RESTART38 STUDIO DI INGEGNERIA

AUTOMATION ANOMALY DETECTION

Data-driven vs Rules-driven approach

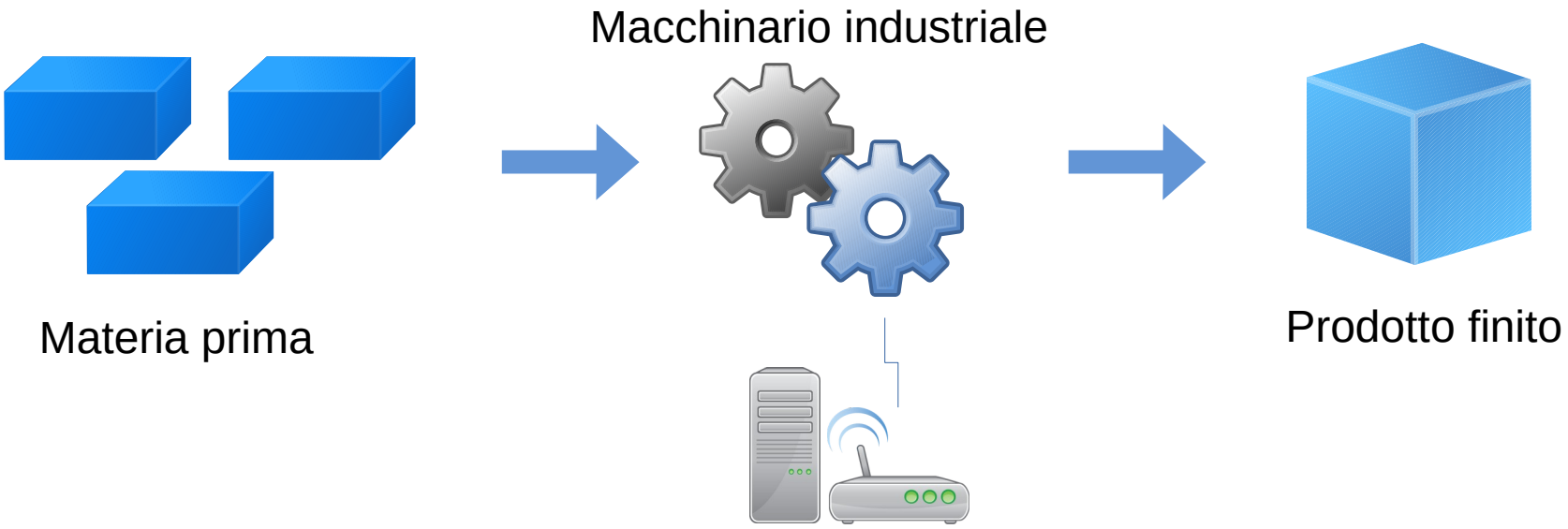
COS'È L'ANOMALY DETECTION?

Wikipedia

Nell'analisi dei dati, l'individuazione delle anomalie è generalmente intesa come l'identificazione di elementi rari, eventi o osservazioni che si discostano in modo significativo dalla maggior parte dei dati e non sono conformi a una nozione ben definita di comportamento normale. Tali esempi possono destare il sospetto di essere generati da un meccanismo diverso, o apparire incoerenti con il resto dell'insieme di dati.

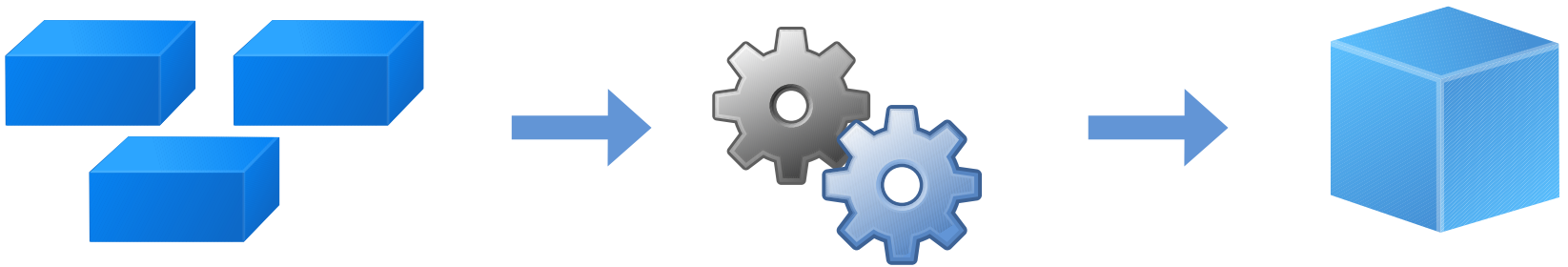
SITUAZIONE E CONTESTO

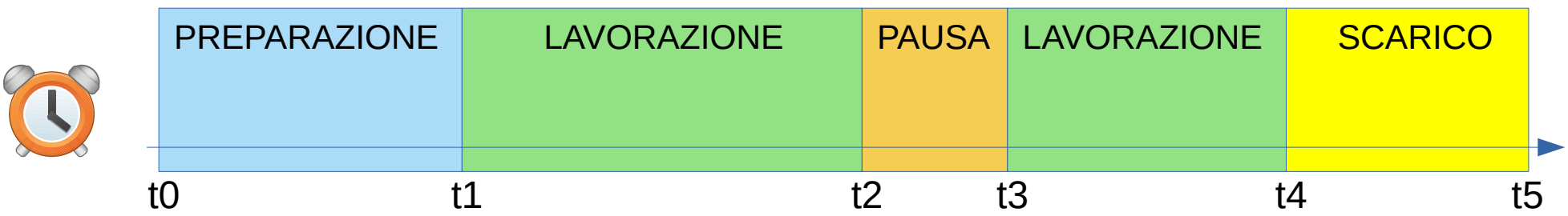
Commessa di trasformazione



SITUAZIONE E CONTESTO

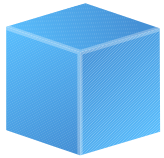
Tempi di lavorazione



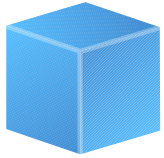


SITUAZIONE E CONTESTO

Costificazione del prodotto

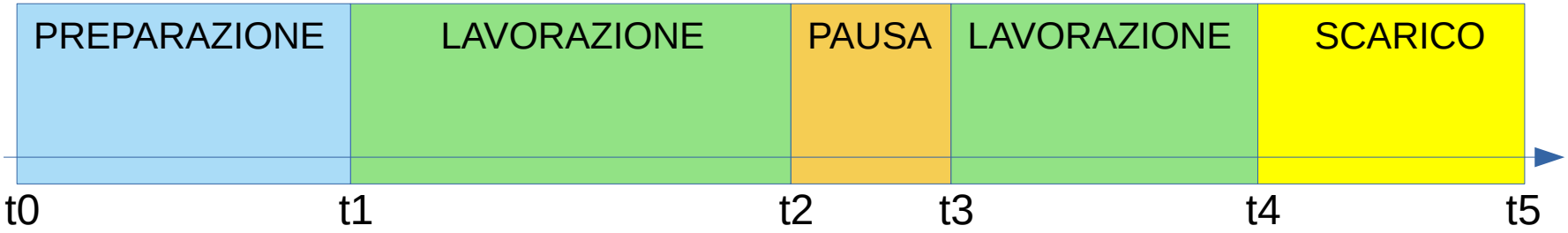


t5-t0 = T → Tempo di produzione



t5’-t0’ = T’ → Tempo di produzione

Una parte del costo del prodotto è direttamente proporzionale a T!!



SITUAZIONE E CONTESTO

Errori di costificazione

Errori di costificazione in eccesso e in difetto avvelenano:

- Valorizzazione di magazzino

- Listino prezzi

- Valutazioni di controllo gestione

ANALISI E METODO

- Acquisizione del dominio
  - Interviste mirate al personale preposto
  - Osservazione sul campo delle modalità operative

- Analisi di dettaglio delle anomalie rilevate durante la revisione sistematica di fine anno

ANALISI E METODO

Risultati dell’analisi

- Elevata variabilità delle lavorazioni possibili
- Presenza di diverse fonti di anomalie di rilevazione
- Identificazione da parte del controllo umano di falsi positivi

ANALISI E METODO

Risultati dell'analisi

- Variabilità delle lavorazioni possibili
  - molte distinte base di materiale per lo stesso prodotto
  - diverse distinte base di ciclo per lo stesso prodotto
  - imballo e altre caratteristiche del prodotto finito differenti per ogni capitolato cliente
  - differenti operatori
  - differenti macchine che possono realizzare lo stesso prodotto

ANALISI E METODO

Risultati dell'analisi

- Diverse fonti di anomalie di rilevazione:
  - macchine con rilevazione manuale + errore/dimenticanza umana
  - errori temporanei di sistema
  - errore/dimenticanza umana nella programmazione della commessa
  - processi di produzione eccezionali non gestiti

ANALISI E METODO

Risultati dell’analisi

Attenzione ai falsi positivi!

Alcune COP identificate come anomale dal controllo manuale sono risultate formalmente corrette. La costificazione che sembrava errata era invece giustificata da eventi e da situazioni imprevisti, ma comunque corretti dal punto di vista della rilevazione dei tempi.

ANALISI E METODO

Dove da qui?

In sintesi: esistono regole che si possono sintetizzare dalla conoscenza di dominio, ma non è certo che riescano a coprire tutte le casistiche.

Utilizziamo due approcci!

Natural Intelligence vs Artifical Intelligence

ANALISI E METODO

I due approcci

Natural Intelligence: un approccio classico tale per cui un analista programmatore sintetizza, dalla conoscenza del dominio già acquisita e da eventuali ulteriori indagini, le condizioni e le regole da codificare mediante un linguaggio di programmazione procedurale che permettono di identificare le informazioni in un set di dati fornito in input. Chiameremo questo approccio **Rules-driven**.

Artificial Intelligence: selezione e applicazione di algoritmi di Machine Learning da parte di un Data Scientist con la finalità di far emergere implicitamente le regole dai dati. Chiameremo questo approccio **Data-driven.**

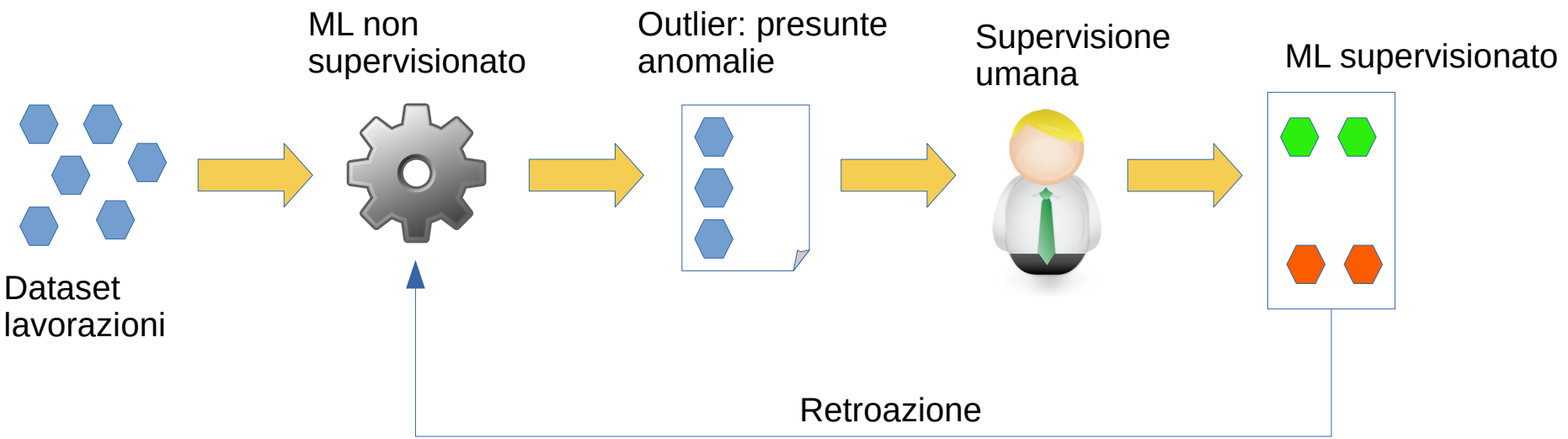
TEAM DATA-DRIVEN

L’Anomaly Detection come molti metodi di ML è caratterizzata da due tipologie di approcci: supervisionato e non supervisionato.

- **Supervisionato**: ogni campione di dato da valutare viene etichettato come anomalo o non anomalo da un supervisore umano. Nel caso del problema specifico che si sta cercando di risolvere questa metodologia non è applicabile per due motivi:
  - costringerebbe un umano specializzato ad un controllo sistematico completo e quotidiano;
  - il metodo sarebbe soggetto all'inaffidabilità di giudizio umano rilevato durante la fase di analisi.
- **Non supervisionato**: si basa sull'identificazione degli outlier, ovvero nel caso specifico di quelle lavorazioni che non rientrano nella massa secondo una valutazione multidimensionale dei parametri definiti come significativi.

TEAM DATA-DRIVEN

Metodo semi-supervisionato



TEAM DATA-DRIVEN

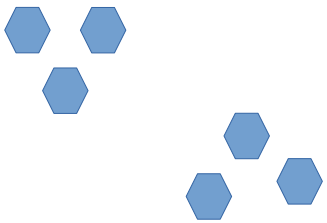
Fasi operative

- **Data Exploration**: studio del dataset alla ricerca di problematiche o per comprendere relazioni tra le variabili.
- **Data Cleaning**: pulizia e preparazione del dataset affinché sia in un formato idoneo al modello utilizzato.
- **Modeling**: definizione e allenamento dei modelli che identificano le anomalie.
- **Evaluation**: confronto tra i modelli scelti per valutarne le performance e selezionare il migliore.
- **Sviluppo**: messa in produzione del modello.

TEAM DATA-DRIVEN

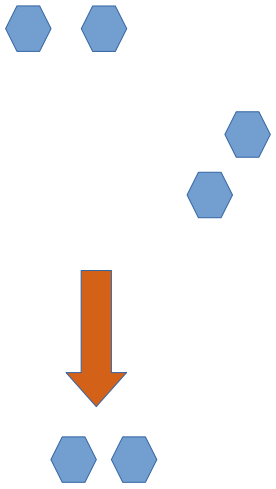
Fasi operative

Data exploration



Vengono identificati due dataset:
- elenco lavorazioni
- elenco giornate di lavoro

Data cleaning



Gli elementi già evidentemente anomali vengono rimossi dai dataset

Modeling

Modello A

Modello B

Modello C

Test dei modelli identificati dal Data Scientist

Evaluation

Modello A

Modello C

Valutazione dei risultati e selezione dei modelli migliori

TEAM DATA-DRIVEN

Modelli utilizzati

L'**Isolation Forest** è una tecnica che si basa sulla costruzione di binary decision trees per identificare le anomalie. Tra i vari metodi è molto utilizzato per la sua leggerezza computazionale e per l'indipendenza dalla struttura del dato in quanto non richiede assunzioni sul dataset di partenza.

IF crea ricorsivamente binary tree passando un sotto insieme del dataset originale. Gli alberi prendono la loro forma grazie a split effettuati su variabili scelte randomicamente che andranno a generare la struttura ramificata.

Le osservazioni che sono maggiori di un threshold definito per quella variabile seguiranno un percorso differente da quelle minori.

In conclusione si avrà un insieme di alberi che smistano le osservazioni in base ai valori che assumono per le variabili selezionate. Più un'osservazione si muove nell'albero meno è probabile che sia un outlier.

TEAM DATA-DRIVEN

Modelli utilizzati

One Class Support Vector Machine (OCSVM) è una variazione del classico Support Vector Machine (SVM) in quanto viene utilizzato per identificare una singola classe internamente al dato definendo come anomale le osservazioni che non ricadono in quella classe.

Il suo funzionamento è molto semplice in quanto crea un confine di decisione entro il quale sono racchiuse le osservazioni ‘normali’. Un suo punto di forza è la possibilità di ottimizzare la definizione di questo boundary mediante la scelta di un kernel, parametro scelto a priori che permette di identificare anche relazioni non-lineari.

A differenza dell’IF risulta essere computazionalmente oneroso al crescere delle dimensioni del dataset.

TEAM RULES-DRIVEN

Elementi valutati:

- Lavorazioni
- Turni di lavoro
- Singole fasi

Tipologia di regole:

- Regole puntuali
- Regole di scostamento

TEAM RULES-DRIVEN

Regole puntuali

if (elemento.parametro1 > sogliaUp || elemento.parametro1 < sogliaDw) { /* elemento è una presunta anomalia */ return 1; }

E’ necessario identificare in modo preciso sia parametro1 sia sogliaUp sia sogliaDw attraverso la comprensione del dominio.

- una specifica tipologia di fase dura più del dovuto
- in un turno di lavoro il tempo di lavorazione è troppo alto o troppo basso
- una specifica lavorazione è priva di una fase obbligatoria

TEAM RULES-DRIVEN

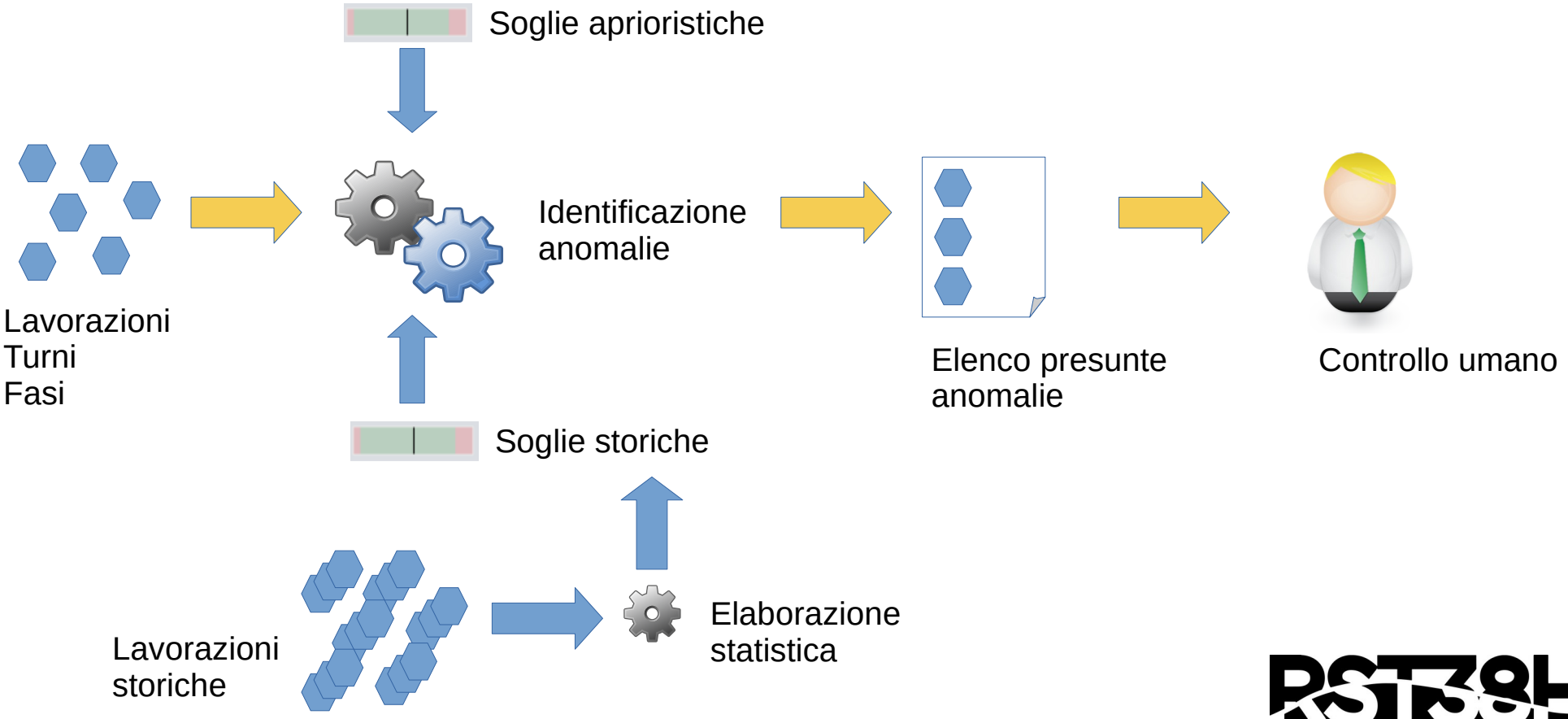
Regole di scostamento

if (elemento.parametro1 > storicoUp || elemento.parametro1 < storicoDw) { /* elemento è una presunta anomalia */ return 1; }

E’ necessario identificare in modo preciso parametro1 ed estrarre dai dati storici le soglie Up e Dw di riferimento utilizzando dei metodi statistici. Le regole di scostamento vengono applicate alle singole lavorazioni sui seguenti parametri:

- velocità di marcia
- kg prodotti nell’unità di tempo
- rapporto tra le durate della fase di lavorazione rispetto alle altre fasi

TEAM RULES-DRIVEN



COSA DICEVA ERACLITO

Tutto cambia

- Valutazione pesata dei dati storici: i dati recenti sono più attendibili.
  - nel tempo possono essere cambiate le metodologie di lavoro, l'esperienza del personale, l'efficienza delle macchine e altri parametri simili
  - valutazione applicata da entrambi i team
- Rivalutazione periodica delle soglie storiche: rules-driven team
- Riaddestramento periodico dei modelli di ML: data-driven team

DATA-DRIVEN VS RULES-DRIVEN

Chi vince?

Ovviamente entrambi!

I due approcci vengono utilizzati assieme per fornire al cliente un risultato migliore.

Inoltre essi collaborano aiutandosi in modo reciproco:

- ML può scartare le anomalie evidenti perché esse sono intercettate dall'approccio rules-driven
- i feedback dell'approccio rules-driven alimentano il sistema di supervisione del ML
- I cluster definiti dal ML contribuiscono a definire le soglie storiche dell'approccio rules-driven

DATA-DRIVEN E RULES-DRIVEN

Lavorazioni anomale

Oggi | Maggio 2024 | 22/05/2024

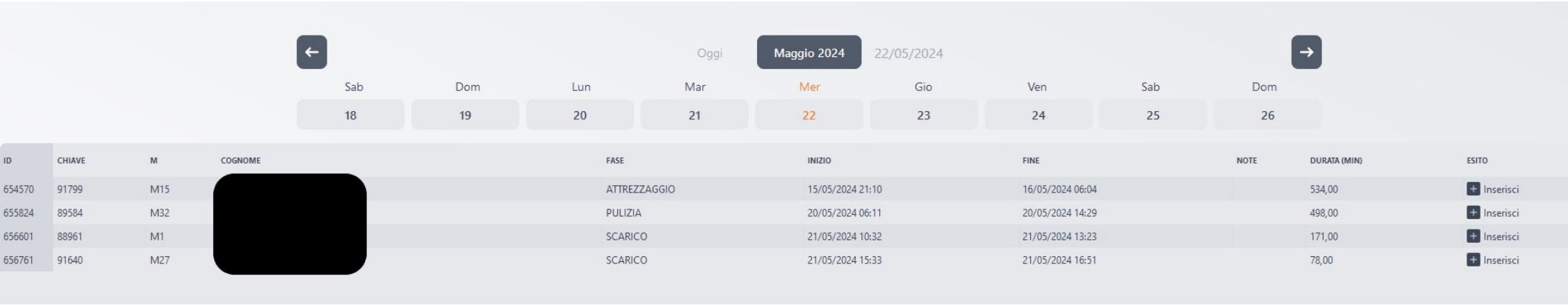
Sab	Dom	Lun	Mar	Mer	Gio	Ven	Sab	Dom
18	19	20	21	22	23	24	25	26

ID RECORD	ID LAVORAZIONE	COMMESSA	TOTALE MIN	MARCIA MIN	KG	CHIAVE	METRI	M	KG/MIN LAVORATI	KG/MIN MARCIA	MT/MIN LAVORATI	MT/MIN MARCIA	MARCIA/TOTALE	MT/MIN	KG/MIN	MARCIA/TOT	MARCIA 0	ANOMALIE RILEVATE	ESITO
76654/1	[redacted]50-700-1090-32	COP/1463/76654/1	209,00	172,00	575,00 / 500,00	90300	5400,00	M32	0,36	0,30	25,00	31,00	82%				No	Pochi kg, Troppe pause	Inserisci
77465/1	[redacted]-1670-8	COP/2276/77465/1	211,00	160,00	333,00 / 300,00	91146	9600,00	M8	0,63	0,48	45,00	60,00	76%				No	Pochi kg, Troppa marcia	Inserisci
77636/1	[redacted]1100-13	COP/2447/77636/1	529,00	397,00	0,00 / 400,00	91340	10000,00	M13	0,00	0,00	18,00	25,00	75%				No	Pochi kg	Inserisci
77880/1	[redacted]-645-11	COP/2691/77880/1	205,00	174,00	149,00 / 200,00	91559	4500,00	M11	1,38	1,17	21,00	25,00	85%				No	Troppo lenta, Troppi kg	Inserisci
77922/1	[redacted]-GEN-0-27	COP/2733/77922/1	107,00	74,00	553,00 / 536,00	91632	0,00	M27	0,19	0,13	0,00	0,00	69%				No	Pochi kg, Troppa marcia	Inserisci
77928/1	[redacted]50-GEN-0-27	COP/2739/77928/1	47,00	14,00	319,00 / 303,00	91638	0,00	M27	0,15	0,04	0,00	0,00	30%				No	Pochi kg, Troppe pause	Inserisci
77930/1	[redacted]70-GEN-0-27	COP/2741/77930/1	161,00	59,00	517,00 / 536,00	91640	0,00	M27	0,31	0,11	0,00	0,00	37%				No	Pochi kg	Inserisci
77934/1	[redacted]50-GEN-0-27	COP/2745/77934/1	106,00	68,00	634,00 / 606,00	91644	0,00	M27	0,17	0,11	0,00	0,00	64%				No	Pochi kg	Inserisci
77970/1	[redacted]-GEN-0-27	COP/2781/77970/1	345,00	160,00	1210,00 / 1213,00	91680	0,00	M27	0,29	0,13	0,00	0,00	46%				No	Pochi kg, Troppe pause	Inserisci
78012/1	[redacted]GEN-0-14	COP/2823/78012/1	161,00	133,00	1925,00 / 1915,00	91734	0,00	M14	0,08	0,07	0,00	0,00	83%				No	Pochi kg	Inserisci
78012/1	[redacted]GEN-0-20	COP/2823/78012/1	63,00	50,00	1925,00 / 1915,00	91734	0,00	M20	0,03	0,03	0,00	0,00	79%				No	Pochi kg	Inserisci
78068/1	[redacted]GEN-0-15	COP/2879/78068/1	913,00	244,00	1781,00 / 1837,00	91799	0,00	M15	0,51	0,14	0,00	0,00	27%				No	Pochi kg, Troppe pause	Inserisci
78201/1	[redacted]GEN-0-15	COP/3012/78201/1	97,00	63,00	545,00 / 537,00	91956	0,00	M15	0,18	0,12	0,00	0,00	65%				No	Pochi kg, Troppe pause	Inserisci
78272/1	[redacted]GEN-0-34	COP/3070/78272/1	467,00	302,00	4997,00 / 6429,00	92010	0,00	M34	0,09	0,06	0,00	0,00	65%				No	Pochi kg, Troppe pause	Inserisci
78277/1	[redacted]010-0-14	COP/3075/78277/1	86,00	49,00	364,00 / 360,00	92029	0,00	M14	0,24	0,13	0,00	0,00	57%				No	Pochi kg	Inserisci
78281/1	[redacted]010-0-14	COP/3079/78281/1	77,00	50,00	622,00 / 612,00	92033	0,00	M14	0,12	0,08	0,00	0,00	65%				No		
78290/1	[redacted]0AA0000-STA-0-29	COP/3088/78290/1	516,00	93,00	104,00 / 103,00	92041	0,00	M29	4,96	0,89	0,00	0,00	18%				No	Pochi kg	Inserisci
78293/1	[redacted]040-STA-0-29	COP/3091/78293/1	473,00	293,00	2388,00 / 2424,00	92044	0,00	M29	0,20	0,12	0,00	0,00	62%				No	Pochi kg	Inserisci

RST38H®
Restart38 Studio di Ingegneria

DATA-DRIVEN E RULES-DRIVEN

Fasi con durata anomala



ID	CHIAVE	M	COGNOME	FASE	INIZIO	FINE	NOTE	DURATA (MIN)	ESITO
654570	91799	M15	[redacted]	ATTREZZAGGIO	15/05/2024 21:10	16/05/2024 06:04		534,00	Inserisci
655824	89584	M32		PULIZIA	20/05/2024 06:11	20/05/2024 14:29		498,00	Inserisci
656601	88961	M1		SCARICO	21/05/2024 10:32	21/05/2024 13:23		171,00	Inserisci
656761	91640	M27		SCARICO	21/05/2024 15:33	21/05/2024 16:51		78,00	Inserisci

DATA-DRIVEN E RULES-DRIVEN

Lavorazioni incomplete

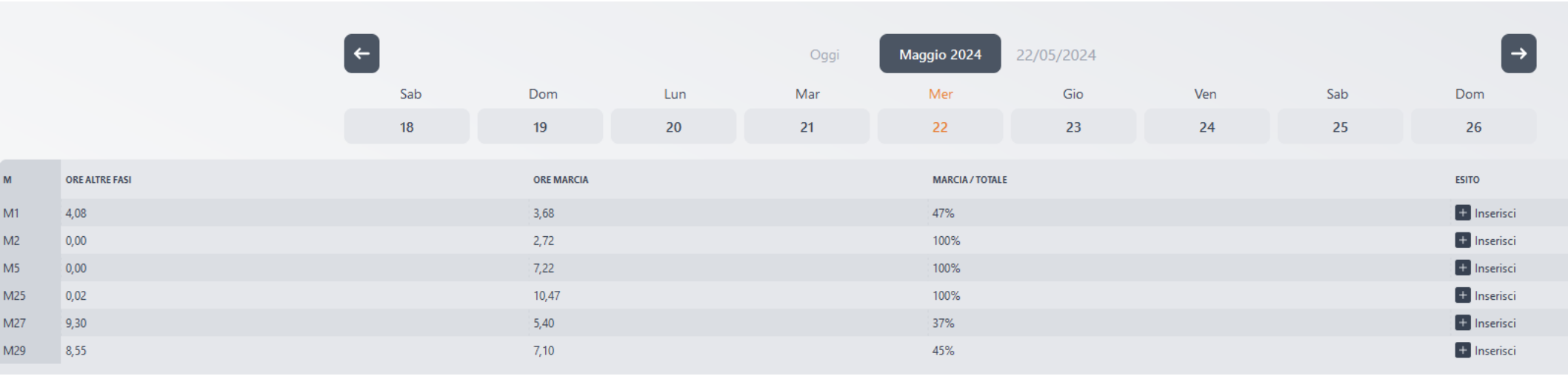
Oggi | Maggio 2024 | 22/05/2024

Sab	Dom	Lun	Mar	Mer	Gio	Ven	Sab	Dom
18	19	20	21	22	23	24	25	26

ID RECORD	ID LAVORAZIONE	COMMESSA	TOTALE MIN	MARCIA MIN	KG	CHIAVE	ESITO
77517/1	LAY29#B02300AO050-700-1879-32	COP/2328/77517/1	158,00	65,00	477,00 / 500,00	91207	Inserisci
77922/1	LAY29#B05400AO070-GEN-0-27	COP/2733/77922/1	107,00	74,00	553,00 / 536,00	91632	Inserisci
76654/1	LAY30S#B00290RI08050-700-1090-32	COP/1463/76654/1	209,00	172,00	575,00 / 500,00	90300	Inserisci
77934/1	LAY30S#B07000RI07050-GEN-0-27	COP/2745/77934/1	106,00	68,00	634,00 / 606,00	91644	Inserisci
78060/1	PY#R00200R019020-235-3970-5	COP/2871/78060/1	433,00	433,00	0,00 / 65,00	91753	Inserisci

DATA-DRIVEN E RULES-DRIVEN

Turni con rapporti anomali



Oggi | Maggio 2024 | 22/05/2024

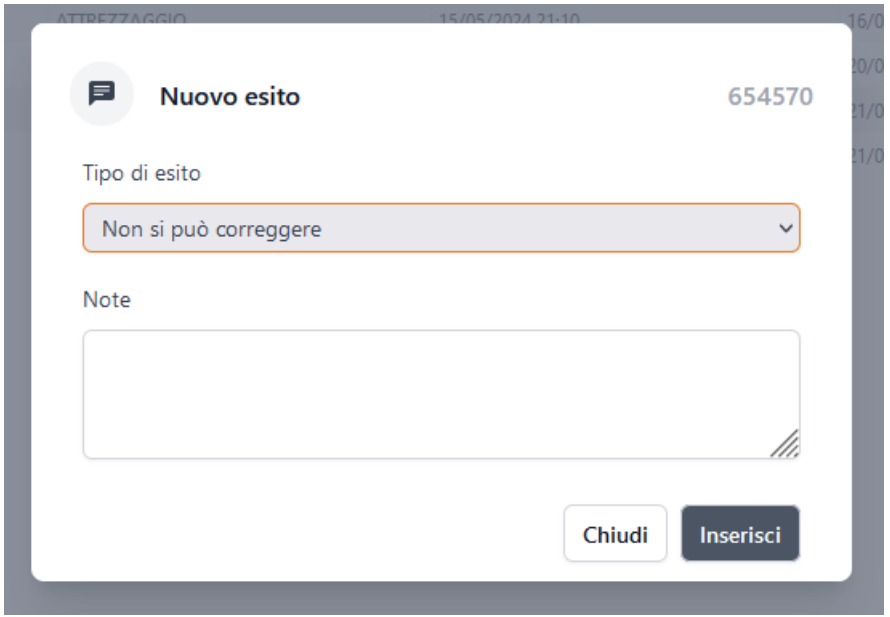
Sab	Dom	Lun	Mar	Mer	Gio	Ven	Sab	Dom
18	19	20	21	22	23	24	25	26

M	ORE ALTRE FASI	ORE MARCIA	MARCIA / TOTALE	ESITO
M1	4,08	3,68	47%	Inserisci
M2	0,00	2,72	100%	Inserisci
M5	0,00	7,22	100%	Inserisci
M25	0,02	10,47	100%	Inserisci
M27	9,30	5,40	37%	Inserisci
M29	8,55	7,10	45%	Inserisci

RST38H®
Restart38 Studio di Ingegneria

DATA-DRIVEN E RULES-DRIVEN

Supervisione umana



- va bene così
- anomalia corretta
- non si può correggere

PROS E CONS

Data-driven	Rules-driven
Intercetta anomalie senza intercettare una regola esplicita	Intercetta solo le anomalie per cui si è definita una regola specifica
Difficoltà di categorizzazione delle anomalie rilevate	Precisa categorizzazione delle anomalie rilevate
Adattamento automatico all'insorgere di nuove tipologie di anomalie	Nuove tipologie di anomalie devono essere intercettate mediante nuove regole
Fase di analisi del dominio è più limitata, l'informazione sorge dai dati	Fase di analisi del dominio è fondamentale, l'informazione sorge dalla comprensione del dominio
Richiede competenze di data analisi, matematiche e di ML	Richiede la capacità di comprendere il dominio

RISULTATI OTTENUTI

Questa pagina è volutamente lasciata in bianco

PROSSIMI STEP

- Addestramento dei modelli di ML e delle soglie storiche
- Raccolta dei dati di supervisione dal pannello web preposto
- Introduzione delle reti neurali e del deep learning
- Analisi dei risultati ottenuti per introdurre a monte accorgimenti finalizzati ad eliminare l’insorgere delle anomalie