritmo FP-Growth adotta un approccio dal basso verso l'alto, partendo dagli elementi che ricorrono meno volte: poiché la struttura dell'albero è ordinata in base al

numero di supporti, gli elementi meno frequenti possono essere individuati nelle foglie dell'albero.

Nella Figura 4-4 si evince che gli elementi che compaiono meno sono $Entertaiment$ e $Sports$ perché il numero di supporto è una transazione: se $Entertaiment$ fosse un elemento frequente, allora l'algoritmo individuerebbe tutti gli insiemi di elementi che terminano con $Entertaiment$, quali $\{Entertaiment\}$ e $\{News, Entertaiment\}$, seguendo il percorso dal basso verso l'alto.

Dal momento che $\{Entertaiment\}$ non è frequente, l'algoritmo salta l'elemento e passa a quello successivo, $\{Sports\}$, e individua tutti i possibili insiemi di oggetti che terminano con $Sports$, quali $\{Sports\}$, $\{Finance, Sports\}, \{News, Sports\}, \{News, Finance, Sports\}$.

Ciò è possibile generando un percorso con prefisso e un FP-Tree condizionato per un elemento, come mostrano nella Figura 4-6.

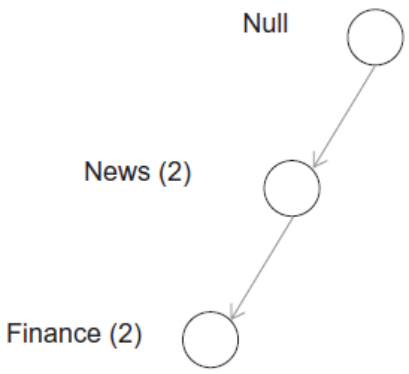

Figura 4-6: FP-Tree condizionato

Il percorso con prefisso di un elemento è un sottoalbero con solo percorsi che contengono l'elemento di interesse. Un FP-Tree condizionale per un elemento, ad esempio $\{Sports\}$, è simile all'FP-Tree, ma con l'elemento $\{Sports\}$, rimosso. Sulla base dell'FP-Tree condizionale, l'algoritmo ripete il processo di ricerca dei nodi foglia. Poiché i nodi foglia dell'albero condizionale di $Sports$ coesistono con $\{Sports\}$, l'algoritmo trova l'associazione con la $Finance$ e genera $\{Finance, Sports\}$.

La generazione di regole in FP-Growth è molto simile all'algoritmo Apriori: poiché l'obiettivo è quello di trovare elementi che si ripetono frequentemente, molte delle transazioni dovrebbero avere essenzialmente lo stesso percorso. Segue che in molte applicazioni pratiche il rapporto di compattazione è molto alto e, grazie all'utilizzo di grafi per mappare le relazioni tra oggetto frequenti, porta risultati efficienti e trova applicazione anche oltre l'analisi delle associazioni (Vijay Kotu 2014).

4.3 Data Mining applicato alla Manutenzione

L'enorme mole di dati a cui attinge il sistema di trasporto ferroviario costituisce un dataset potenzialmente interessante se combinato e analizzato: non è il trattamento dei dati che di per sé costituisce una innovazione, ma la consapevolezza e la capacità, grazie alla potenza di calcolo disponibile e all'aumento di informazioni utilizzabili, di metterla in atto in contesti applicativi per risolvere problemi concreti.

Infatti, all'aumentare della numerosità e disomogeneità del parco rotabili, le aziende di trasporto ferroviario hanno dovuto fronteggiare, nel corso degli anni, le disfunzioni dovute ai limiti degli approcci tradizionali alla manutenzione.

Al giorno d'oggi e in alcuni ambiti, è una consuetudine molto frequente posizionare processori e sensori sulle macchine per ricevere in tempo reale i dati sull'andamento del sistema, generando così un flusso che non può essere gestito da database tradizionali e che risulta difficile da manipolare in quanto vasto ed eterogeneo.

Allo stesso tempo, le nuove metodologie di analisi riescono, in tempi ragionevoli, a estrarre informazioni utili al management e ad agire, in conseguenza degli input raccolti, per effettuare aggiustamenti o riparazioni prima che accada il guasto vero e proprio.

La commistione di questi due elementi, monitoraggio della condizione dell'apparecchiatura da un lato e nuovi strumenti per l'estrazione ed elaborazione dei dati attraverso le tecniche di data mining dall'altro, porta alla realizzazione della manutenzione predittiva, che cerca di prevedere guasti futuri e di facilitare il processo decisionale che ne consegue.

Nel presente lavoro il metodo utilizzato è quello delle *Assotiation Rules*, che viene applicato ai dati di guasto per scoprire l'impatto di un determinato errore sugli altri e per individuare potenziali elementi che, se analizzati da tecnici specializzati, possano tradurre i guasti di oggi nell'efficienza di domani.

5 Elaborazione e Analisi dei Dati: il caso studio

Il lavoro di tesi si è sviluppato in tre blocchi principali: il primo relativo al preprocessing dei dati, il secondo volto all'estrazione delle regole e l'ultimo, invece, per l'analisi e visualizzazione dei risultati.

Figura 5-1: fasi dell'analisi dei dati

La fase di *preprocessing* ha molta importanza in quanto i dati grezzi, se utilizzati direttamente nel mining, possono causare confusione e produrre risultati imprecisi. Per questo i dati ricevuti in input vengono puliti e armonizzati prima della fase di mining vera e propria.

L'*estrazione delle regole* rappresenta il punto centrale del lavoro di tesi. Verrà eseguita in RapidMiner che, attraverso la definizione della corretta sequenza di operatori, permetterà la trasformazione e l'estrazione delle associazioni. Si tratta di un passaggio fondamentale in quanto la bontà delle regole estratte sarà dipendente dai parametri impostati in questa fase.

L'*analisi e visualizzazione* delle regole rappresenta la fase finale del lavoro: si analizzeranno in maniera critica le regole estratte, con l'intento di individuare quelle associazioni potenzialmente utili alla diagnostica.

5.1 Data-set iniziale

Prima di procedere con la descrizione della fase di pulizia e per poi arrivare allo sviluppo dell'algoritmo di estrazione dati, è conveniente specificare il contenuto e il formato dei dati di partenza, per ripercorre al meglio le criticità e i punti di forza del percorso sperimentale.

Il data-set in esame si riferisce allo storico di eventi diagnostici (*Diagnostic Data Record - DDS*), generati da un piccolo gruppo di treni E464. La loro estrazione è stata effettuata dagli operatori della sala di controllo attraverso un tool interno che raccoglie le informazioni diagnostiche inviate dalle motrici al sistema di terra.

I dati delle cinque locomotive a disposizione fanno riferimento a due intervalli di tempo distinti:

- per le locomotive 152, 292, 481 e 514 i DDS sono relativi ad uno storico che va da Maggio 2015 a Marzo 2016;
- per la locomotiva 519 i DDS sono più recenti e relativi ad un periodo che va da Gennaio 2019 a Maggio 2019.

Inoltre, per i due gruppi sopradescritti, il numero di processori considerato in fase di estrazione è diverso. Il dettaglio è riportato in Tabella 5-1.

Tabella 5-1: classificazione set di dati

MOD. LOCOMOTIVA	NUM. LOCOMOTIVA	PERIODO	PROCESSORI
E464	152, 292, 481, 514	Mag. 2015 - Mar. 2016	DCU1, FLG2
	519	Gen. 2019 - Mag. 2019	CSA1, CSA2, DCU1, FLG2, MCG, VCUMON

Il risultato dell'estrazione è rappresentato da un insieme di circa *45 file CSV* (*Comma-Separated Values*), contenenti i record degli eventi diagnostici e le relative variabili ambientali dei rotabili in esame.

E464_-152-292-481-514_20160401_010040_From_0_To_50000_DA_0	01/04/2016 02:49	File con valori separati da virgola (CSV) ...	10.007 KB
E464_-152-292-481-514_20160401_052250_From_0_To_50000_DA_0	01/04/2016 07:12	File con valori separati da virgola (CSV) ...	9.651 KB
E464_-152-292-481-514_20160401_052306_From_0_To_50000_DA_0	01/04/2016 07:56	File con valori separati da virgola (CSV) ...	7.976 KB
E464_-152-292-481-514_20160401_052436_From_0_To_50000_DA_0	01/04/2016 08:22	File con valori separati da virgola (CSV) ...	5.437 KB
E464_-152-292-481-514_20160401_082842_From_0_To_50000_DA_0	01/04/2016 08:44	File con valori separati da virgola (CSV) ...	7.664 KB
E464_-519_20190330_133147_From_0_To_50000_DA_Samples_0	30/03/2019 13:45	File con valori separati da virgola (CSV) ...	8.102 KB
E464_-519_20190330_133917_From_0_To_50000_DA_Samples_0	30/03/2019 14:06	File con valori separati da virgola (CSV) ...	10.117 KB
E464_-519_20190330_134108_From_0_To_50000_DA_Samples_0	30/03/2019 14:20	File con valori separati da virgola (CSV) ...	7.949 KB
E464_-519_20190330_134351_From_0_To_50000_DA_Samples_0	30/03/2019 14:32	File con valori separati da virgola (CSV) ...	6.953 KB
E464_-519_20190330_134601_From_0_To_50000_DA_Samples_0	30/03/2019 14:43	File con valori separati da virgola (CSV) ...	6.362 KB

Figura 5-2: file CSV di origine (sottogruppo)

L'unione dei singoli file CSV, secondo quanto descritto nei paragrafi successivi, va a costituire un database di circa 315.000 voci, ognuna delle quali composta da circa 500 variabili. Un piccolo estratto è riportato in Figura 5-3.

B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
VehicleID	DepotFlag	Process	EventId	SubSystem	StartTime	EndTime	Duration[seconds]	ErrorCode	Priority	Name	Description	VCUState	DCUState	Speed	ConfigLoco	Baseline	SampleTime	NumeroTreno
519	False	VCUMON	111		02/01/2019 05:28	02/01/2019 05:28	1	2949231	C	MON_Alm-1-Pitto-att	Nuovo pittogramma attivo	4260	40000	0	30	005.004.000.000	384	
519	False	VCUMON	118		02/01/2019 05:28	02/01/2019 05:28	3	2949238	C	BNA_BNA-Abilitato	Banco di manovra abilitato					005.004.000.000	-128	
519	False	VCUMON	118		02/01/2019 05:28	02/01/2019 05:28	3	2949238	C	BNA_BNA-Abilitato	Banco di manovra abilitato					005.004.000.000	0	
519	False	VCUMON	118		02/01/2019 05:28	02/01/2019 05:28	3	2949238	C	BNA_BNA-Abilitato	Banco di manovra abilitato					005.004.000.000	128	
519	False	VCUMON	111		02/01/2019 05:29	02/01/2019 05:29	1	2949231	C	MON_Alm-1-Pitto-att	Nuovo pittogramma attivo	4260	40000	0	30	005.004.000.000	-384	
519	False	VCUMON	111		02/01/2019 05:29	02/01/2019 05:29	1	2949231	C	MON_Alm-1-Pitto-att	Nuovo pittogramma attivo	4260	40000	0	30	005.004.000.000	-256	
519	False	VCUMON	111		02/01/2019 05:29	02/01/2019 05:29	1	2949231	C	MON_Alm-1-Pitto-att	Nuovo pittogramma attivo	4260	40000	0	30	005.004.000.000	-128	
519	False	VCUMON	111		02/01/2019 05:29	02/01/2019 05:29	1	2949231	C	MON_Alm-1-Pitto-att	Nuovo pittogramma attivo	4260	40000	0	30	005.004.000.000	0	
519	False	VCUMON	111		02/01/2019 05:29	02/01/2019 05:29	1	2949231	C	MON_Alm-1-Pitto-att	Nuovo pittogramma attivo	4260	40000	0	30	005.004.000.000	128	
519	False	VCUMON	111		02/01/2019 05:29	02/01/2019 05:29	1	2949231	C	MON_Alm-1-Pitto-att	Nuovo pittogramma attivo	4260	40000	0	30	005.004.000.000	256	
519	False	VCUMON	111		02/01/2019 05:29	02/01/2019 05:29	1	2949231	C	MON_Alm-1-Pitto-att	Nuovo pittogramma attivo	4260	40000	0	30	005.004.000.000	384	
519	False	VCUMON	122		02/01/2019 05:35	02/01/2019 05:35	3	2949242	C	CAZ_Imp-Enable	Sblocco impulsi					005.004.000.000	-128	
519	False	VCUMON	122		02/01/2019 05:35	02/01/2019 05:35	3	2949242	C	CAZ_Imp-Enable	Sblocco impulsi					005.004.000.000	0	
519	False	VCUMON	122		02/01/2019 05:35	02/01/2019 05:35	3	2949242	C	CAZ_Imp-Enable	Sblocco impulsi					005.004.000.000	128	
519	False	VCUMON	122		02/01/2019 05:47	02/01/2019 05:53	356	2949242	C	CAZ_Imp-Enable	Sblocco impulsi					005.004.000.000	-128	
519	False	VCUMON	122		02/01/2019 05:47	02/01/2019 05:53	356	2949242	C	CAZ_Imp-Enable	Sblocco impulsi					005.004.000.000	0	
519	False	VCUMON	122		02/01/2019 05:47	02/01/2019 05:53	356	2949242	C	CAZ_Imp-Enable	Sblocco impulsi					005.004.000.000	128	
519	False	VCUMON	113		02/01/2019 05:47	02/01/2019 05:53	347	2949233	C	TELEDIA_Speed>5km/h	Velocita' maggiore di 5km/h					005.004.000.000	-128	
519	False	VCUMON	113		02/01/2019 05:47	02/01/2019 05:53	347	2949233	C	TELEDIA_Speed>5km/h	Velocita' maggiore di 5km/h					005.004.000.000	0	
519	False	VCUMON	113		02/01/2019 05:47	02/01/2019 05:53	347	2949233	C	TELEDIA_Speed>5km/h	Velocita' maggiore di 5km/h					005.004.000.000	128	
519	False	VCUMON	114		02/01/2019 05:47	02/01/2019 05:48	51	2949234	C	TELEDIA_SforzoSviTrax>0	Sforzo di trazione reso >0					005.004.000.000	-128	
519	False	VCUMON	114		02/01/2019 05:47	02/01/2019 05:48	51	2949234	C	TELEDIA_SforzoSviTrax>0	Sforzo di trazione reso >0					005.004.000.000	0	
519	False	VCUMON	114		02/01/2019 05:47	02/01/2019 05:48	51	2949234	C	TELEDIA_SforzoSviTrax>0	Sforzo di trazione reso >0					005.004.000.000	128	
519	False	VCUMON	114		02/01/2019 05:49	02/01/2019 05:50	25	2949234	C	TELEDIA_SforzoSviTrax>0	Sforzo di trazione reso >0					005.004.000.000	-128	
519	False	VCUMON	114		02/01/2019 05:49	02/01/2019 05:50	25	2949234	C	TELEDIA_SforzoSviTrax>0	Sforzo di trazione reso >0					005.004.000.000	0	
519	False	VCUMON	114		02/01/2019 05:49	02/01/2019 05:50	25	2949234	C	TELEDIA_SforzoSviTrax>0	Sforzo di trazione reso >0					005.004.000.000	128	
519	False	VCUMON	115		02/01/2019 05:50	02/01/2019 05:50	7	2949235	C	TELEDIA_SforzoSviFre<0	Sforzo di trazione reso <0					005.004.000.000	-128	
519	False	VCUMON	115		02/01/2019 05:50	02/01/2019 05:50	7	2949235	C	TELEDIA_SforzoSviFre<0	Sforzo di trazione reso <0					005.004.000.000	0	
519	False	VCUMON	115		02/01/2019 05:50	02/01/2019 05:50	7	2949235	C	TELEDIA_SforzoSviFre<0	Sforzo di trazione reso <0					005.004.000.000	128	
519	False	VCUMON	114		02/01/2019 05:50	02/01/2019 05:51	58	2949234	C	TELEDIA_SforzoSviTrax>0	Sforzo di trazione reso >0					005.004.000.000	-128	
519	False	VCUMON	114		02/01/2019 05:50	02/01/2019 05:51	58	2949234	C	TELEDIA_SforzoSviTrax>0	Sforzo di trazione reso >0					005.004.000.000	0	
519	False	VCUMON	114		02/01/2019 05:50	02/01/2019 05:51	58	2949234	C	TELEDIA_SforzoSviTrax>0	Sforzo di trazione reso >0					005.004.000.000	128	
519	False	VCUMON	115		02/01/2019 05:52	02/01/2019 05:52	19	2949235	C	TELEDIA_SforzoSviFre<0	Sforzo di trazione reso <0					005.004.000.000	-128	
519	False	VCUMON	115		02/01/2019 05:52	02/01/2019 05:52	19	2949235	C	TELEDIA_SforzoSviFre<0	Sforzo di trazione reso <0					005.004.000.000	0	
519	False	VCUMON	115		02/01/2019 05:52	02/01/2019 05:52	19	2949235	C	TELEDIA_SforzoSviFre<0	Sforzo di trazione reso <0					005.004.000.000	128	
519	False	VCUMON	115		02/01/2019 05:52	02/01/2019 05:52	16	2949235	C	TELEDIA_SforzoSviFre<0	Sforzo di trazione reso <0					005.004.000.000	-128	
519	False	VCUMON	115		02/01/2019 05:52	02/01/2019 05:52	16	2949235	C	TELEDIA_SforzoSviFre<0	Sforzo di trazione reso <0					005.004.000.000	0	
519	False	VCUMON	115		02/01/2019 05:52	02/01/2019 05:52	16	2949235	C	TELEDIA_SforzoSviFre<0	Sforzo di trazione reso <0					005.004.000.000	128	
519	False	VCUMON	111		02/01/2019 05:52	02/01/2019 05:52	1	2949231	C	MON_Alm-1-Pitto-att	Nuovo pittogramma attivo	4400	40050	62,26991	2	005.004.000.000	-384	
519	False	VCUMON	111		02/01/2019 05:52	02/01/2019 05:52	1	2949231	C	MON_Alm-1-Pitto-att	Nuovo pittogramma attivo	4400	40050	62,26991	2	005.004.000.000	-256	
519	False	VCUMON	111		02/01/2019 05:52	02/01/2019 05:52	1	2949231	C	MON_Alm-1-Pitto-att	Nuovo pittogramma attivo	4400	40050	62,26991	2	005.004.000.000	-128	
519	False	VCUMON	111		02/01/2019 05:52	02/01/2019 05:52	1	2949231	C	MON_Alm-1-Pitto-att	Nuovo pittogramma attivo	4400	40050	62,26991	2	005.004.000.000	0	
519	False	VCUMON	111		02/01/2019 05:52	02/01/2019 05:52	1	2949231	C	MON_Alm-1-Pitto-att	Nuovo pittogramma attivo	4400	40050	62,26991	2	005.004.000.000	128	
519	False	VCUMON	111		02/01/2019 05:52	02/01/2019 05:52	1	2949231	C	MON_Alm-1-Pitto-att	Nuovo pittogramma attivo	4400	40050	62,26991	2	005.004.000.000	256	
519	False	VCUMON	111		02/01/2019 05:52	02/01/2019 05:52	1	2949231	C	MON_Alm-1-Pitto-att	Nuovo pittogramma attivo	4400	40050	62,26991	2	005.004.000.000	384	
519	False	VCUMON	115		02/01/2019 05:53	02/01/2019 05:53	3	2949235	C	TELEDIA_SforzoSviFre<0	Sforzo di trazione reso <0					005.004.000.000	-128	
519	False	VCUMON	115		02/01/2019 05:53	02/01/2019 05:53	3	2949235	C	TELEDIA_SforzoSviFre<0	Sforzo di trazione reso <0					005.004.000.000	0	
519	False	VCUMON	115		02/01/2019 05:53	02/01/2019 05:53	3	2949235	C	TELEDIA_SforzoSviFre<0	Sforzo di trazione reso <0					005.004.000.000	128	

Figura 5-3: estratto database DDS

5.2 Pulizia e Armonizzazione dei Dati

I dati ricevuti in input contengono diverse impurità e non sono nel formato più adeguato e funzionale per il lavoro da svolgere. Per questo motivo, prima di iniziare lo sviluppo di un algoritmo per la loro analisi, risulta imprescindibile una fase di pulizia e trasformazione. Fondamentalmente ci si concentrerà sulla rimozione di dati rumorosi o incompleti dalla raccolta, oltre che all'armonizzazione delle variabili, che in alcuni casi sono riportate in formati differenti (es. incongruenze sul separatore decimale; data/ora in formati differenti, ecc.…).

Il data-set, inoltre, facendo riferimento estrazioni distinte su diversi processori, contiene convenzioni di denominazione delle variabili difformi che causano ridondanze nel database unificato. Per di più l'ordinamento e il numero delle variabili (colonne) è discordante tra i diversi file CSV. È necessario, quindi, eseguire un'ulteriore elaborazione per rimuovere le ridondanze e le incongruenze dall'integrazione dei dati ed evitare di comprometterne l'affidabilità.

Il lavoro di pulizia e armonizzazione è stato eseguito durante la fase di importazione dei dati in Excel, tramite l'Editor di Power Query.

Figura 5-4: flusso di estrazione dati e armonizzazione

Le maggiori difformità sono emerse tra i gruppi di dati esportati nei due periodi differenti. Per snellire il processo di importazione e ridurre eventuali errori si è deciso, pertanto, di effettuare una prima importazione dei dati con due query

indipendenti: una per i DDS relativi al periodo Mag. 2015 - Mar. 2016, l'altra per quelli relativi a Gen. 2019 - Mag. 2019.

Figura 5-5: query DDS

Dalle query precedenti risultano, quindi, due file contenenti ciascuno circa 150,000 record. Ogni stringa è costituita da circa 500 variabili ambientali, molte delle quali potenzialmente inutilizzabili. Alla luce di questa osservazione, si è deciso di ridurre il numero di variabili, eliminando:

- le colonne contenenti meno di 1000 valori;
- le colonne contenenti variabili/dati poco influenti e/o di difficile interpretazione.

Al termine della selezione restano *70 variabili ambientali*, riportate in Tabella 5-2, con le relative unità di misura, ove presenti.

Tabella 5-2: variabili e dati ambientali

NOME VARIABILE	UNITÀ	NOME VARIABILE	UNITÀ
Vehicle		Pressione H2O	[hPa]
Depot/Test		Effort	[kN]
Process		GPS Longitudine (minuti)	
Event Id		GPS Latitudine (minuti)	
Start Time		GPS Longitudine (gradi)	
End Time		GPS Latitudine (gradi)	
Duration (seconds)	[sec]	Sforzo realizzato da DCU	[kN]
Error Code		Sforzo richiesto verso DCU	[kN]
Priority		Tensione di linea istantanea	[V]
Name		Tensione di batteria attuale	[V]
Description		Temperatura riduttore 4	[°C]
Speed	[km/h]	Temperatura riduttore 3	[°C]
ConfigLoco		Temperatura riduttore 2	[°C]
Numero Treno		Temperatura riduttore 1	[°C]
Corrente di linea	[A]	Temperatura PT100/2 riduttore 4	[°C]

NOME VARIABILE	UNITÀ	NOME VARIABILE	UNITÀ
Tensione di linea	[V]	Temperatura PT100/1 riduttore 4	[°C]
Potenza di linea	[MW]	Temperatura PT100/2 riduttore 3	[°C]
Tensione di DC-Link	[V]	Temperatura PT100/1 riduttore 3	[°C]
Velocita' di riferimento del treno	[km/h]	Temperatura PT100/2 riduttore 2	[°C]
Media velocita' asse acquisita	[fsn2]	Temperatura PT100/1 riduttore 2	[°C]
Sforzo traz/fren richiesto	[kN]	Temperatura PT100/2 riduttore 1	[°C]
Sforzo traz/fren sviluppato	[kN]	Temperatura PT100/1 riduttore 1	[°C]
Riduzione % del Traction Contr	[%]	Pressione in Condotta Generale	[mBar]
Tensione semifiltro sup. INV1	[V]	Pressione Cilindro Freno Anteriore	[mBar]
Tensione semifiltro inf. INV2	[V]	Pressione Cilindro Freno Posteriore	[mBar]
Potenza reostato di frenatura	[kW]	Temperatura esterna climatizzatore	[°C]
Temperatura motore 1 sonda 1	[°C]	Conducibilita' acqua raffreddamento	[µS]
Temperatura motore 1 sonda 2	[°C]	Pressione acqua raffreddamento	[hPa]
Temperatura motore 2 sonda 1	[°C]	Temperatura acqua raffreddamento	[°C]
Temperatura motore 2 sonda 2	[°C]	Position	
Temperatura motore 3 sonda 1	[°C]	Temperatura Motore 4	[°C]
Temperatura motore 3 sonda 2	[°C]	Temperatura Motore 3	[°C]
Temperatura motore 4 sonda 1	[°C]	Temperatura Motore 2	[°C]
Temperatura motore 4 sonda 2	[°C]	Temperatura Motore 1	[°C]
Conducibilita' H2O	[µS]	Temperatura H2O	[°C]

Seguentemente si passa all'armonizzazione delle variabili dei due data-set estratti. L'obiettivo è quello di rimuovere le difformità tra i due database e coadiuvare la formazione di un grande archivio unificato.

Le principali discordanze riscontrate tra i due gruppi di dati sono relative ai formati data/ora e ai separatori decimali per le variabili reali.

In particolare, le colonne "*Start Time*" ed "*End Time*" sono in formato americano (es. 10/31/2016 09:56 PM) nei dati meno recenti, mentre sono in formato italiano (es. 31/10/2016 21:56) in quelli più attuali. Per alcune variabili reali i separatori decimali utilizzati sono in alcuni casi la virgola (es. 0,3) e in altri il punto (es. 0.3).

Query 1-DDS
Mag. 2015- Mar. 2016

Start Time	End Time
02/01/2019 03:41:53	02/01/2019 03:41:54
02/01/2019 03:41:53	02/01/2019 03:41:54
02/01/2019 03:41:53	02/01/2019 03:41:54

Speed	Corrente di linea
54,16579	-285,155518
54,16579	-283,446533
54,16579	-302,245331
77,98995	54,687359
77,98995	56,396339

Query 2- DDS
Gen. 2019- Mag. 2019

Start Time	End Time
1/27/2016 8:56:02 AM	
1/27/2016 9:14:54 AM	1/27/2016 9:14:54 AM
1/27/2016 9:15:17 AM	1/27/2016 9:15:17 AM

Speed	Corrente di linea
19.92	962,155762
33.4	895,505493
11.33	967,282654
45.04	859,616943
44.04	1.153,561523

Figura 5-6: alcune delle difformità sulle variabili

L'armonizzazione, sempre eseguita attraverso l'Editor di Power Query in Excel, è stata effettuata manualmente andando a indicare, nelle colonne di competenza, il formato corretto per ognuna delle variabili.

Figura 5-7: modifica tipo dati con impostazioni locali

Terminata la fase di armonizzazione, si è passato all'unione dei due database, eseguita anch'essa tramite query in Excel. Il risultato è un set di dati unico, contenente i DDS registrati in un anno e mezzo da cinque locomotive E464, operative sul territorio nazionale tra Maggio 2015 e Maggio 2019.

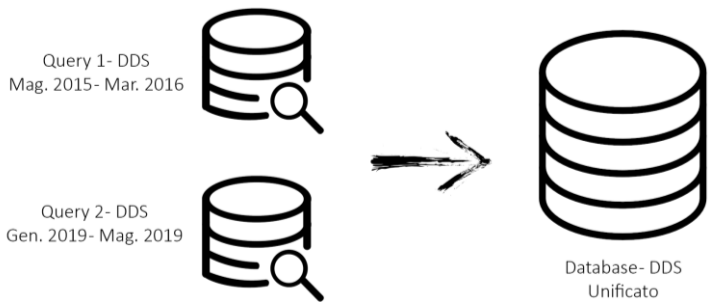

Figura 5-8: database DDS unificato

5.3 Creazione Modello Associazioni

L'estrazione delle regole di associazione da un set di dati è implementata attraverso l'algoritmo *FP-Growth* in *RapidMiner*.

5.3.1 RapidMiner

RapidMiner è una piattaforma software per il data science sviluppata dall'omonima azienda. Fornisce un ambiente integrato per la preparazione dei dati, machine learning, deep learning, text mining e analisi predittiva.

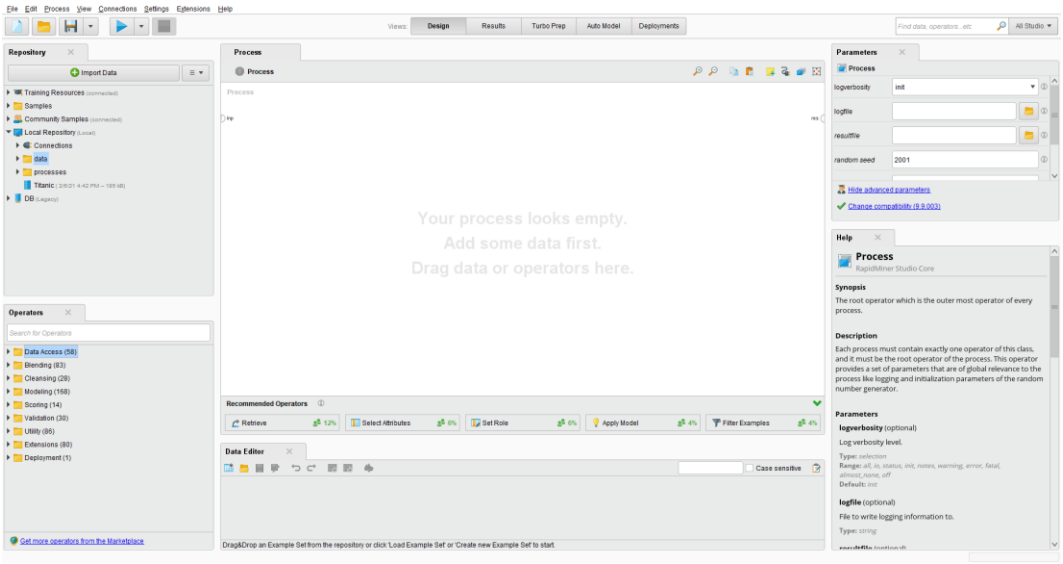

Figura 5-9: RapidMiner GUI

Viene utilizzato sia in ambito aziendale e commerciale, per la ricerca, l'istruzione, la formazione, la prototipazione rapida e lo sviluppo di applicazioni.

RapidMiner, permette di importare dati da diverse sorgenti, quali fogli Excel, tabelle in Access e CSV. Ha al suo interno tutti gli operatori necessari ad implementare un intero workflow che va dalla fase di preparazione dei dati fino a quella di validazione del modello e relativa ottimizzazione (Hofmann e Klinkenberg 2013).

RapidMiner è scritto in Java e offre una GUI per progettare ed eseguire flussi di lavoro analitici, come mostrato in Figura 5-9. Questi workflow vengono chiamati "*Processi*" e si compongono di più "*Operatori*": ogni operatore esegue un

singolo compito all'interno del processo e l'output costituisce l'input del successivo operatore (Figura 5-10).

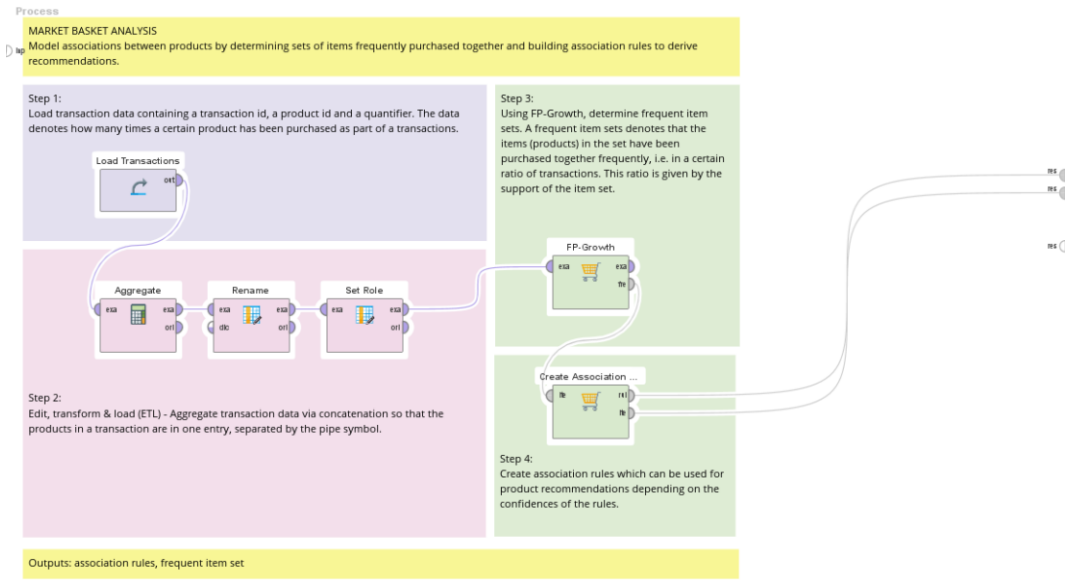

Figura 5-10: flusso di lavoro in RapidMiner. Operatori in un processo di Basket Analysis.

5.3.2 Impostazione del programma e del lavoro

In questo paragrafo si esaminerà in modo dettagliato l'iter sperimentale adottato, a partire dall'importazione e preparazione dei dati fino alla modellazione all'interno del software.

5.3.2.1 Passo 0: Importazione dei dati

Preliminarmente all'analisi vera e propria, si inseriscono i dati in RapidMiner: il trasferimento può avvenire con diverse modalità e a partire da file di formato differente. Nell'ambito di questo lavoro di tesi si è scelto di importare i file nell'archivio centrale di RapidMiner, chiamato *Repository*: questa soluzione, specialmente quando i dati provengono da file come XLS o CSV, permette di semplificare di gran lunga la progettazione del processo analitico, poiché memorizza i metadati descrittivi insieme ai dati.

L'importazione è eseguita a partire dall'Excel elaborato nel §5.2, contenente i DDS e le relative variabili ambientali delle 5 locomotive E464. A livello pratico,

è sufficiente cliccare su *Import Data* nel pannello *Repository* (Figura 5-11) e seguire i passi della procedura guidata. Una volta completata l'importazione, i dati verranno memorizzati con il nome scelto nella sottocartella *Local Repository*.

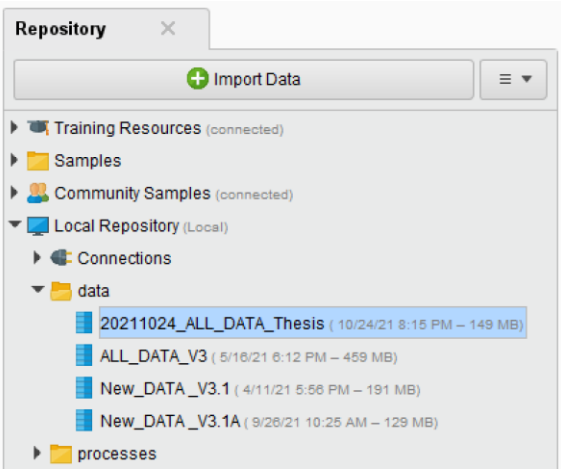

Figura 5-11: pannello "Repository"

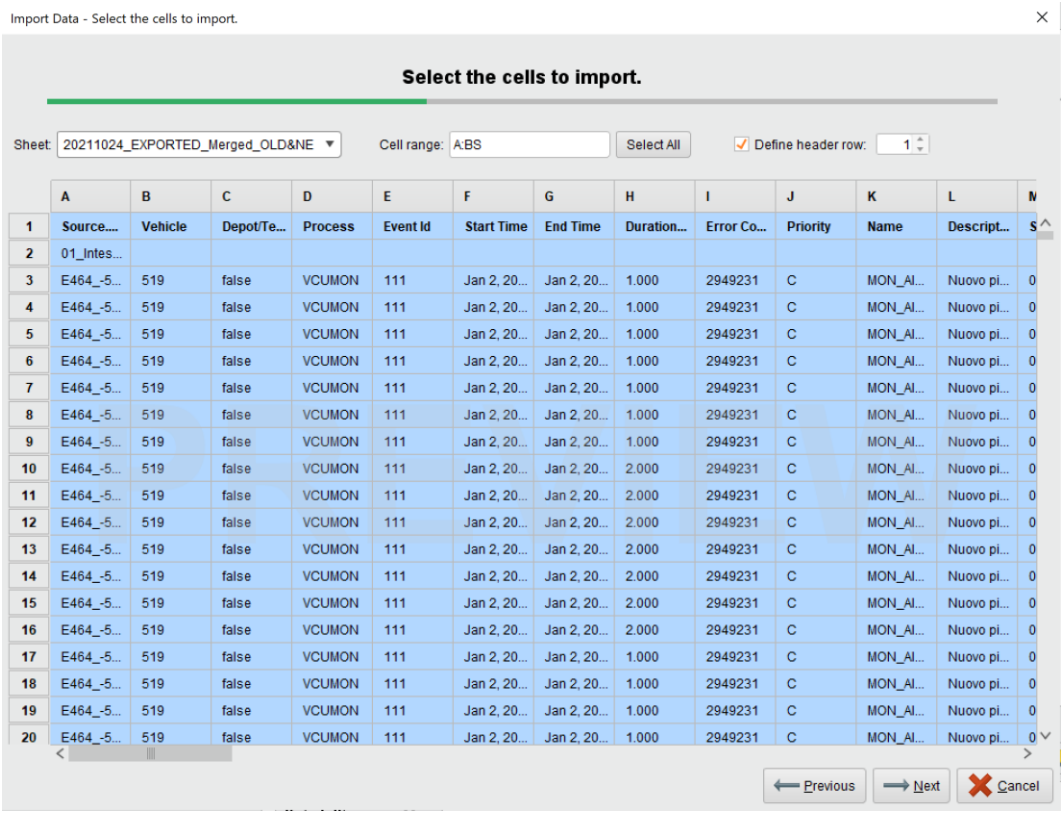

Figura 5-12: Procedura guidata per l'importazione dei dati

Uno dei passi più significativi della procedura d'importazione è quello di verificare che le variabili vengano interpretate da RapidMiner nel formato più consono (es. integer, polynomial, text, real, ecc…): l'immissione corretta dei formati è essenziale se si vogliono garantire risultati accurati. Per cambiare tipologia è sufficiente, in fase di importazione, cliccare nell'apposito menù a tendina di fianco all'header di ogni colonna (Figura 5-13).

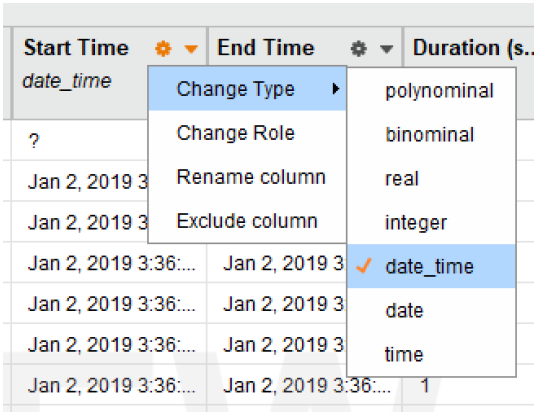

Figura 5-13: cambio del formato di dati

Importati i dati, questi possono essere trascinati dal repository all'interno del processo vero e proprio. Qui verrà creato un operatore di caricamento (Figura 5-14) che caricherà effettivamente i dati solo quando si eseguirà il processo.

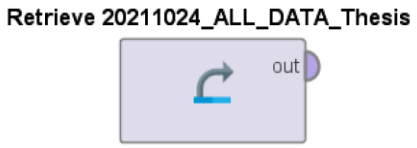

Figura 5-14: retrieve, operatore di caricamento

5.3.2.2 Passo 1: Preparazione dei dati

Il processo di analisi delle associazioni richiede che la tabella di dati in input sia in un determinato formato, cioè deve essere composta da dati binomiali (vero o falso) con le variabili di ogni DDS distribuiti nelle diverse colonne.

Row No.	Temperatur...	Temperatur...	Temperatur...	Temperatur...
89	false	true	false	false
90	false	true	false	false
91	false	true	false	false
92	false	false	true	false
93	false	false	true	false
94	false	true	false	false

Figura 5-15: tabella di input all'operatore FP-Growth

I set di dati devono quindi essere convertiti nel formato descritto in precedenza, e mostrato in Figura 5-15, mediante l'utilizzo degli operatori di trasformazione dei dati.

La conversione in formato binomiale viene eseguita attraverso l'operatore *Nominal to Binominal* e l'output che ne deriva è poi collegato all'operatore FP-Growth che genererà gli insiemi di elementi frequenti.

In Figura 5-16 è riportata la sequenza di operatori selezionati per il processo di mining delle Regole d'Associazione in RapidMiner. L'algoritmo scelto per la ricerca degli itemset frequenti è l'FP-Growth, uno dei più diffusi nel Data Mining in quanto più efficiente e veloce dei suoi concorrenti.

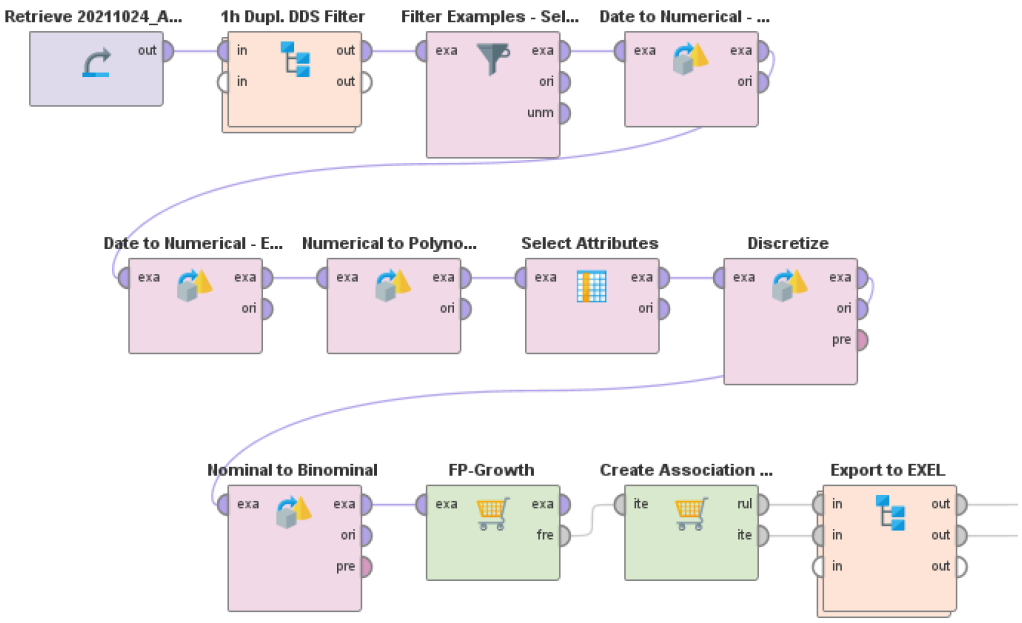

Figura 5-16: processo di analisi con algoritmo FP-Growth

Il funzionamento degli operatori utilizzati è riportato di seguito:

Select Attributes: questo operatore seleziona un sottoinsieme di attributi di un ExampleSet e rimuove gli altri attributi, cioè fornisce diversi tipi di filtro per facilitare la selezione degli attributi. Le opzioni sono, ad esempio: selezione diretta degli attributi, selezione tramite un'espressione regolare o selezione dei soli attributi senza valori mancanti. Inoltre, il parametro "inverti selezione" inverte la selezione. Gli attributi speciali, cioè attributi con ruoli (id, label, weight), sono di default ignorati nel processo di selezione e rimarranno sempre nell'ExampleSet risultante. Il parametro "include special attributes" ha la funzione di estendere l'effetto della selezione anche agli attributi speciali.

In definitiva, solo gli attributi selezionati sono trasmessi alla porta di output, il resto viene rimosso dall'ExampleSet.

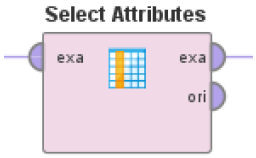

Figura 5-17: operatore Select Attribute

Nell'ambito del lavoro di tesi, Select Attributes viene utilizzato per selezionare un sottoinsieme di attributi dal database fornito e quelli non ritenuti utili ai fini dell'analisi vengono rimossi.

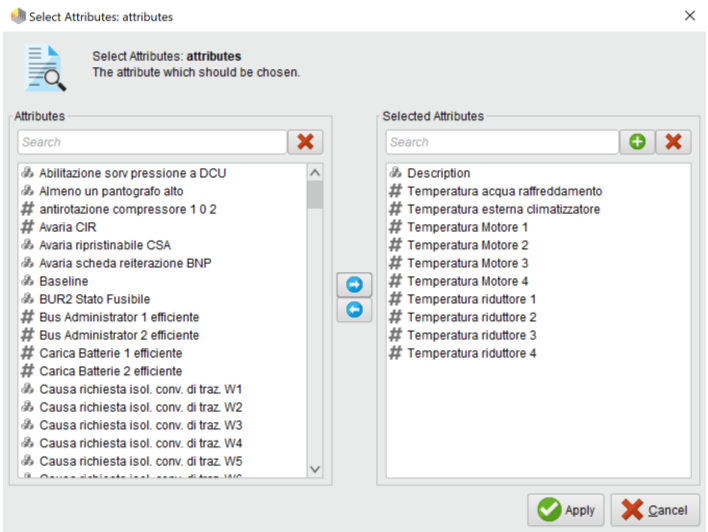

Figura 5-18: Pannello Select Attributes

1h Dupl. DDS Filter: il sottoprocesso è stato realizzato per rimuovere quei DDS che compaiono innumerevoli volte in intervalli di tempo limitati.

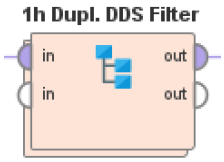

Figura 5-19: sottoprocesso 1h Dupl. DDS Filter

Capita frequentemente, infatti, che la macchina invii in modo ripetuto e persistente dei DDS relativi alla stessa anomalia nell'arco di poche ore e in alcune situazioni l'invio di questi record persiste anche per tutta la durata del viaggio.

La presenza numerosa di queste segnalazioni finirebbe con il condizionare i risultati: sarebbero individuate molte regole con supporto elevato ma di fatto con poca utilità e quelle con supporto inferiore, invece, rimarrebbero nascoste. Per ovviare a questo problema, si è deciso di applicare un filtro atto a ridurre il numero di questi DDS "duplicati": ad esempio se l'anomalia "Effort > 0" venisse ripetuta 360 volte nell'arco di 2h, il filtro eliminerebbe i valori ripetuti lasciando un unico DDS per tipologia in ogni ora.

Filter Examples: l'operatore seleziona quali elementi di un ExampleSet vengono mantenuti e quali vengono rimossi, restituendo gli Example (righe dei singoli DDS) che corrispondono alla condizione data. Le regole sono fissate dall'utente ma esistono anche diverse opzioni predefinite. La funzione può ridurre il numero di Exemple ma non ha effetto sul numero di Attributi (colonne).

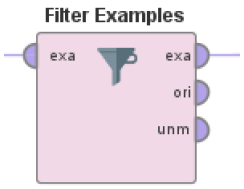

Figura 5-20: operatore Filter Examples

Filter Examples viene utilizzato nel lavoro di tesi per eliminare quei DDS contenenti valori mancanti nelle variabili considerate nell'analisi.

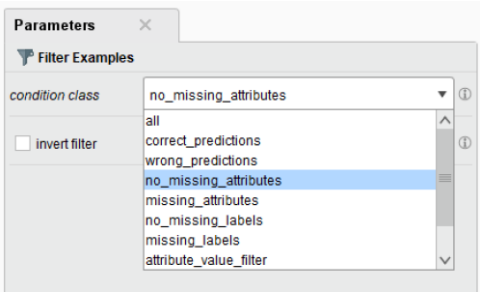

Figura 5-21: pannello Filter Examples

Date to Numerical: questo operatore converte un attributo di tipo data in attributo di tipo numerico, cioè permette di specificare esattamente quale elemento della data o dell'ora deve essere estratto definendo l'elemento di riferimento. Per esempio, è possibile estrarre il giorno rispetto alla settimana, rispetto al mese o rispetto all'anno.

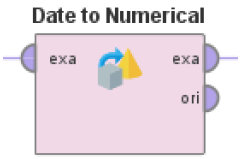

Figura 5-22: operatore Date to Numerical

Supponiamo che la data sia 15/feb/2012: il giorno relativo al mese sarebbe 15 perché è il 15° giorno del mese e il giorno relativo all'anno sarebbe 46 perché questo è il 46° giorno dell'anno. Tutte le componenti della data e dell'ora possono essere estratte in relazione alle componenti madri più comuni.

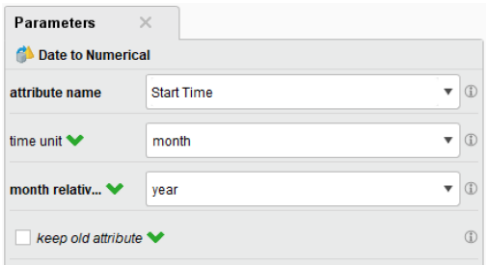

Figura 5-23: pannello Date to Numerical

Nell'ambito della tesi Date to Numerical viene utilizzato in alcuni run di analisi per convertire i riferimenti temporali di ogni DDS (Start Time, End Time) in valori numerici (es. i valori di Start Time relativi ai mesi di Gennaio acquisteranno valore 1, quelli del mese di Marzo valore 3). L'obiettivo è quello di verificare se determinati guasti o alert si ripetono più frequentemente in determinati periodi dell'anno (es. nei mesi freddi) escludendo l'ulteriore discretizzazione definita dal considerare lo specifico in cui essi avvengono.

Numerical to Polynominal: l'operatore converte gli attributi numerici selezionati in attributi di tipo polinomiale e mappa anche tutti i valori nei loro corrispondenti polinomiali, in pratica a ogni valore numerico associa un corrispondente valore polinomiale. Poiché gli attributi numerici possono avere un enorme quantità di valori diversi, anche in un piccolo intervallo, la conversione potrebbe generare un enorme numero di valori polinomiali. Ciò potrebbe non risultare molto utile e potrebbe aumentare significativamente l'uso della memoria. Quindi, nel caso in cui il numero di valori numerici sia elevato, sarà meglio utilizzare gli operatori di discretizzazione.

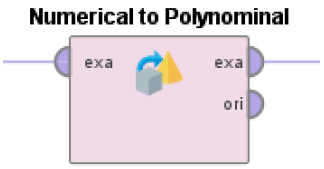

Figura 5-24: operatore Numerical to Polynominal

Nell'ambito del lavoro di tesi Numerical to Polynominal è stato utilizzato per la conversione di particolari valori numerici (es. numero treno, numero veicolo, config loco, ecc…) in valori polinomiali. Risulta infatti inesatto considerare questi valori come numerici e suddividerli in gruppi.

Discretize by Binning: questo operatore discretizza gli attributi numerici selezionati in un numero di bin specificato dall'utente. Vengono generati automaticamente dei bin di uguale ampiezza, convertiti poi in attributi nominali.

Figura 5-25: operatore Discretize by Binning

Il parametro "number of bins" viene utilizzato per specificare il numero di bin richiesto. La gamma di valori numerici è partizionata in intervalli di uguale dimensione, dove ogni intervallo rappresenta un bin.

Gli intervalli sono nominati automaticamente secondo il range valori che cadono in quel determinato intervallo.

I parametri "min value" e "max value" sono utilizzati per definire i limiti dell'intervallo. Nel caso in cui ci siano dei valori inferiori al parametro "min value", per questi valori viene creato un intervallo separato. Allo stesso modo, se ci sono dei valori maggiori del parametro "max value", viene creato un intervallo separato. La discretizzazione per binning viene eseguita solo sui valori che sono all'interno dei confini specificati.

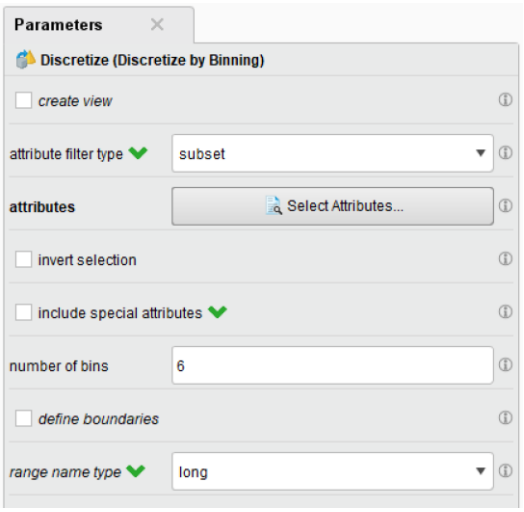

Figura 5-26: pannello Discretize by Binning

Nominal to Binominal: l'operatore trasforma gli attributi nominali selezionati in attributi di tipo binomiale, mappando anche tutti i valori in formato binomiale, ossia in vero e falso (true/false).

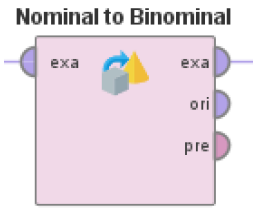

Figura 5-27: operatore Nominal to Binominal

Gli attributi numerici dell'ExampleSet in input rimangono invariati.

L'operatore Nominal to Binominal è fondamentale per la conversione degli attributi nei DDS nel formato richiesto dall'operatore FP-Growth (vedi Figura 5-15).

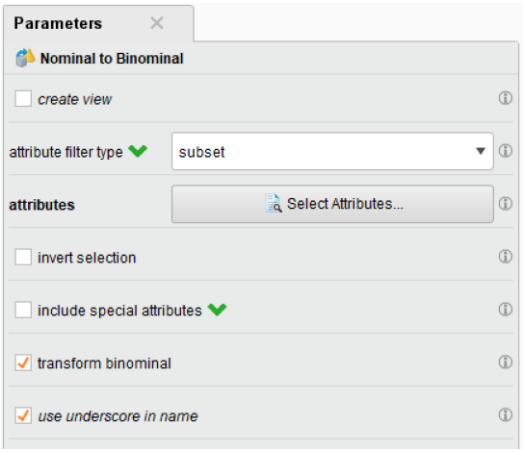

Figura 5-28: pannello Nominal to Binominal

5.3.2.3 Passo 2: Operatore di modellazione e parametri

L'operatore **FP-Growth** in RapidMiner genera, dal set di dati in input, tutti i frequent item set (insiemi di elementi frequenti) che soddisfano certi parametri, utilizzando la struttura dati FP-tree.

Figura 5-29: Operatore FP-Growth

L'operatore è disponibile nella cartella Modeling > Associations, come mostrato in Figura 5-30.

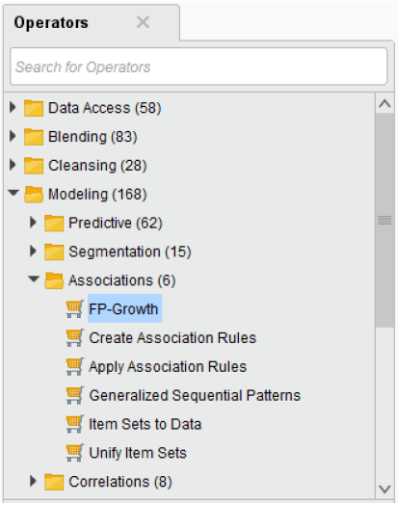

Figura 5-30: Localizzazione operatore FP-Growth

Può lavorare in due modalità: una con un numero specificato di item set ad alto supporto (default) e l'altra con criteri di supporto minimo fissati.

I seguenti parametri possono essere impostati nella scheda dell'operatore influenzando il comportamento del modello:

- *Min Support:* definisce la soglia minima di supporto. Tutti gli insiemi di elementi frequenti che superano questa soglia saranno forniti nell'output. Diminuendo questo valore cresce il numero di itemset nel risultato.
- *Max Number of Itemsets:* numero massimo di elementi in un insieme di elementi. Specificando questo parametro si limita il numero eccessivo di elementi in un insieme di elementi.
- *Must Contain List:* espressione regolare per filtrare gli itemset in modo che contengano gli elementi specificati.

- *Find Min Number of Itemsets:* questa opzione permette all'operatore FP-Growth di abbassare la soglia di supporto, fino a che i risultati non conterranno il numero minimo di itemset specificato. La soglia di supporto viene diminuita del 20% ad ogni tentativo fino a che l'obiettivo non viene raggiunto.
- *Min Number of Itemsets:* valore del numero minimo di set di elementi da generare.
- *Max Number of Retries:* numero massimo di tentativi consentiti per raggiungere il numero minimo di insiemi di elementi. Questo parametro è disponibile solo quando è selezionato "find min number of itemsets". Nella fase di decrescita automatica del valore di supporto minimo/frequenza minima, determina quante volte l'operatore può diminuire il valore prima di rinunciare. Aumentando questo numero si ottengono più risultati.

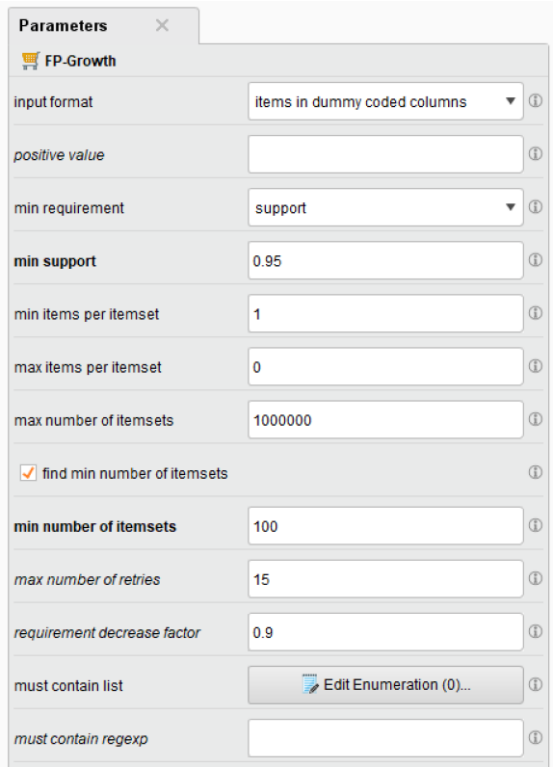

Figura 5-31: Parametri operatore FP-Growth

L'operatore **Create Association Rules** genera un insieme di regole di associazione dall'insieme dato di itemset frequenti. Le regole di associazione sono dichiarazioni if/then che aiutano a scoprire relazioni tra dati apparentemente non correlati.

Figura 5-32: operatore Create Association Rules

Un esempio di una regola di associazione potrebbe essere "Se un cliente compra del prosciutto, ha l'80% di probabilità di acquistare anche del pane". Una regola di associazione ha due parti, un antecedente (if) e un conseguente (then). Un antecedente è un elemento (o un insieme di elementi) trovato nei dati. Un conseguente è un elemento (o un insieme di elementi) che si trova in combinazione con l'antecedente.

Le regole di associazione sono create analizzando i dati per i pattern if/then frequenti, utilizzando i criteri di supporto e confidenza per identificare le relazioni più importanti.

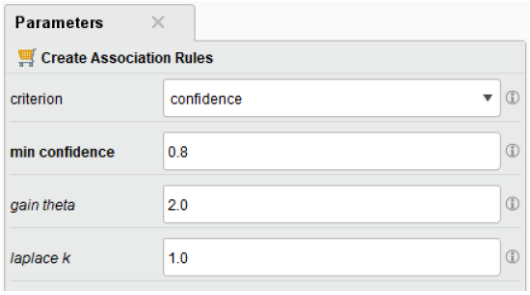

Figura 5-33: parametri operatore Create Association Rules

Il supporto è un'indicazione di quanto frequentemente gli elementi appaiono nel database. La confidenza indica il numero di volte in cui le dichiarazioni if/then sono state trovate vere. I pattern if/then frequenti sono estratti usando l'operatore FP-Growth. Create Association Rules prende questi itemset frequenti e genera regole di associazione.

5.4 Applicazione al Caso Studio

Il caso studio riguarda l'analisi dei dati di guasto delle locomotive E464 della flotta di Trenitalia. Ogni macchina, durante la marcia, genera degli avvisi (DDS) che vengono inviati in tempo reale nelle sedi preposte al controllo.

Questi avvisi sono descrittori di anomalie, non necessariamente guasti, nei sistemi di bordo monitorati. Ad esempio, un'anomalia potrebbe essere il superamento della soglia di temperatura limite del liquido di raffreddamento, o il raggiungimento del livello minimo di olio nel serbatoio dei motoriduttori.

Le anomalie non generano direttamente nessun blocco o richiamo della macchina, ma vengono analizzate da un insieme di algoritmi diagnostici, che eventualmente emetteranno un avviso di blocco o un avviso manutentivo che verrà espletato al successivo rientro della locomotiva.

Gli algoritmi di analisi dei DDS sono ad oggi generati dagli operatori della sala di controllo, che li realizzano basandosi sulla loro esperienza personale.

L'obiettivo principale di questo lavoro è quello supportare il team tecnico nello sviluppo degli algoritmi diagnostici, ricercando nei dati correlazioni "nascoste" che possano essere sfuggite anche all'occhio esperto dei tecnici.

5.4.1 Definizione dei Run di Analisi

La preparazione del processo di estrazione prevede il settaggio dei diversi parametri degli operatori: le variabili in gioco sono numerose e sono stati effettuati innumerevoli tentativi volti a individuarne la combinazione migliore.

A seguito della calibrazione, si è scelto il run che ha mostrato le regole più significative e, per questo, verrà analizzato più in dettaglio nel capitolo dedicato ai risultati. In riferimento al processo di analisi in Figura 5-16, si riportano di seguito le impostazioni utilizzate per i singoli operatori:

- Sottogruppo *1h Dupl. DDS Filter*: settato per rimuovere tutti quei DDS che risultano "duplicati" in un intervallo di 1h.
- Operatore *Filter Exemples*: non utilizzato nel run selezionato.
- Operatore *Date to Numerical*: predisposto per svincolare le variabili temporali Start Time ed End Time dall'ora, dal giorno e dall'anno in cui sono stati generati.
- Il mese resta l'unico parametro associato ai DDS, in quanto rappresentante sufficientemente accurato delle condizioni ambientali di periodo, che si ripetono ciclicamente anno per anno e quindi comuni ai due set di dati.
- Operatore *Numerical to Polynominal*: per convertire i valori numerici, che identificano i mesi nelle variabili Start Time ed End Time, in valori polinomiali.
- Operatore *Select Attributes*: per escludere quegli attributi che non apportano valore aggiunto. Per il run selezionato sono stati esclusi:
  - "GPS Latitudine" e "GPS Longitudine", in quanto ridondanti con la variabile "Positions";
  - "EventID", "ErrorCode" e "Name" in quanto ridondanti con l'attributo "Description";
- Operatore *Discretize by Binning*: per suddividere i valori di ogni variabile numerica in cinque bins fornendo come output degli attributi nominali.
- Operatore *Nominal to Binominal*: trasforma tutte le variabili da nominali a binomiali, unica tipologia accettata dall'operatore FP-Growth.

5.4.2 Settaggio operatori FP-Growth e Create Association Rules

Il settaggio dei due operatori FP-Growth e Create Association Rules è uno dei passaggi critici per l'estrazione delle regole d'associazione. Due sono i parametri più importanti, *supporto* e *confidenza*, in quanto influenzano direttamente la quantità di regole estratte nel processo di mining. La quantità di regole da

analizzare risulta, infatti, inversamente proporzionale al valore delle soglie di supporto e di confidenza utilizzate.

Nel presente lavoro di tesi, i valori iniziali di supporto e confidenza sono stati impostati rispettivamente a 0.1 e 0.8, ma sono stati gradualmente ridotti per incrementare il numero di regole da analizzare. Il processo di riduzione è stato interrotto al raggiungimento di un numero di regole soddisfacenti.

Non esiste, infatti, un criterio globale per la definizione dei valori di soglia, che risultano, caso per caso, legati allo specifico dataset che si sta analizzando.

In letteratura si annoverano molteplici analisi settate su valori di soglia differenti e valutati al fine di ottenere delle regole che risultassero quanto più utili possibile (Verma, et al. 2014)

È evidente che bisogna trovare il giusto compromesso, in quanto uno spropositato numero di regole genera confusione ma, allo stesso tempo, un numero ridotto (usando valori di soglia più alti) può far sì che non vengano individuate associazioni potenzialmente importanti.

Al termine del processo di iterazione per l'identificazione delle soglie minime, per l'operatore FP-Growth (Figura 5-29) è stato scelto un valore di supporto minimo (min support) di 0.0165.

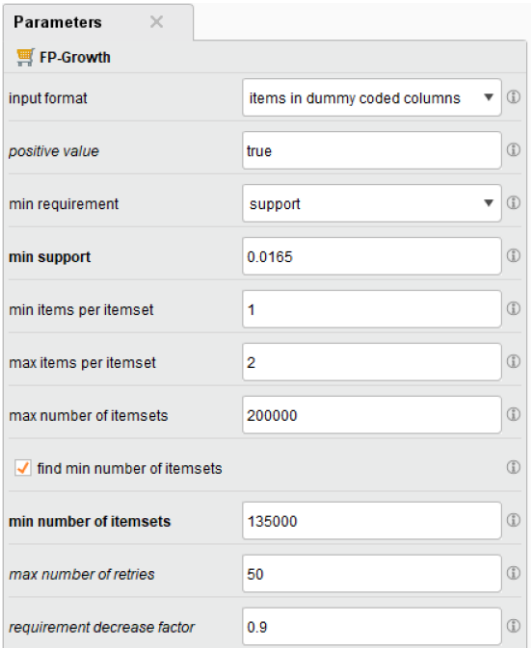

Figura 5-34: settaggi operatore FP-Growth per il caso studio

In ogni caso, l'algoritmo è stato impostato per la ricerca di un minimo numero di itemsets che prescinde dal valore di soglia impostato, nell'eventualità che il numero di associazioni trovato sia troppo basso.

Mentre per l'operatore Create Association Rules il valore minimo di confidenza (min confidence) utilizzato è stato 0.1.

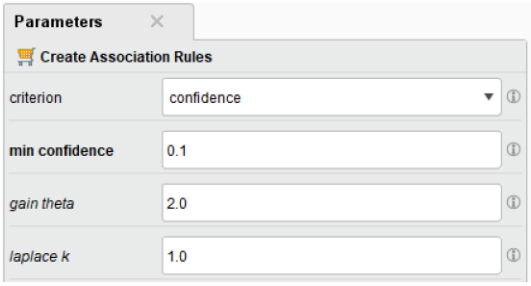

Figura 5-35: settaggi operatore Create Association Rules per il caso studio

6 Presentazione dei risultati

In questo capitolo sono presentati i risultati del lavoro di tesi. Il mining delle Association Rules, effettuato sui DDS raccolti per le cinque locomotive in esame, ha fornito circa 135'000 regole.

In generale, un numero così alto di possibili associazioni comporta il rischio di trovare molte correlazioni spurie, cioè collezioni di elementi che si presentano con una frequenza inaspettata nei dati, ma lo fanno solo per caso. È quindi necessario leggere criticamente i risultati ottenuti, per non incorrere in errori di interpretazione.

Nel caso studio molte regole risultano superflue e di poco rilievo, in quanto descrivono condizioni reali ma che non portano informazioni utili alla diagnostica, altre sembrano potenzialmente valide, ma è difficoltoso valutarne l'effettiva utilità senza il giudizio del personale tecnico della sala di controllo. Per questo è importante, successivamente al mining, passare al vaglio le regole per focalizzarsi su quelle più pertinenti.

6.1 Panoramica dei Risultati Ottenuti

I risultati si presentano sotto forma di tabella: ogni riga rappresenta una regola, i cui descrittori sono gli elementi individuati sulle colonne. I principali sono:

- Antecedente (Premises), un elemento (o un insieme di elementi) trovato nei dati;
- Conseguente (Conclusion), un elemento (o un insieme di elementi) che si trova in combinazione con l'antecedente;
- Supporto (Support), un'indicazione di quanto frequentemente gli elementi appaiono nel database.
- Confidenza (Confidence), indica il numero di volte in cui le affermazioni if/then sono state trovate vere.

Ogni singola regola di associazione è quindi composta da due parti: un antecedente (if) e un conseguente (then), con i valori di *supporto* e *confidenza* a identificare le relazioni più importanti.

Come riportato nel §5.4.1, per l'analisi è stato scelto il run che ha mostrato le regole più significative. Per una lettura più agevole, nella Tabella 6-1 ne è riportato un piccolo estratto contenente le prime 100 voci delle 135'000 totali, elencate in ordine decrescente in riferimento al valore del supporto.

Tabella 6-1: estratto regole di associazione per il run selezionato

ID	Premises	Conclusion	Support	Confidence
1	Duration (seconds)_range1 [-∞ - 136416.800]	Speed_range1 [-∞ - 32.800]	0.519	0.635
2	Speed_range1 [-∞ - 32.800]	Duration (seconds)_range1 [-∞ - 136416.800]	0.519	0.778
3	Speed_range1 [-∞ - 32.800]	Numero Treno_0,00	0.443	0.663
4	Numero Treno_0,00	Speed_range1 [-∞ - 32.800]	0.443	0.816
5	Duration (seconds)_range1 [-∞ - 136416.800]	Numero Treno_0,00	0.424	0.519
6	Numero Treno_0,00	Duration (seconds)_range1 [-∞ - 136416.800]	0.424	0.783
7	Pressione acqua raffreddamento_range5 [2691.399 - ∞]	Tensione di batteria attuale_range5 [29.083 - ∞]	0.361	0.954
8	Tensione di batteria attuale_range5 [29.083 - ∞]	Pressione acqua raffreddamento_range5 [2691.399 - ∞]	0.361	0.992
9	Duration (seconds)_range1 [-∞ - 136416.800]	Priority_C	0.346	0.423
10	Priority_C	Duration (seconds)_range1 [-∞ - 136416.800]	0.346	0.890
11	Numero Treno_0,00	Priority_B	0.343	0.632

ID	Premises	Conclusion	Support	Confidence
12	Priority_B	Numero Treno_0,00	0.343	0.795
13	Pressione acqua raffreddamento_range5 [2691.399 - ∞]	Tensione di linea istantanea_range5 [3401.163 - ∞]	0.342	0.903
14	Tensione di linea istantanea_range5 [3401.163 - ∞]	Pressione acqua raffreddamento_range5 [2691.399 - ∞]	0.342	0.982
15	Temperatura PT100/1 riduttore 3_range2 [30.015 - 109.996]	Temperatura PT100/2 riduttore 3_range2 [30.015 - 109.996]	0.340	0.989
16	Temperatura PT100/2 riduttore 3_range2 [30.015 - 109.996]	Temperatura PT100/1 riduttore 3_range2 [30.015 - 109.996]	0.340	0.993
17	Duration (seconds)_range1 [-∞ - 136416.800]	Priority_B	0.340	0.416
18	Priority_B	Duration (seconds)_range1 [-∞ - 136416.800]	0.340	0.788
19	Temperatura PT100/1 riduttore 3_range2 [30.015 - 109.996]	Temperatura PT100/1 riduttore 4_range2 [30.015 - 109.996]	0.340	0.987
20	Temperatura PT100/1 riduttore 4_range2 [30.015 - 109.996]	Temperatura PT100/1 riduttore 3_range2 [30.015 - 109.996]	0.340	0.997
21	Temperatura PT100/1 riduttore 4_range2 [30.015 - 109.996]	Temperatura PT100/2 riduttore 4_range2 [30.015 - 109.996]	0.340	0.996
22	Temperatura PT100/2 riduttore 4_range2 [30.015 - 109.996]	Temperatura PT100/1 riduttore 4_range2 [30.015 - 109.996]	0.340	0.997
23	Temperatura PT100/1 riduttore 3_range2 [30.015 - 109.996]	Temperatura PT100/2 riduttore 4_range2 [30.015 - 109.996]	0.339	0.986
24	Temperatura PT100/2 riduttore 4_range2 [30.015 - 109.996]	Temperatura PT100/1 riduttore 3_range2 [30.015 - 109.996]	0.339	0.996
25	Temperatura PT100/2 riduttore 3_range2 [30.015 - 109.996]	Temperatura PT100/2 riduttore 4_range2 [30.015 - 109.996]	0.338	0.987
26	Temperatura PT100/2 riduttore 4_range2 [30.015 - 109.996]	Temperatura PT100/2 riduttore 3_range2 [30.015 - 109.996]	0.338	0.993
27	Temperatura PT100/2 riduttore 3_range2 [30.015 - 109.996]	Temperatura PT100/1 riduttore 4_range2 [30.015 - 109.996]	0.338	0.986
28	Temperatura PT100/1 riduttore 4_range2 [30.015 - 109.996]	Temperatura PT100/2 riduttore 3_range2 [30.015 - 109.996]	0.338	0.991
29	Sforzo richiesto verso DCU_range2 [-28.400 - 28.200]	Sforzo realizzato da DCU_range2 [-30.400 - 25.200]	0.335	0.980
30	Sforzo realizzato da DCU_range2 [-30.400 - 25.200]	Sforzo richiesto verso DCU_range2 [-28.400 - 28.200]	0.335	0.983
31	Tensione di batteria attuale_range5 [29.083 - ∞]	Tensione di linea istantanea_range5 [3401.163 - ∞]	0.329	0.904
32	Tensione di linea istantanea_range5 [3401.163 - ∞]	Tensione di batteria attuale_range5 [29.083 - ∞]	0.329	0.945
33	Speed_range1 [-∞ - 32.800]	Priority_B	0.329	0.493
34	Priority_B	Speed_range1 [-∞ - 32.800]	0.329	0.763
35	Sforzo realizzato da DCU_range2 [-30.400 - 25.200]	Effort_range2 [-30.400 - 25.200]	0.325	0.956
36	Effort_range2 [-30.400 - 25.200]	Sforzo realizzato da DCU_range2 [-30.400 - 25.200]	0.325	0.993
37	Duration (seconds)_range1 [-∞ - 136416.800]	Pressione acqua raffreddamento_range5 [2691.399 - ∞]	0.322	0.394
38	Pressione acqua raffreddamento_range5 [2691.399 - ∞]	Duration (seconds)_range1 [-∞ - 136416.800]	0.322	0.851
39	Sforzo richiesto verso DCU_range2 [-28.400 - 28.200]	Effort_range2 [-30.400 - 25.200]	0.320	0.937
40	Effort_range2 [-30.400 - 25.200]	Sforzo richiesto verso DCU_range2 [-28.400 - 28.200]	0.320	0.978
41	Pressione acqua raffreddamento_range5 [2691.399 - ∞]	Temperatura PT100/1 riduttore 3_range2 [30.015 - 109.996]	0.313	0.826
42	Temperatura PT100/1 riduttore 3_range2 [30.015 - 109.996]	Pressione acqua raffreddamento_range5 [2691.399 - ∞]	0.313	0.909
43	Speed_range1 [-∞ - 32.800]	Pressione in Condotta Generale_range1 [-∞ - 1500]	0.313	0.468
44	Pressione in Condotta Generale_range1 [-∞ - 1500]	Speed_range1 [-∞ - 32.800]	0.313	0.988
45	Duration (seconds)_range1 [-∞ - 136416.800]	Tensione di batteria attuale_range5 [29.083 - ∞]	0.312	0.382
46	Tensione di batteria attuale_range5 [29.083 - ∞]	Duration (seconds)_range1 [-∞ - 136416.800]	0.312	0.857

ID	Premises	Conclusion	Support	Confidence
47	Pressione acqua raffreddamento_range5 [2691.399 - ∞]	Temperatura PT100/2 riduttore 3_range2 [30.015 - 109.996]	0.312	0.823
48	Temperatura PT100/2 riduttore 3_range2 [30.015 - 109.996]	Pressione acqua raffreddamento_range5 [2691.399 - ∞]	0.312	0.909
49	Pressione acqua raffreddamento_range5 [2691.399 - ∞]	Temperatura PT100/1 riduttore 4_range2 [30.015 - 109.996]	0.311	0.820
50	Temperatura PT100/1 riduttore 4_range2 [30.015 - 109.996]	Pressione acqua raffreddamento_range5 [2691.399 - ∞]	0.311	0.911
51	Pressione acqua raffreddamento_range5 [2691.399 - ∞]	Temperatura PT100/2 riduttore 4_range2 [30.015 - 109.996]	0.310	0.819
52	Temperatura PT100/2 riduttore 4_range2 [30.015 - 109.996]	Pressione acqua raffreddamento_range5 [2691.399 - ∞]	0.310	0.910
53	Pressione acqua raffreddamento_range5 [2691.399 - ∞]	ConfigLoco_11	0.308	0.814
54	ConfigLoco_11	Pressione acqua raffreddamento_range5 [2691.399 - ∞]	0.308	0.924
55	Tensione di batteria attuale_range5 [29.083 - ∞]	Temperatura PT100/1 riduttore 3_range2 [30.015 - 109.996]	0.303	0.833
56	Temperatura PT100/1 riduttore 3_range2 [30.015 - 109.996]	Tensione di batteria attuale_range5 [29.083 - ∞]	0.303	0.882
57	Tensione di batteria attuale_range5 [29.083 - ∞]	Temperatura PT100/2 riduttore 3_range2 [30.015 - 109.996]	0.302	0.830
58	Temperatura PT100/2 riduttore 3_range2 [30.015 - 109.996]	Tensione di batteria attuale_range5 [29.083 - ∞]	0.302	0.882
59	Tensione di batteria attuale_range5 [29.083 - ∞]	Temperatura PT100/1 riduttore 4_range2 [30.015 - 109.996]	0.301	0.827
60	Temperatura PT100/1 riduttore 4_range2 [30.015 - 109.996]	Tensione di batteria attuale_range5 [29.083 - ∞]	0.301	0.884
61	Pressione acqua raffreddamento_range5 [2691.399 - ∞]	Sforzo richiesto verso DCU_range2 [-28.400 - 28.200]	0.301	0.795
62	Sforzo richiesto verso DCU_range2 [-28.400 - 28.200]	Pressione acqua raffreddamento_range5 [2691.399 - ∞]	0.301	0.881
63	Tensione di batteria attuale_range5 [29.083 - ∞]	Temperatura PT100/2 riduttore 4_range2 [30.015 - 109.996]	0.301	0.826
64	Temperatura PT100/2 riduttore 4_range2 [30.015 - 109.996]	Tensione di batteria attuale_range5 [29.083 - ∞]	0.301	0.883
65	Duration (seconds)_range1 [-∞ - 136416.800]	Temperatura PT100/1 riduttore 3_range2 [30.015 - 109.996]	0.301	0.368
66	Temperatura PT100/1 riduttore 3_range2 [30.015 - 109.996]	Duration (seconds)_range1 [-∞ - 136416.800]	0.301	0.874
67	Duration (seconds)_range1 [-∞ - 136416.800]	Temperatura PT100/2 riduttore 3_range2 [30.015 - 109.996]	0.300	0.367
68	Temperatura PT100/2 riduttore 3_range2 [30.015 - 109.996]	Duration (seconds)_range1 [-∞ - 136416.800]	0.300	0.875
69	Pressione acqua raffreddamento_range5 [2691.399 - ∞]	Sforzo realizzato da DCU_range2 [-30.400 - 25.200]	0.300	0.791
70	Sforzo realizzato da DCU_range2 [-30.400 - 25.200]	Pressione acqua raffreddamento_range5 [2691.399 - ∞]	0.300	0.880
71	Duration (seconds)_range1 [-∞ - 136416.800]	Temperatura PT100/1 riduttore 4_range2 [30.015 - 109.996]	0.299	0.365
72	Temperatura PT100/1 riduttore 4_range2 [30.015 - 109.996]	Duration (seconds)_range1 [-∞ - 136416.800]	0.299	0.876
73	Duration (seconds)_range1 [-∞ - 136416.800]	Temperatura PT100/2 riduttore 4_range2 [30.015 - 109.996]	0.299	0.365
74	Temperatura PT100/2 riduttore 4_range2 [30.015 - 109.996]	Duration (seconds)_range1 [-∞ - 136416.800]	0.299	0.876
75	Tensione di batteria attuale_range5 [29.083 - ∞]	ConfigLoco_11	0.296	0.812
76	ConfigLoco_11	Tensione di batteria attuale_range5 [29.083 - ∞]	0.296	0.886
77	Duration (seconds)_range1 [-∞ - 136416.800]	Tensione di linea istantanea_range5 [3401.163 - ∞]	0.296	0.362
78	Tensione di linea istantanea_range5 [3401.163 - ∞]	Duration (seconds)_range1 [-∞ - 136416.800]	0.296	0.849
79	Priority_C	Pressione acqua raffreddamento_range5 [2691.399 - ∞]	0.290	0.746
80	Pressione acqua raffreddamento_range5 [2691.399 - ∞]	Priority_C	0.290	0.766

ID	Premises	Conclusion	Support	Confidence
81	Numero Treno_0,00	Pressione in Condotta Generale_range1 [-∞ - 1500]	0.290	0.534
82	Pressione in Condotta Generale_range1 [-∞ - 1500]	Numero Treno_0,00	0.290	0.915
83	Duration (seconds)_range1 [-∞ - 136416.800]	ConfigLoco_11	0.287	0.351
84	ConfigLoco_11	Duration (seconds)_range1 [-∞ - 136416.800]	0.287	0.861
85	Tensione di linea istantanea_range5 [3401.163 - ∞]	ConfigLoco_11	0.286	0.822
86	ConfigLoco_11	Tensione di linea istantanea_range5 [3401.163 - ∞]	0.286	0.858
87	Pressione acqua raffreddamento_range5 [2691.399 - ∞]	Effort_range2 [-30.400 - 25.200]	0.286	0.755
88	Effort_range2 [-30.400 - 25.200]	Pressione acqua raffreddamento_range5 [2691.399 - ∞]	0.286	0.872
89	Priority_C	Tensione di batteria attuale_range5 [29.083 - ∞]	0.285	0.734
90	Tensione di batteria attuale_range5 [29.083 - ∞]	Priority_C	0.285	0.783
91	Tensione di batteria attuale_range5 [29.083 - ∞]	Sforzo richiesto verso DCU_range2 [-28.400 - 28.200]	0.285	0.782
92	Sforzo richiesto verso DCU_range2 [-28.400 - 28.200]	Tensione di batteria attuale_range5 [29.083 - ∞]	0.285	0.833
93	Tensione di linea istantanea_range5 [3401.163 - ∞]	Temperatura PT100/1 riduttore 3_range2 [30.015 - 109.996]	0.284	0.814
94	Temperatura PT100/1 riduttore 3_range2 [30.015 - 109.996]	Tensione di linea istantanea_range5 [3401.163 - ∞]	0.284	0.824
95	Tensione di linea istantanea_range5 [3401.163 - ∞]	Sforzo richiesto verso DCU_range2 [-28.400 - 28.200]	0.283	0.813
96	Sforzo richiesto verso DCU_range2 [-28.400 - 28.200]	Tensione di linea istantanea_range5 [3401.163 - ∞]	0.283	0.829
97	Tensione di batteria attuale_range5 [29.083 - ∞]	Sforzo realizzato da DCU_range2 [-30.400 - 25.200]	0.283	0.778
98	Sforzo realizzato da DCU_range2 [-30.400 - 25.200]	Tensione di batteria attuale_range5 [29.083 - ∞]	0.283	0.832
99	Tensione di linea istantanea_range5 [3401.163 - ∞]	Temperatura PT100/2 riduttore 3_range2 [30.015 - 109.996]	0.283	0.813
100	Temperatura PT100/2 riduttore 3_range2 [30.015 - 109.996]	Tensione di linea istantanea_range5 [3401.163 - ∞]	0.283	0.826

6.2 Analisi dei Risultati Ottenuti

Sebbene la ricerca delle regole di associazione sia facilitata dall'utilizzo di software che implementano algoritmi ad hoc, non è sempre immediato ottenere delle correlazioni che possano avere un concreto riscontro pratico. Infatti, nonostante molte delle regole individuate siano sostenute da un valore di supporto e di confidenza medio-alto, che ne avvalora la bontà e l'attendibilità, analizzandole più in dettaglio si osserva che, seppur corrette, queste sembrano descrivere semplicemente le condizioni di funzionamento nominale della macchina, senza evidenziare anomalie e/o possibili guasti.

Questo è un risultato più che prevedibile, dal momento che un DDS è costituito da molte variabili ambientali di cui solo alcune, essendo alterate, sono legate all'anomalia registrata. Le altre, ricadendo nei range di valori ottimali di funzionamento, sono in numero nettamente maggiore e contribuiscono in modo rilevante alla generazione di correlazioni che, tuttavia, non forniscono regole utili alla diagnostica.

In Tabella 6-2 sono riportate alcune delle regole che rientrano in questa casistica.

Tabella 6-2: alcune regole interessanti dai run 1 e 2

Regola N.	Premises	Conclusion	Support	Confidence
1	Tensione di batteria attuale_range5 [29.083 - ∞]	Pressione acqua raffreddamento_range5 [2691.399 - ∞]	0.361	0.992
2	Pressione in Condotta Generale_range1 [-∞ - 1500]	Speed_range1 [-∞ - 32.800]	0.313	0.988
3	Temperatura PT100/1 riduttore 4_range2 [30.015 - 109.996]	Temperatura PT100/1 riduttore 3_range2 [30.015 - 109.996], Temperatura PT100/2 riduttore 3_range2 [30.015 - 109.996]	0.338	0.990
4	Sforzo richiesto verso DCU_range2 [-28.400 - 28.200]	Sforzo realizzato da DCU_range2 [-30.400 - 25.200]	0.335	0.980
5	Tensione di linea istantanea_range5 [3401.163 - ∞]	Tensione di batteria attuale_range5 [29.083 - ∞]	0.329	0.945
6	Speed_range1 [-∞ - 32.800]	Pressione in Condotta Generale_range1 [-∞ - 1500]	0.313	0.468

Regola 1: Tensione di batteria attuale_range5 [29.083 - ∞] → Pressione acqua raffreddamento_range5 [2691.399 - ∞].

L'associazione presenta un supporto del 36.1% e una confidenza pari al 99.2% ma, nonostante questi valori indichino che la regola è stata trovata vera molte volte, sembra risultare poco utile in quanto associa due variabili della macchina che sono indipendenti tra loro.

Regola 2: Pressione in Condotta Generale_range1 [-∞ - 1500] → Speed_range1 [-∞ - 32.800].

L'associazione presenta un supporto del 31.3% e una confidenza del 98.8%. In questo caso la correlazione tra gli elementi è presente, ma sta a dimostrare che quando la pressione in condotta generale è bassa (ossia la macchina è in frenata o ferma) allora la velocità è altrettanto bassa.

Infatti, il sistema frenante della E464 è costituito da una condotta pneumatica (detta "condotta generale") che attraversa tutto il treno e che, quando il treno è sfrenato, presenta una pressione di circa 5 bar. Riducendo la pressione, in maniera più o meno accentuata, si ottiene la frenatura graduale del treno. Per sfrenare i veicoli occorre riportare la pressione in condotta al suo valore nominale.

Regola 3: Temperatura PT100/1 riduttore 4_range2 [30.015 - 109.996] → Temperatura PT100/1 riduttore 3_range2 [30.015 - 109.996], Temperatura PT100/2 riduttore 3_range2 [30.015 - 109.996].

L'associazione presenta un supporto del 33.8% e una confidenza del 99.0%. Anche in questo caso gli elementi sono dipendenti: di norma si trovano alla stessa temperatura perché i riduttori lavorano contemporaneamente e allo stesso modo durante la marcia. In altri termini, la regola rappresenta una condizione ottimale; sarebbe stata interessante se avesse associato temperature in range differenti, perché avrebbe segnalato un eventuale problema nei riduttori del carrello anteriore/posteriore.

Regola 4: Sforzo richiesto verso DCU_range2 [-28.400 - 28.200] → Sforzo realizzato da DCU_range2 [-30.400 - 25.200]

L'associazione presenta un supporto del 33.5% e una confidenza del 98%. Gli elementi sono dipendenti, in quanto uno rappresenta lo sforzo richiesto verso la Drive Control Unit e l'altro quello effettivamente realizzato. Anche in questo caso è descritto un comportamento ordinario della macchina, in quanto lo sforzo richiesto è pari a quello realizzato.

Regola 5: Tensione di linea istantanea_range5 [3401.163 - ∞] → Tensione di batteria attuale_range5 [29.083 - ∞]

L'associazione presenta un supporto del 32.9% e una confidenza del 94.5%. Una connessione tra gli elementi è presente, anche se non sono strettamente dipendenti. La carica della batteria avviene, infatti, solo quando è applicata tensione alla locomotiva, ma questa potrebbe risultare carica anche quando il pantografo è abbassato. Si presume, quindi, che durante la macia, a meno di malfunzionamenti, la tensione della batteria sia sempre nel suo range massimo/ottimale. Il comportamento è anche qui nominale.

Oltre all'analisi critica e diretta, eseguita studiando quelle associazioni con i valori di supporto e confidenza più alti, è stata percorsa una strada alternativa per cercare di distinguere le regole che rilevano il normale funzionamento da quelle che evidenziano potenziali anomalie.

Dall'osservazione delle statistiche relative alle singole variabili, si osserva che la distribuzione dei valori all'interno dei cinque range, stabiliti in precedenza e di egual dimensione, non è equa: ci sono intervalli che ricorrono più frequentemente rispetto ad altri. Questo potrebbe far presupporre che tali intervalli siano ordinari per il funzionamento della locomotiva e, di conseguenza, che quelli meno frequenti siano caratteristici di un eventuale anomalia.

Come punto di partenza per questa analisi si è scelto quindi di considerare quelle regole che contengono variabili in intervalli meno ricorrenti.

A seguire un esempio numerico di quanto descritto: in Figura 6-1 è riportata la statistica dei valori assunti dalla variabile *Temperatura motore 4 sonda 2*.

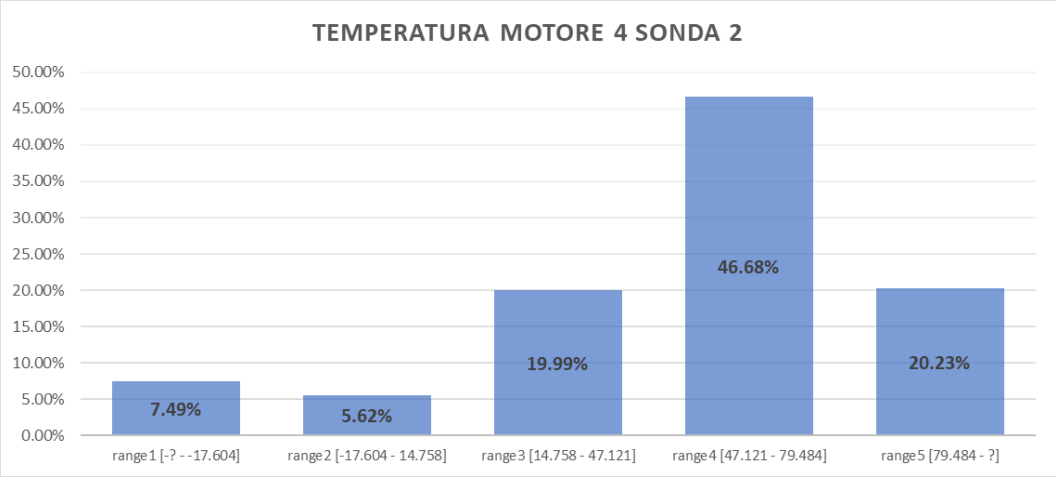

Figura 6-1: distribuzione delle Temperature Motore 4 Sonda 2 sui vari range

Gli intervalli ad essa associati sono cinque, di egual estensione, così ripartiti:

Tabella 6-3: range per la variabile Temperatura Motore 4 Sonda 2

NOME	INTERVALLO	FREQUENZA
range1	– ∞ ÷ -17.604 °C	7.49%
range2	-17.604 °C ÷ 14.758 °C	5.62%
range3	14.758 °C ÷ 47.121 °C	19.99%
range4	47.121 °C ÷ 79.484 °C	46.68%
range5	79.484 °C ÷ + ∞	20.23%

Si osserva che ai range 3 e 4 corrispondono valori di frequenza maggiori: questo fa presupporre che le temperature ad essi associate siano descrittrici del normale funzionamento della macchina. Di conseguenza, si può assumere che gli intervalli meno frequenti rappresentino condizioni atipiche, con l'eccezione, per questa specifica variabile, dei range 1 e 2, in quanto, temperature più basse potrebbero corrispondere alle fasi iniziali di funzionamento del motore.

Applicando quindi un filtro, che escluda quei range poco interessanti, si individuano le seguenti associazioni per la variabile di cui all'esempio:

Tabella 6-4: regole per Temperatura motore 4 sonda 2 - range5

Premises	Conclusion	Support	Confidence
Temperatura motore 4 sonda 2_range5 [79.484 - ∞]	Riduzione % del Traction Contr_range1 [-∞ - 19.997]	0.031	0.988
Temperatura motore 4 sonda 2_range5 [79.484 - ∞]	Priority_B	0.031	0.982
Temperatura motore 4 sonda 2_range5 [79.484 - ∞]	Pressione H2O_range5 [2701.165 - ∞]	0.031	0.973
Temperatura motore 4 sonda 2_range5 [79.484 - ∞]	Tensione semifiltro inf. INV2_range5 [1568.404 - ∞]	0.029	0.922
Temperatura motore 4 sonda 2_range5 [79.484 - ∞]	Tensione di DC-Link_range5 [3072.893 - ∞]	0.029	0.919
Temperatura motore 4 sonda 2_range5 [79.484 - ∞]	Duration (seconds)_range1 [-∞ - 136416.800]	0.029	0.913
Temperatura motore 4 sonda 2_range5 [79.484 - ∞]	Tensione di linea_range5 [3123.478 - ∞]	0.028	0.904
Temperatura motore 4 sonda 2_range5 [79.484 - ∞]	Tensione semifiltro sup. INV1_range5 [1569.674 - ∞]	0.028	0.904
Temperatura motore 4 sonda 2_range5 [79.484 - ∞]	Potenza reostato di frenatura_range3 [-1108.542 - 8.447]	0.028	0.901
Temperatura motore 4 sonda 2_range5 [79.484 - ∞]	Temperatura motore 2 sonda 1_range5 [82.647 - ∞]	0.028	0.887
Temperatura motore 4 sonda 2_range5 [79.484 - ∞]	Vehicle_152	0.026	0.833
Temperatura motore 4 sonda 2_range5 [79.484 - ∞]	Description_Sonda temp. T2.4 non affidabile	0.025	0.806
Temperatura motore 4 sonda 2_range5 [79.484 - ∞]	Numero Treno_0,00	0.019	0.591
Temperatura motore 4 sonda 2_range5 [79.484 - ∞]	Conducibilita' H2O_range3 [1.635 - 2.452]	0.019	0.591
Temperatura motore 4 sonda 2_range5 [79.484 - ∞]	Sforzo traz/fren richiesto_range2 [-29.014 - 26.944]	0.014	0.460
Temperatura motore 4 sonda 2_range5 [79.484 - ∞]	Sforzo traz/fren sviluppato_range2 [-30.971 - 24.615]	0.014	0.448
Temperatura motore 4 sonda 2_range5 [79.484 - ∞]	Temperatura motore 3 sonda 1_range5 [98.987 - ∞]	0.013	0.427
Temperatura motore 4 sonda 2_range5 [79.484 - ∞]	End Time_range3 [5.400 - 7.600]	0.013	0.424
Temperatura motore 4 sonda 2_range5 [79.484 - ∞]	Temperatura motore 3 sonda 2_range5 [97.952 - ∞]	0.013	0.409
Temperatura motore 4 sonda 2_range5 [79.484 - ∞]	Potenza di linea_range3 [0.286 - 1.670]	0.013	0.403
Temperatura motore 4 sonda 2_range5 [79.484 - ∞]	Corrente di linea_range3 [78.418 - 489.208]	0.013	0.400
Temperatura motore 4 sonda 2_range5 [79.484 - ∞]	Potenza di linea_range2 [-1.099 - 0.286]	0.012	0.373
Temperatura motore 4 sonda 2_range5 [79.484 - ∞]	Corrente di linea_range2 [-332.372 - 78.418]	0.012	0.370
Temperatura motore 4 sonda 2_range5 [79.484 - ∞]	Start Time_7	0.011	0.334
Temperatura motore 4 sonda 2_range5 [79.484 - ∞]	End Time_range4 [7.600 - 9.800]	0.010	0.322
Temperatura motore 4 sonda 2_range5 [79.484 - ∞]	Temperatura motore 1 sonda 1_range5 [97.952 - ∞]	0.009	0.299
Temperatura motore 4 sonda 2_range5 [79.484 - ∞]	Temperatura motore 1 sonda 2_range5 [97.952 - ∞]	0.009	0.293
Temperatura motore 4 sonda 2_range5 [79.484 - ∞]	Temperatura H2O_range5 [43.230 - ∞]	0.009	0.290
Temperatura motore 4 sonda 2_range5 [79.484 - ∞]	Media velocita' asse acquisita_range2 [-135.720 - -67.030]	0.009	0.287
Temperatura motore 4 sonda 2_range5 [79.484 - ∞]	Velocita' di riferimento del treno_range4 [96.312 - 128.416]	0.008	0.245

Il primo step per l'analisi è quello di cercare se tra i conseguenti c'è una *description* associata all'antecedente *"Temperatura motore 4 sonda 2_range5"*. Si procede focalizzandosi su questa voce perché è quella che potrebbe contenere informazioni più esplicite sull'eventuale anomalia. In questo caso, la description individuata *"Sonda temp. T2.4 non affidabile"* lascerebbe presupporre che la lettura data dalla sonda sia non corretta.

In realtà, dalle altre associazioni evidenziate in Tabella 6-4 si evince che anche le altre sonde rilevano una temperatura elevata (range5). Segue che o tutte le sonde stanno leggendo una temperatura falsata oppure la *description* non risulta attendibile. Ad avvalorare questa ipotesi si ha che contestualmente anche la temperatura dell'acqua di raffreddamento *"Temperatura H2O"*, e la relativa pressione *"Pressione H2O"*, sono di valore elevato. A sostegno della lettura corretta da parte della sonda, abbiamo inoltre che la marcia del treno è a velocità relativamente sostenuta (*Velocità di riferimento del treno_range4 [96.3km/h ÷ 128.4 km/h]*); a velocità elevate, infatti, si sviluppa più calore. Un'ulteriore associazione a riprova della tesi è *"Temperatura motore 4 sonda 2_range5"* con "*Start Time_7*",

dove sette indica il mese di luglio, caratterizzato da temperature piuttosto elevate.

Con le osservazioni precedenti, quindi, si può escludere quasi con certezza un'avaria della sonda, resta però il fatto che temperature così alte sono presenti e possano essere dovute ad altre possibili cause, non individuate dal processo di mining (es. livello basso dell'acqua di raffreddamento).

La stessa logica è stata applicata di seguito: dall'osservazione delle statistiche sono stati scelti gli intervalli meno ricorrenti per le variabili ritenute più significative. I range considerati si riportano nella Tabella 6-5.

Tabella 6-5: range considerati per ogni variabile

ID	VARIABILE	RANGE CONSIDERATI
1	Tensione di linea	range1, range4
2	Tensione di linea istantanea	range1, range4
3	Tensione batteria attuale	range3, range4
4	Tensione semifiltro sup. INV1	range4, range1
	Tensione semifiltro inf. INV2	
5	Riduzione % del traction control	range2, range3, range4, range5
6	Temperatura motore 1	range5
	Temperatura motore 2	
	Temperatura motore 3	
	Temperatura motore 4	
7	Temperatura motore 1 sonda 1	range5
	Temperatura motore 1 sonda 2	
	Temperatura motore 2 sonda 1	
	Temperatura motore 2 sonda 2	
	Temperatura motore 3 sonda 1	
	Temperatura motore 3 sonda 2	
	Temperatura motore 4 sonda 1	
	Temperatura motore 4 sonda 2	
8	Temperatura riduttore 1	range5
	Temperatura riduttore 2	
	Temperatura riduttore 3	
	Temperatura riduttore 4	
9	Temperatura PT100/1 riduttore 1	range5
	Temperatura PT100/2 riduttore 1	
	Temperatura PT100/1 riduttore 2	
	Temperatura PT100/2 riduttore 2	
	Temperatura PT100/1 riduttore 3	
	Temperatura PT100/2 riduttore 3	
	Temperatura PT100/1 riduttore 4	
	Temperatura PT100/2 riduttore 4	
10	Temperatura esterna climatizzatore	range1, range4, range5
11	Temperatura H2O	range5

ID	VARIABILE	RANGE CONSIDERATI
	Temperatura acqua di raffreddamento	
12	Pressione H2O	range1
	Pressione acqua di raffreddamento	

Non in tutti i casi sono emerse delle associazioni contenenti le variabili nei range selezionati. Per i diversi gruppi si riporta una breve discussione delle regole individuate.

Gruppo 1: ID1 – Tensione di linea / ID2 – Tensione di linea istantanea

Si è scelto di raggruppare questi due parametri perché rappresentano la stessa grandezza. La differenza sta nel processore che riceve in input queste misure e che genera in output il DDS denominandole in maniera diversa: le voci non compaiono mai contemporaneamente nel singolo record (quando c'è l'una non c'è l'altra).

Gli istogrammi in Figura 6-2 e in Figura 6-3 sono concordi nel mostrare che i valori più comuni ricadono nel range4 e nel range5 che, di conseguenza, sono stati considerati come intervalli di normale funzionamento della macchina.

Per la ricerca delle associazioni si considera il range1 perché è lontano dall'intervallo di funzionamento ottimale e contiene un numero abbastanza grande di valori.

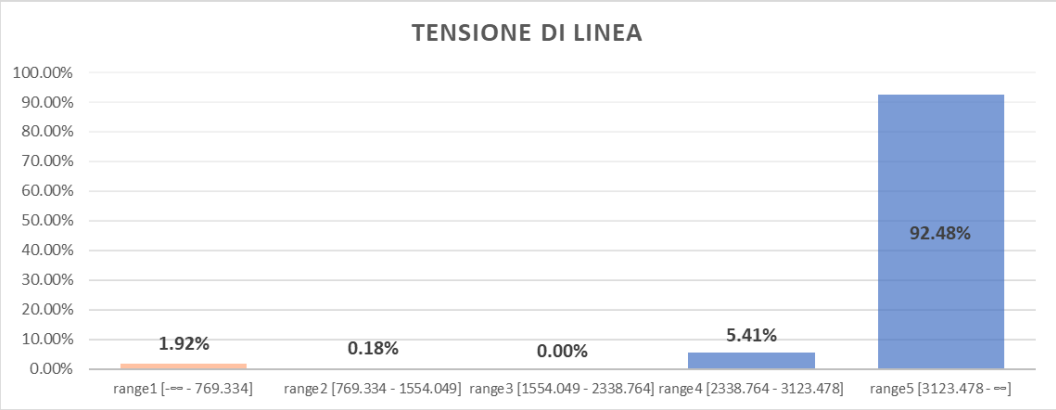

Figura 6-2: istogramma della Tensione di Linea

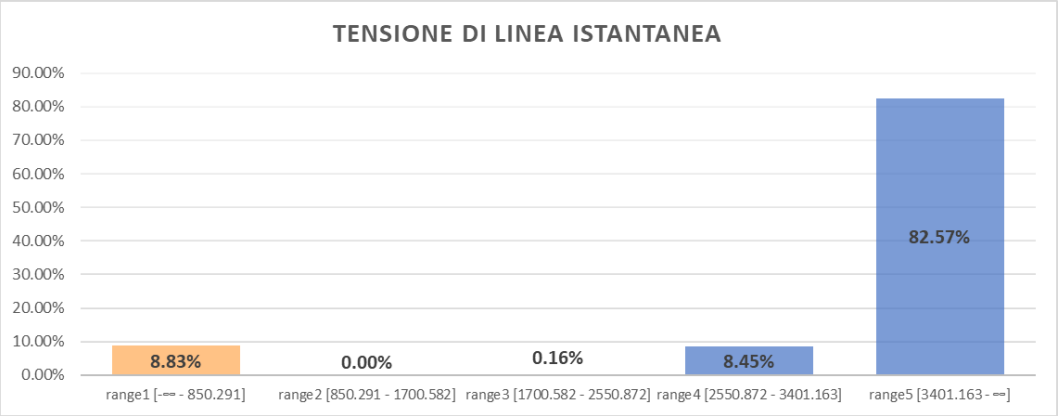

Figura 6-3: istogramma della Tensione di Linea Istantanea

Le associazioni più rilevanti per il Gruppo 1 sono riportate in Tabella 6-6,

Tabella 6-6: regole significative estratte dal Gruppo 1

Regola N.	Premises	Conclusion	Support	Confidence
1	Tensione di linea istantanea_range1 [-∞ - 850.291]	Pressione in Condotta Generale_range1 [-∞ - 1500]	3.08%	82.62%
2	Tensione di linea istantanea_range1 [-∞ - 850.291]	Tensione di batteria attuale_range4 [25.985 - 29.083]	0.73%	19.65%
3	Tensione di linea_range1 [-∞ - 769.334]	Speed_range1 [-∞ - 32.800]	0.28%	93.75%
4	Tensione di linea_range1 [-∞ - 769.334]	Description_Semifiltro INV1 non caricato	0.10%	34.38%

dalle quali si evince che, quando la tensione di linea assume i valori compresi nel range1, i conseguenti sembrano descrivere uno stato del treno non in marcia, cioè fermo e/o in deposito, con il pantografo abbassato. Le conclusion *"Pressione in Condotta Generale_range1 [-∞ - 1500]"* e *"Speed_range1 [-∞ - 32.800]"* possono indicare, infatti, che la motrice è in stazionamento, in quanto la pressione in condotta generale e la velocità sono nell'intervallo più basso. Inoltre, lo stato *"Tensione di batteria attuale_range4 [25.985 - 29.083]"* potrebbe segnalare che la batteria non è in carica: ciò conferma la condizione di pantografo abbassato.

In ultima analisi la regola *"Tensione di linea_range1 [-∞ - 769.334]"* → *"Description_Semifiltro INV1 non caricato"*, necessiterebbe dell'interpretazione di un tecnico in grado attribuirle un significato adeguato.

Gruppo 2: ID3 – Tensione batteria attuale

L'istogramma in Figura 6-4 mostra che i valori del range5 sono i più frequenti e quindi teoricamente relativi al normale funzionamento della motrice.

Per la ricerca delle associazioni si considerano range1, range2, range3 e range4 perché poco comuni. Da notare che il range2 non contiene valori e di conseguenza non ha generato regole.

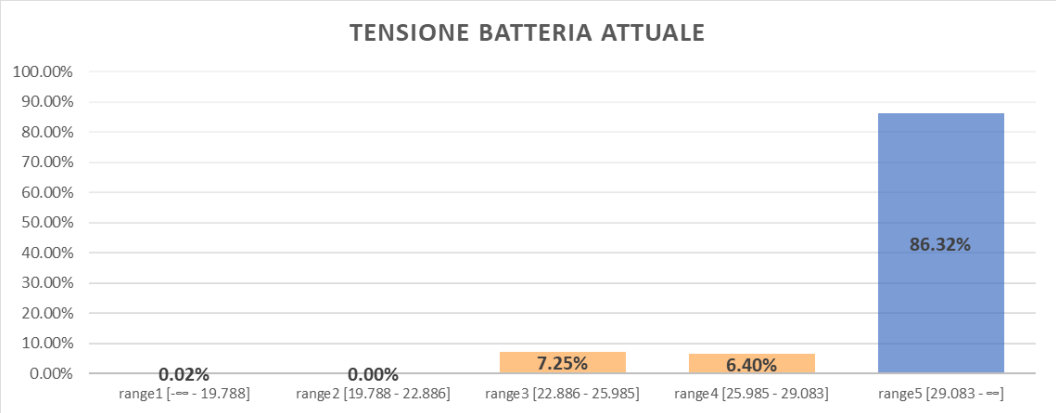

Figura 6-4: istogramma della Tensione Batteria Attuale

Le associazioni più rilevanti per il Gruppo 2 sono riportate in Tabella 6-7.

Tabella 6-7: regole significative estratte dal Gruppo 2

Regola N.	Premises	Conclusion	Support	Confidence
1	Tensione di batteria attuale_range3 [22.886 - 25.985]	Speed_range1 [-∞ - 32.800]	3.02%	98.77%
2	Tensione di batteria attuale_range3 [22.886 - 25.985]	Pressione acqua raffreddamento_range1 [-∞ - 672.850]	3.02%	98.77%
3	Tensione di batteria attuale_range3 [22.886 - 25.985]	Tensione di linea istantanea_range1 [-∞ - 850.291]	2.90%	94.79%
4	Tensione di batteria attuale_range3 [22.886 - 25.985]	Numero Treno_0,00	2.65%	86.81%
5	Tensione di batteria attuale_range3 [22.886 - 25.985]	Pressione in Condotta Generale_range1 [-∞ - 1500]	2.51%	82.21%
6	Tensione di batteria attuale_range1 [-∞ - 19.788]	Numero Treno_0,00	0.01%	100.00%
7	Tensione di batteria attuale_range1 [-∞ - 19.788]	Pressione in Condotta Generale_range1 [-∞ - 1500]	0.01%	100.00%
8	Tensione di batteria attuale_range1 [-∞ - 19.788]	Temperatura riduttore 1_range1 [-∞ - 27.002]	0.01%	100.00%
9	Tensione di batteria attuale_range1 [-∞ - 19.788]	Temperatura PT100/2 riduttore 4_range1 [-∞ - 30.015]	0.01%	100.00%
10	Tensione di batteria attuale_range1 [-∞ - 19.788]	Temperatura PT100/1 riduttore 3_range1 [-∞ - 30.015]	0.01%	100.00%

Regola N.	Premises	Conclusion	Support	Confidence
11	Tensione di batteria attuale_range1 [-∞ - 19.788]	Temperatura PT100/1 riduttore 2_range1 [-∞ - 25.943]	0.01%	100.00%
12	Tensione di batteria attuale_range1 [-∞ - 19.788]	Temperatura PT100/1 riduttore 1_range1 [-∞ - 25.411]	0.01%	100.00%

I risultati sono affini a quelli derivanti dal Gruppo 1 dal quale si evince che il treno è fermo e/o in deposito. Questo è confermato dal fatto che tutti i conseguenti rappresentano valori tipici di una motrice non in marcia:

- *Speed_range1 [-∞ - 32.800]*: velocità bassa o nulla;
- *Pressione in Condotta Generale_range1 [-∞ - 1500]*: condotta generale scarica, conseguente sistema frenante azionato;
- *Tensione di linea istantanea_range1 [-∞ - 850.291]*: tensione di linea bassa o nulla, implica pantografo potenzialmente abbassato;
- *Pressione acqua raffreddamento_range1 [-∞ - 672.850]*: bassa pressione nel circuito di raffreddamento, segue che la pompa non è azionata;
- *Numero Treno_0,00*: treno senza identificativo, quindi non in macia;
- *Temperatura PT100/N riduttore M_range1 [-∞ - 30.015]*: bassa temperatura del riduttore sottintende assenza di funzionamento a riprova che il treno fermo.

Gruppo 3: ID4 Tensione semifiltro sup. INV1 / Tensione semifiltro inf. INV2

Si è scelto di raggruppare questi due parametri perché fanno parte della stessa famiglia, i due filtri, infatti, hanno entrambi la funzione di livellare la tensione in ingresso agli inverter.

Gli istogrammi in Figura 6-5 e in Figura 6-6 mostrano entrambi che il range5 è l'insieme di valori più comuni e, per questo motivo, è stato considerato come intervallo di lavoro ottimale. Per questo, si ricercano le associazioni nell'intervallo range1 ÷ range4 poiché rappresentano valori non comuni.

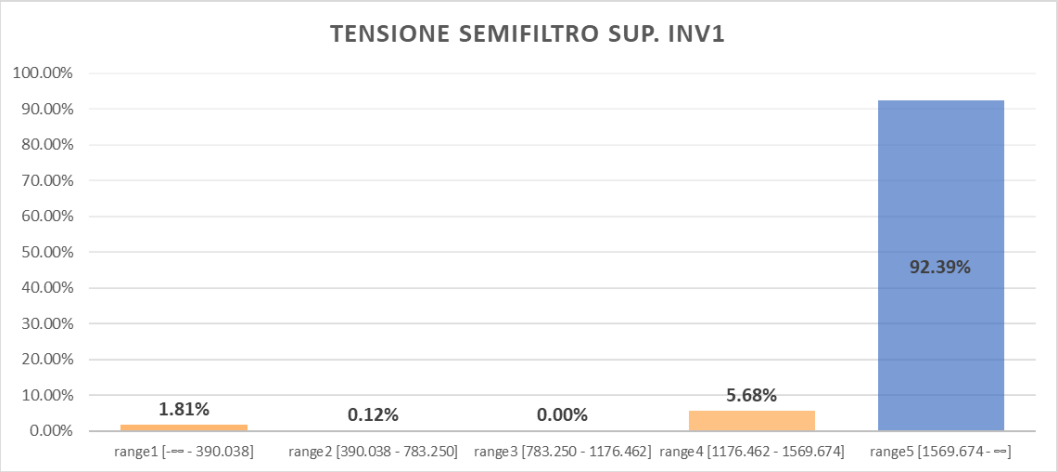

Figura 6-5: istogramma della Tensione Semifiltro Sup. INV1

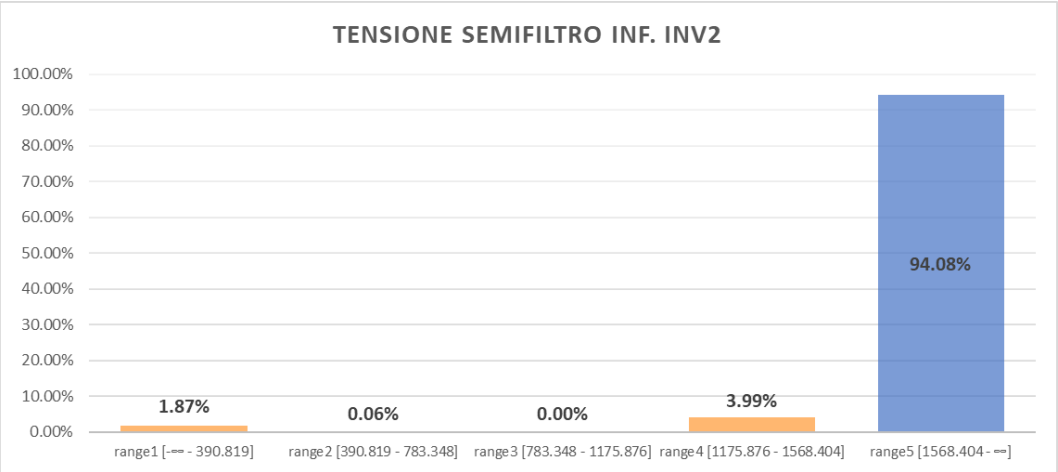

Figura 6-6: istogramma della Tensione Semifiltro Inf. INV2

Le associazioni più rilevanti per il Gruppo 3 sono riportate in Tabella 6-7.

Tabella 6-8: regole significative estratte dal Gruppo 3

Regola N.	Premises	Conclusion	Support	Confidence
1	Tensione semifiltro inf. INV2_range1 [-∞ - 390.819]	Velocita' di riferimento del treno_range1 [-∞ - 32.104]	0.28%	96.77%
2	Tensione semifiltro inf. INV2_range1 [-∞ - 390.819]	Tensione semifiltro sup. INV1_range1 [-∞ - 390.038]	0.28%	96.77%
3	Tensione semifiltro sup. INV1_range1 [-∞ - 390.038]	Tensione semifiltro inf. INV2_range1 [-∞ - 390.819]	0.28%	100.00%
4	Tensione semifiltro inf. INV2_range1 [-∞ - 390.819]	Speed_range1 [-∞ - 32.800]	0.27%	93.55%
5	Tensione semifiltro inf. INV2_range1 [-∞ - 390.819]	Tensione di DC-Link_range1 [-∞ - 766.941]	0.27%	93.55%
6	Tensione semifiltro sup. INV1_range1 [-∞ - 390.038]	Speed_range1 [-∞ - 32.800]	0.26%	93.33%
7	Tensione semifiltro inf. INV2_range1 [-∞ - 390.819]	Tensione di linea_range1 [-∞ - 769.334]	0.25%	87.10%

Regola N.	Premises	Conclusion	Support	Confidence
8	Tensione semifiltro sup. INV1_range1 [-∞ - 390.038]	Tensione di linea_range1 [-∞ - 769.334]	0.25%	90.00%
9	Tensione semifiltro inf. INV2_range1 [-∞ - 390.819]	Temperatura motore 1 sonda 1_range1 [-∞ - 24.488]	0.12%	41.94%
10	Tensione semifiltro inf. INV2_range1 [-∞ - 390.819]	Temperatura motore 3 sonda 1_range1 [-∞ - 24.747]	0.12%	41.94%
11	Tensione semifiltro inf. INV2_range1 [-∞ - 390.819]	Temperatura H2O_range2 [15.329 - 24.630]	0.10%	35.48%

Come per il Gruppo 2 le associazioni descrivono uno status di treno è fermo e/o in deposito, confermato dal fatto che tutti i conseguenti rappresentano valori tipici di una motrice non in marcia:

- *Speed_range1 [-∞ - 32.800] / Velocita' di riferimento del treno_range1 [-∞ - 32.104]*: velocità bassa o nulla;
- *Tensione di linea_range1 [-∞ - 850.291] / Tensione di DC-Link_range1 [-∞ - 766.941]*: tensione di linea bassa o nulla, implica pantografo potenzialmente abbassato;
- *Temperatura H2O_range2 [15.329 - 24.630]*: bassa temperatura dell'acqua di raffreddamento, potrebbe indicare che il treno è fermo o appena partito;
- *Temperatura motore 1 sonda 1_range1 [-∞ - 24.488] / Temperatura motore 3 sonda 1_range1 [-∞ - 24.747]*: bassa temperatura dei motori a riprova che questi non stanno funzionando.

Gruppo 4: ID5 – Riduzione % del Traction Control

L'istogramma in Figura 6-7 mostra come più frequenti i valori del range1, rappresentanti una bassa o nulla riduzione dello sforzo di trazione richiesto. Il Traction Control è l'elemento che impedisce il pattinamento della motrice in fase di accelerazione e risulta utile principalmente in condizioni critiche, come pioggia o ghiaccio.

In questo caso, si considerano i range2, range3, range4 e range5 perché segnalano un intervento importante del sistema di controllo di trazione.

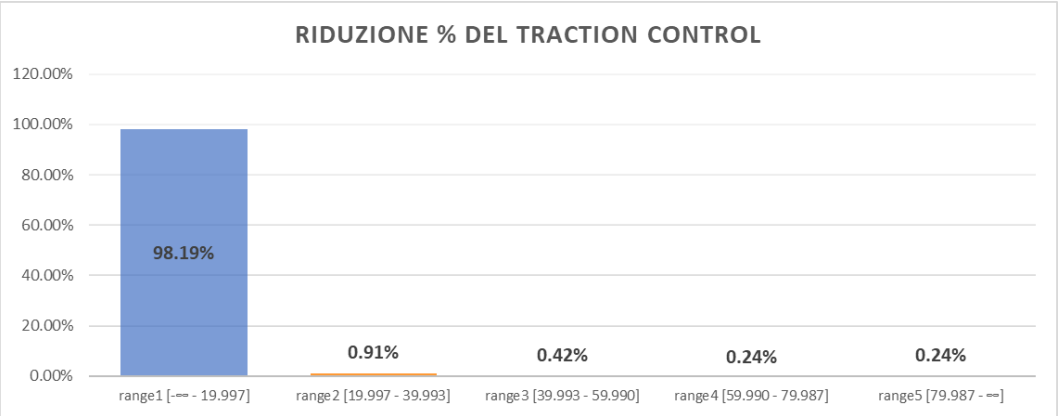

Figura 6-7: istogramma della Riduzione % del Traction Control

Le associazioni più rilevanti per il Gruppo 4 sono riportate in Tabella 6-9.

Tabella 6-9: regole significative estratte dal Gruppo 3

Regola N.	Premises	Conclusion	Support	Confidence
1	Riduzione % del Traction Contr_range2 [19.997 - 39.993]	Start Time_11	0.05%	33.33%
2	Riduzione % del Traction Contr_range2 [19.997 - 39.993]	Start Time_12	0.04%	26.67%
3	Riduzione % del Traction Contr_range2 [19.997 - 39.993]	Temperatura motore 1 sonda 2_range2 [24.488 - 48.976]	0.03%	20.00%
4	Riduzione % del Traction Contr_range2 [19.997 - 39.993]	Temperatura motore 3 sonda 1_range2 [24.747 - 49.494]	0.03%	20.00%
5	Riduzione % del Traction Contr_range2 [19.997 - 39.993]	Temperatura motore 3 sonda 2_range2 [24.488 - 48.976]	0.03%	20.00%
6	Riduzione % del Traction Contr_range2 [19.997 - 39.993]	Temperatura motore 2 sonda 2_range2 [23.429 - 46.857]	0.03%	20.00%
7	Riduzione % del Traction Contr_range2 [19.997 - 39.993]	Temperatura motore 4 sonda 1_range2 [22.628 - 45.256]	0.03%	20.00%
8	Riduzione % del Traction Contr_range2 [19.997 - 39.993]	Numero Treno_23.672,00	0.02%	13.33%

Nonostante le regole individuate abbiano un supporto basso sono supportate da un buon valore di confidenza e sono rappresentative di una situazione più che realistica.

Dalle associazioni, infatti, si evince che il Traction Control è intervenuto con frequenza nei mesi invernali, nei quali è comune che si verifichino situazioni di neve o ghiaccio sulle rotaie. Un'ulteriore conferma è data dalle sonde dei quattro motori, i quali, benché in marcia, registrano valori di temperatura molto bassi.

Inoltre, il numero treno 23'672 è associato alla tratta *Pescara – Sulmona*, linea ferroviaria tra i monti abruzzesi, noti per il loro clima rigido nei mesi invernali.

Gruppo 5: ID6 – Temperatura motore N / ID7 – Temperatura motore N sonda M

Questo gruppo rappresenta i valori di temperatura delle diverse sonde posizionate su ognuno dei quattro motori. Le considerazioni scaturite dall'analisi sono affini a quelle descritte nell'esempio e riportate in Tabella 6-4.

Gruppo 6: ID8 – Temperatura riduttore N / ID9 Temperatura PT100/N riduttore M

Il gruppo 6 è costituito da dodici variabili, utilizzate per il controllo della temperatura dell'olio dei motoriduttori. I due sottogruppi, ID8 e ID9, rappresentano gli stessi valori: la temperatura del riduttore N, infatti, deriva dal primo dei due termistori PT100/N posizionati su ogni riduttore M. Si è scelto di tener conto di entrambi i gruppi in quanto, in modo del tutto casuale, accade spesso che ci siano alternativamente valori mancanti.

Gli istogrammi in Figura 6-8 e in Figura 6-9 sono esplicativi delle variabili di ogni sottogruppo.

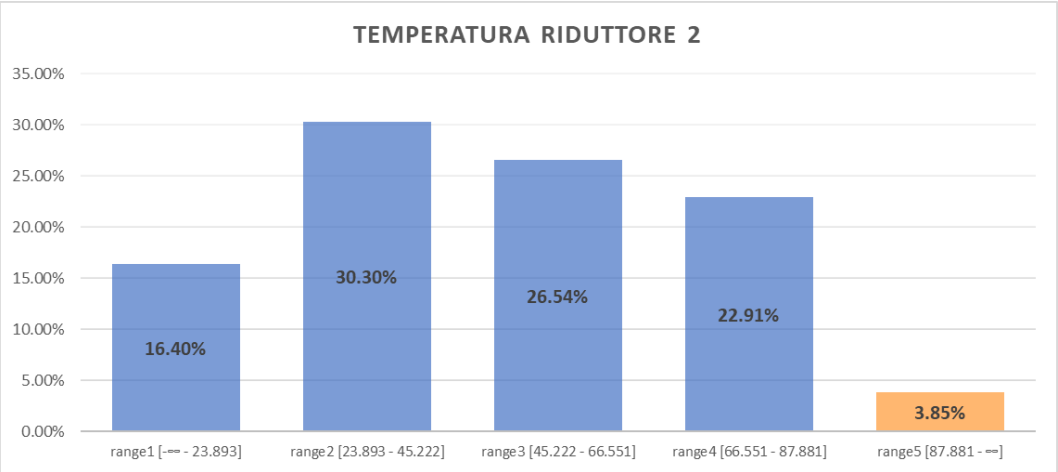

Figura 6-8: istogramma della Temperatura Riduttore 2

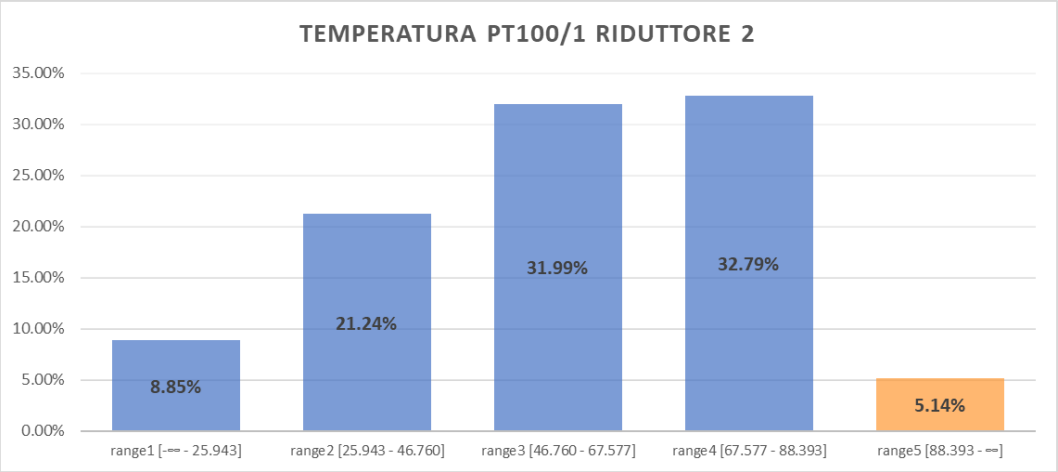

Figura 6-9: istogramma della Temperatura PT100/1 Riduttore 2

Per questo gruppo si ricercano le associazioni costituite dalle variabili in range5 in quanto tra i problemi principali che possono insorgere durante il servizio di un motoriduttore il riscaldamento eccessivo è sicuramente una casistica da attenzionare. I valori nei range3 e range4 sono comuni e possono essere considerati come valori d'esercizio assolutamente nella norma. I valori nel range1 sono meno frequenti ma non da attenzionare in quanto possono essere rappresentativi delle fasi iniziali di funzionamento.

Le associazioni più rilevanti per il Gruppo 6 sono riportate in Tabella 6-10.

Tabella 6-10: regole significative estratte dal Gruppo6

Regola N.	Premises	Conclusion	Support	Confidence
1	Temperatura PT100/2 riduttore 2_range5 [77.375 - ∞]	Pressione Cilindro Freno Posteriore_range1 [-∞ - 896]	5.73%	81.47%
2	Temperatura PT100/2 riduttore 2_range5 [77.375 - ∞]	Pressione Cilindro Freno Anteriore_range1 [-∞ - 895]	5.70%	81.07%
3	Temperatura riduttore 4_range5 [76.281 - ∞]	Description_RSV03: Richiesta taglio trazione da SCMT	5.30%	65.70%
4	Temperatura PT100/2 riduttore 2_range5 [77.375 - ∞]	Description_RSV03: Richiesta taglio trazione da SCMT	4.83%	68.67%
5	Temperatura PT100/2 riduttore 1_range5 [81.612 - ∞]	Pressione in Condotta Generale_range4 [4500 - 6000]	3.18%	80.14%
6	Temperatura PT100/2 riduttore 1_range5 [81.612 - ∞]	Description_RSV03: Richiesta taglio trazione da SCMT	2.45%	61.70%
7	Temperatura PT100/2 riduttore 1_range5 [81.612 - ∞]	Temperatura esterna climatizzatore_range3 [23.971 - 39.349]	2.13%	53.66%

Dalla Tabella 6-10 il conseguente *"Description_RSV03: Richiesta taglio trazione da SCMT"* appare in tre casi con un supporto che va dal 2.5% al 5.3% e una confidenza nel range 61.7% ÷ 65.7%, in associazione ad una temperatura dei riduttori compresa nel range5. Dall'analisi potrebbe sembrare, quindi, che il Sistema di Controllo Marcia Treno intervenga, riducendo la velocità del treno, in concomitanza di un surriscaldamento dei motoriduttori. L'associazione potrebbe però essere di tipo spurio in quanto, dalle informazioni in possesso, il sistema dovrebbe intervenire nel momento in cui l'agente di condotta porta il treno in una condizione di marcia non sicura (non rispetto del segnalamento, superamento della velocità massima ammessa, marcia su binario illegale, ecc…) e non in caso di anomalie legate alle temperature di esercizio. Sarebbe comunque interessante, in ogni caso, sottoporre questa regola al giudizio di personale esperto per valutarne l'eventuale validità.

Dalle altre regole, in ogni caso, si può ipotizzare che il treno sia in movimento (*Pressione Cilindro Freno Posteriore_range1 [-∞ - 896]*, *Pressione Cilindro Freno Anteriore_range1 [-∞ - 895]* e *Pressione in Condotta Generale_range4 [4500 - 6000]*) in un periodo abbastanza caldo (*Temperatura esterna climatizzatore_range3 [23.971 - 39.349]*), il che giustifica le temperature raggiunte dai riduttori.

7 Conclusioni

Il lavoro di tesi ha avuto come obiettivo principale quello di desumere informazioni utili allo sviluppo di nuovi algoritmi diagnostici, per massimizzare efficienza ed efficacia del sistema di Manutenzione Predittiva di Trenitalia.

L'approccio utilizzato è di tipo data-driven e ha visto, come punto di partenza, un'ampia base di dati derivante dal monitoraggio di alcune macchine della flotta di locomotive E464, che rappresentano il fulcro del trasporto regionale giornaliero in Italia.

Il punto di forza del metodo adottato è l'utilizzo del data mining applicato alla ricerca di regole di associazione, al fine di estrarre correlazioni ricorrenti e quanto più affidabili tra gli elementi che compongono i DDS.

La prima problematica riscontrata è stata gestire e rendere omogenei i dati derivanti dai diversi monitoraggi: la mancanza di un separatore univoco per i numeri decimali e il formato delle date non standardizzato ha richiesto una fase preliminare di armonizzazione senza la quale i record sarebbero stati inutilizzabili.

Oltre a ciò, l'assenza di dati riscontrata in molte delle variabili costituenti i singoli DDS, ha visto necessario l'abbassamento della soglia di supporto minimo al fine di estrarre tutte quelle regole che altrimenti sarebbero state perse. Tale impostazione ha però generato una moltitudine di correlazioni che ha reso difficoltoso e lungo il lavoro di analisi dei risultati.

In prima battuta, coerentemente con quanto riportato in letteratura, sono state selezionate e analizzate direttamente le regole con valori di supporto più alti. Successivamente l'indagine si è focalizzata sull'analisi di quelle variabili contenenti valori non ordinari, con l'obiettivo di individuare associazioni più interessanti e potenzialmente legate a guasti o malfunzionamenti.

L'esito, in entrambi i casi, è stato negativo: come si evince dall'analisi dei risultati, molte delle regole sono descrittrici del normale funzionamento della macchina

e non apportano informazioni aggiuntive e utili ai fini della diagnostica predittiva. Tale risultato è più che prevedibile, in quanto un DDS è costituito da molte variabili ambientali di cui solo una piccola parte è effettivamente legata all'anomalia registrata.

In secondo luogo, l'analisi dei risultati e la conseguente ricerca di regole significative solleva una questione basilare: chi porta avanti questa ricerca deve avere una conoscenza approfondita della motrice e dei suoi sottoinsiemi, nonché dell'apparato di diagnostica. Infatti, le informazioni sull'argomento reperibili in letteratura sono nulle in quanto legate sostanzialmente al know-how dell'impresa.

Ragion per cui, l'approccio di ricerca analizzato deve essere considerato come una tecnica esplorativa preliminare che non può, allo stato attuale, sostituirsi al complesso di conoscenze ed esperienze tecniche aziendali. Resta evidente, in ogni caso, il potenziale sull'insieme dei DDS considerato: il mining ha portato a una moltitudine di regole, a cui però è stato difficile assegnare il giusto valore a causa della mancanza di una chiave di lettura adeguata. Uno dei possibili sviluppi futuri potrebbe essere, quindi, quello di analizzare le associazioni individuate con il supporto di personale specializzato, in grado di valutarne, con cognizione di causa, l'utilità nello sviluppo di algoritmi di diagnostica.

Un'ulteriore e significativo contributo per il proseguito del lavoro potrebbe derivare dalla ricerca di pattern sequenziali nell'insieme dei DDS. Come già illustrato in precedenza, questi ultimi sono descrittori di anomalie e un singolo guasto comporta solitamente la generazione di più DDS, che nella maggior parte dei casi si presentano in una sequenza ben definita o nell'arco di intervalli di tempo ravvicinati. Gli algoritmi attualmente utilizzati per la diagnostica si basano su questo principio: individuano determinate catene di DDS a cui, per esperienza, sono stati associati specifici alert o avvisi manutentivi. L'individuazione di questi modelli potrebbe, quindi, contribuire in modo sostanziale alla ricerca di nuovi algoritmi che esulano dall'esperienza del personale.

In conclusione, la realizzazione di una efficace Manutenzione Predittiva è un gioco di squadra in cui l'esperto del sistema ferroviario riveste un ruolo essenziale. È la conoscenza di dominio, infatti, a guidare i data scientist nella costruzione degli algoritmi corretti che verranno poi implementati. Il successo di una soluzione di manutenzione predittiva ferroviaria consiste quindi nella scelta oculata dei sistemi del treno da analizzare, nella costruzione di un opportuno ecosistema di dati e nella giusta combinazione di esperti in campo ferroviario e data scientist.

Bibliografia

Agnoli, Alberto, e Guido Del Gobbo. «Sistema Telediagnostica.» *Manutenzione & Trasporti*, 2016: 15 - 17.

Agrawal, R., T. Imielinski, e A. Swami. «Mining Association Rules Between Sets of Items in Large Databases.» *SIGMOD Conference*. 1993. 207-216.

Avignone, Angela. «Differenze di efficienza nel trasporto regionale italiano.» Torino, 2017-2018.

Bentivogli, Chiara, e Eugenia Panicara. «Regolazione decentrata e servizio concentrato: le ferrovie regionali viaggiano su un binario stretto?» *Rivista di politica economica*, 2012: 51-100.

Cambini, Carlo, e Beniamino Buzzo Margari. *Le gare ferroviarie locali. Rapporto III - Documento di ricerca.* HERMES, 2005.

Castoldi, Bruno, e Gianluca Perlini. «Telediagnostica e control room.» *Manutenzione&Trasporti*, 2016: 8-9.

Daniel T. Larose, Chantal D. Larose. *Data Mining and Predictive Analytics.* Wiley, 2015.

De Agostini. *Data Mining.* 5 Giugno 2020. https://www.sapere.it/enciclopedia/data+mining.html.

Giunta, Marinella. *Infrastrutture ferroviarie.* Reggio Calabria: Università Mediterranea di Reggio Calabria, 2018.

Guidi Buffarini, Giuseppe. «Parte 1: Sintesi storica dello sviluppo.» *L'elettrificazione ferroviaria in Italia dal 1900 al 2000*. Roma, 2007.

Hofmann, Markus, e Ralf Klinkenberg. «Rapidminer: data mining use cases and business analytics applications.» *Chapman & Hall/CRC Data mining and knowledge discovery series*, 2013.

Legambiente. *La situazione e gli scenari del trasporto ferroviario in Italia.* Pendolaria, 2019.

Masini, Paolo. «Dynamic Maintenance Management System: la manutenzione dinamica per il miglioramento continuo di Trenitalia.» *La manutenzione predittiva, CBM (Condition Based Maintenance)*. Napoli, 2015.

Ministero, delle infrastrutture e della mobilità sostenibile. *Infrastrutture - Ferrovie*. s.d. https://www.mit.gov.it/temi/infrastrutture/ferrovie.

Pucci , Giovanni, Paolo Mattera, e Lorenzo Berlincioni. *Telediagnostica*. 1 Luglio 2016. https://www.manutenzione-online.com/articolo/telediagnostica.

Sarnataro, A. «Applicazione di un software relazionale al sistema informativo della manutenzione.» *Ingegneria ferroviaria (Rivista di tecnica e di economia dei trasporti)* 51, n. 12 (1996): 894-901.

Verma, A., S. Das Khan, J. Maiti, e O. B. Krishna. «Identifying patterns of safety related incidents in a steel plant using association rule mining of incident investigation reports.» *Saf. Science*, 2014: 89-98.

Vijay Kotu, Bala Deshpande, PhD. *Predictive Analytics and Data Mining - Concepts and Practice with RapidMiner*. Morgan Kaufmann, 2014.

Viola , Vera. «Napoli-Portici, la prima linea ferroviaria italiana a doppio binario.» *Il sole 24 ore*, 2009: https://www.ilsole24ore.com/art/napoli-portici-prima-linea-ferroviaria-italiana-doppio-binario-AC7Dyzo.