UNIVERSITÀ POLITECNICA DELLE MARCHE

FACOLTÀ DI INGEGNERIA

Corso di Laurea Triennale in Ingegneria Gestionale

DIPARTIMENTO DI INGEGNERIA INDUSTRIALE E SCIENZE MATEMATICHE (DIISM)

ALGORITMI DI MACHINE LEARNING PER L'ANOMALY DETECTION DI IMPIANTI MULTIFASE NEL SETTORE OIL&GAS

Machine learning algorithms for multiphase plant anomaly detection in the oil&gas sector

RELATORE:

Prof. Maurizio Bevilacqua

CORRELATORE:

Prof. Giovanni Mazzuto

TESI LAUREA DI:

Federico Ciccotosto

Anno Accademico 2021/2022

SOMMARIO

In questa tesi è stata trattata una piccola parte del mondo del machine learning, l'anomaly detection. Inizialmente è stata fatta una breve introduzione sul machine learning, sulle tipologie di apprendimento automatico che possiamo trovare e sono inoltre stati esposti gli step di una attività di machine learning. Successivamente si è passati alla spiegazione dell' identificazione delle anomalie ed è stata fatta una breve trattazione delle tipologie di anomaly detection. È stato poi descritto il caso studio, un impianto multifase collocato nel Dipartimento di Ingegneria Industriale e Scienze Matematiche (DIISM) dell'Università Politecnica delle Marche e le problematiche che si intendeva affrontare. Gli algoritmi utilizzati sono stati Angle Based Outlier Detection, Isolation Forest, k-Nearest Neighbors Detector, Local Outlier Factor, Class Outlier Factor, One-class SVM, Minimum Covariance Determinant, Stochastic Outlier Selection, DBSCAN. Sono stati valutati i risultati ottenuti e infine presentati sotto forma di grafici e tabelle.

INDICE

INDICE DELLE FIGURE

INDICE DELLE TABELLE

1. INTRODUZIONE

Valutare la presenza di anomalie in un generico sistema può essere una attività molto utile. Tuttavia non è sempre semplice rilevare una anomalia. L'operatore potrebbe effettuare la ricerca manuale delle anomalie ma queste potrebbero richiedere molto tempo per essere individuate e molto tempo per emergere e spesso hanno già fatto danni rilevanti quando vengono trovate. La connettività e il flusso di informazioni e dati tra dispositivi e sensori consente un'abbondanza di dati disponibili. Il fattore chiave è quindi essere in grado di utilizzare queste enormi quantità di dati disponibili ed estrarre effettivamente informazioni utili, consentendo di ridurre i costi, ottimizzare la capacità e ridurre al minimo i tempi di fermo. Oggi è possibile effettuare il rilevamento delle anomalie tramite algoritmi di machine learning in quanto, con l'aumento dei dati a disposizione, il rilevamento manuale diventa inefficace.

L'obiettivo di questa tesi è quello di utilizzare questi algoritmi al fine di stabilire quali dati sono considerati "normali" e quali "anomali" utilizzando il dataset composto dalle letture effettuate sull'impianto sperimentale multifase collocato nel Dipartimento di Ingegneria Industriale e Scienze Matematiche (DIISM) dell'Università Politecnica delle Marche (Ancona, Italia).

2. MACHINE LEARNING

2.1 Introduzione machine learning

Un modello matematico è uno strumento che permette di passare dalla semplice raccolta e organizzazione dei dati osservati ad analisi, interpretazione e previsione del comportamento futuro del sistema, utilizzando per questo un linguaggio matematico. In passato la modellazione matematica era essenzialmente legata alla fisica. Oggi la modellazione matematica è una disciplina che svolge un ruolo fondamentale in molte scienze, sia della natura sia della società. Il machine learning permette di non dover considerare come è fatto il sistema da modellare potendo quindi prescindere dalla complessità del sistema stesso.

Il machine learning è un sottoinsieme dell'intelligenza artificiale ed è una collezione di tecniche che permettono di estrarre informazioni dai dati. Esso può essere supervisionato, semi-supervisionato o non supervisionato.

- Apprendimento supervisionato: nell'apprendimento supervisionato, il set di dati include esempi etichettati . In altre parole, l'insieme contiene le risposte corrette (note come target) di ciascun esempio. L'obiettivo è sviluppare un modello che prenda come input un vettore di caratteristiche e produca una variabile target. Due tipici compiti dell'apprendimento supervisionato sono la classificazione e la regressione.
- Apprendimento non-supervisionato: nell'apprendimento non supervisionato il set di dati è una raccolta di esempi senza etichetta. Questo tipo di apprendimento è utile

soprattutto quando si desidera eseguire il clustering, riduzione della dimensionalità o rilevamento anomalie.

- Apprendimento semi-supervisionato: alcuni algoritmi possono gestire dati di addestramento parzialmente etichettati, di solito molti dati senza etichetta e alcuni con etichetta. Un esempio è Google Photos che riconosce automaticamente che la stessa persona A compare nelle foto 1,5,11 e un'altra persona compare nelle foto 2,5,7. Questa è la parte non supervisionata dell'algoritmo. Ora il sistema ha bisogno che gli venga indicato il nome di quelle persone. In questo modo il sistema sarà in grado di rinominare ogni persona in ogni foto.
- Apprendimento con rinforzo: il sistema di apprendimento può osservare l'ambiente, selezionare, eseguire azioni e dare premi o penalità in cambio, deve imparare da solo quale è la migliore strategia per avere la ricompensa più grande. La strategia definisce quali azioni deve fare l'algoritmo di apprendimento quando si trova in una determinata situazione.

Alcune applicazioni del machine learning sono:

- Riconoscimento vocale: i modelli ML possono essere addestrati e utilizzati nel contesto dell'elaborazione del linguaggio naturale per elaborare il linguaggio umano e convertirlo in un formato scritto (noto anche come sintesi vocale). Siri, Alexa e Google Assistant sono esempi perfetti.
- Sistema di suggerimenti: gli algoritmi di Machine Learning vengono utilizzati dalle aziende per fornire agli utenti consigli basati su ciò che potrebbero aver acquistato in passato o in base a ciò che utenti simili hanno acquistato.
- Computer vision: un'altra area dell'apprendimento automatico e dell'intelligenza artificiale è la computer vision che consente ai computer di estrarre informazioni da

immagini e video digitali. Alcune applicazioni della computer vision includono le carte per la guida autonoma e l'imaging medico. [1]

2.2 Workflow attività Machine learning

Figura 2.1: workflow attività di machine learning

1. Raccolta e esplorazione dei dati.
2. Pretrattamento dei dati.
3. Ricerca del modello migliore per il tipo di dati.
4. Addestrare e testare il modello.
5. Valutazione.

2.2.1 Raccolta e esplorazione dei dati

Il processo di raccolta dati varia in base al tipo di progetto che si vuole realizzare. Possiamo infatti avere flussi di dati in tempo reale e flussi di dati "batch", dati interni o esterni all'azienda, dati provenienti da database, file, sensori e molte altre fonti simili, ma i dati raccolti non possono essere utilizzati direttamente per eseguire il processo di analisi poiché potrebbero esserci molti dati mancanti, valori estremamente grandi, disorganizzati

dati di testo o dati rumorosi. Pertanto, per risolvere questo problema viene eseguita la preparazione dei dati. [2]

2.2.2 Pretrattamento dei dati

È necessario andare a pretrattare i dati perché i dati raccolti sono dati grezzi e cioè con delle imperfezioni che devono essere valutate. Un set di dati viene considerato grezzo se contiene:

-Dati incompleti: manca il valore di alcuni attributi, o mancano del tutto alcuni attributi.
-Dati inaccurati: cioè dati che contengono valori errati o che si discostano da valori attesi.
-Dati inconsistenti: questo tipo di dati potrebbe essere raccolto a causa di errori umani (errori con il nome o valori) o duplicazione di dati.

Principali tecniche nella fase di pre-elaborazione:

- Data cleaning: riempire i campi con i valori mancanti, rimuovere la componente rumorosa nei dati, rimuovere i valori non realistici. Possibili approcci quando si hanno dati mancanti sono quelli di ignorare le istanze con valori mancanti, riempire i valori mancanti manualmente, usare la media dell'attributo o predire il valore dell'istanza mancante sulla base delle altre istanze note attraverso algoritmi di machine learning.
- Data trasformation: preparare i dati per l'uso con alcuni particolari algoritmi di analisi. Ad esempio molto spesso è necessario modificare la scala dei dati in modo che cadano in intervalli stabiliti (normalizzazione) oppure è necessario aggregare dati, cioè, ad esempio, è possibile trasformare una serie temporale di transazioni in conteggi delle vendite giornaliere o, viceversa, costruire degli attributi a partire da quelli esistenti per aiutare l'algoritmo di analisi.

- Data reduction: ridurre la mole dei dati in input senza compromettere la validità delle analisi. Attività tipiche che vengono eseguite sono la riduzione della dimensionalità, riduzione della numerosità e aggregazione stessa.

2.2.3 Ricerca del modello migliore per il tipo di dati

Un modello è l'output di un algoritmo di apprendimento. Quest'ultimo trova nei dati di addestramento i pattern e i trend e genera un modello ML che acquisisce questi pattern. L'obiettivo principale è quello di addestrare il modello più performante possibile, utilizzando i dati pre-elaborati.

Possiamo avere modelli di apprendimento supervisionato quando abbiamo un insieme di esempi di addestramento (i dati di input) in cui i segnali di output desiderati (le etichette) sono già noti e modelli di apprendimento non supervisionato in cui abbiamo dati senza etichetta o dati dalla struttura ignota.

La classificazione è una sottocategoria dell'apprendimento con supervisione in cui l'obiettivo è predire le etichette sulla base delle precedenti osservazioni. Tali etichette delle classi sono valori discreti e non ordinati, che sanciscono l'appartenenza a un gruppo delle istanze. L'esempio del rilevamento dello spam nella posta elettronica rappresenta un tipico esempio di un compito di classificazione binario. Un secondo tipo di apprendimento con supervisione è la predizione di risultati continui, un compito chiamato anche analisi a regressione. Nell'analisi a regressione, abbiamo un certo numero di variabili predittrici (descrittive) e una variabile di risposta continua (il risultato) e tentiamo di trovare una relazione fra tali variabili che ci consenta di predire un risultato. Notiamo che nel campo del machine learning, le variabili predittrici sono comunemente chiamate features o caratteristiche e le variabili di risposta sono normalmente chiamate variabili target.

Alcuni dei più importanti sottocampi dell'apprendimento non supervisionato sono:

- Visualizzazione dei dati.
- Clustering.
- Dimensionality reduction.
- Anomaly detection.

Gli algoritmi di visualizzazione richiedono in input dati anche complessi senza etichetta e forniscono in output una rappresentazione 2D o 3D dei dati stessi. Questi algoritmi cercano di preservare il più possibile la struttura, in questo modo è possibile capire come i dati sono organizzati e identificare patterns.

Il clustering è una tecnica di analisi esplorativa dei dati che ci consente di organizzare una massa di informazioni in sottogruppi significativi (cluster) senza alcuna precedente conoscenza delle loro appartenenze a gruppi.

La riduzione della dimensionalità senza supervisione è un approccio comunemente usato nella pre-elaborazione delle caratteristiche con lo scopo di eliminare il rumore dai dati, di degradare le prestazioni predittive di determinati algoritmi e di comprimere i dati in un sottospazio a dimensionalità inferiore mantenendo nel contempo la maggior parte delle informazioni rilevanti. Talvolta, questa tecnica può essere utile anche per la visualizzazione dei dati.

L'anomaly detection consiste nel riconoscere quali sono le istanze anomale da quelle normali/di regime. Questa tipologia di problema verrà trattata nel capitolo 3.

2.2.4 Addestrare, testare e validare il modello

Il dataset viene suddiviso in training set, test set e validation set.

- Training set: un insieme di dati utilizzati per l'apprendimento, ovvero per adattarsi ai parametri del modello. Solitamente è molto maggiore del test e validation set.
- Validation set: una serie di dati utilizzati per ottimizzare i parametri di un modello.
- Test set: un insieme di dati utilizzati solo per valutare le prestazioni di un modello.

Il tasso di errore che commettiamo sui nuovi dati viene chiamato errore di generalizzazione e valutando il modello attraverso il test set si avrà una stima di questo errore. Questo valore esprime quanto bene il modello performerà su nuove istanze che non ha mai visto prima.

Un altro aspetto importante è l'overfitting e l'underfitting. Overfitting significa che il modello si adatta troppo ai dati di addestramento ma non generalizza bene su nuovi esempi. Le soluzioni possibili sono quelle di:

-ridurre il numero di attributi nel training set o vincolare il modello;

-fornire più dati di training;

-ridurre il rumore nel training set.

La procedura di vincolare un modello per renderlo più semplice è chiamata regolarizzazione e permette di trovare il giusto compromesso tra adattare bene il modello ai dati e mantenere semplice il modello stesso per essere sicuri che esso generalizzerà bene su nuovi dati. La regolarizzazione viene gestita attraverso gli iperparametri.

L'underfitting ,invece, si verifica quando il modello è troppo semplice per apprendere la struttura sottostante dei dati.

2.2.5 Valutazione

La valutazione del modello è parte integrante del processo di sviluppo del modello stesso. Aiuta a trovare il miglior modello che rappresenti i nostri dati e quanto bene

funzionerà il modello scelto in futuro. Per migliorarlo potremmo mettere a punto gli iperparametri del modello, provare a migliorare l'accuratezza e anche guardare la matrice di confusione per cercare di aumentare il numero di veri positivi e veri negativi. [3]

3. ANOMALY DETECTION

Che cos'è il rilevamento delle anomalie?

Il rilevamento delle anomalie è il compito di identificare gli elementi, gli eventi o le osservazioni rare che sollevano sospetti differendo in modo significativo dalla maggior parte dei dati. Tipicamente gli elementi anomali si tradurranno in qualche tipo di problema come frode bancaria, un difetto strutturale, problemi medici o errori in un testo. Un problema centrale è che non c'è una definizione unica che permette di valutare quanto sono simili due dati. Esistono tre tipi di anomalie:

1. Anomalia puntuale: una singola istanza di dati devia dal resto del dataset. La figura 3.1 mostra i punti anomali. N1 e N2 sono le regioni che hanno un comportamento normale mentre O1 e O2 sono i punti anomali.
2. Anomalia contestuale: quando una singola istanza è considerata un'anomalia solo in specifici contesti. In questi casi, il contesto deve essere specificato come parte della definizione del problema. Questo tipo di anomalia si presenta, solitamente, nei dati influenzati dal tempo. Un classico esempio è la rilevazione di una temperatura di 5° a Chieti che può essere considerata un'anomalia se rilevata durante la stagione estiva (es. nel mese di Giugno) ma nella norma se si considera la stagione invernale (es. nel mese di Dicembre).
3. Anomalia collettiva: quando un gruppo di istanze di dati assumono un comportamento anomalo rispetto al resto del dataset ma non individualmente.

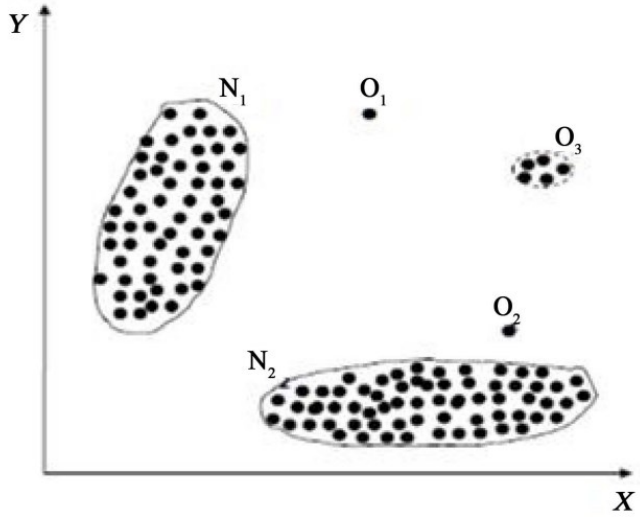


Figura 3.1:esempio di anomalia puntuale [4]

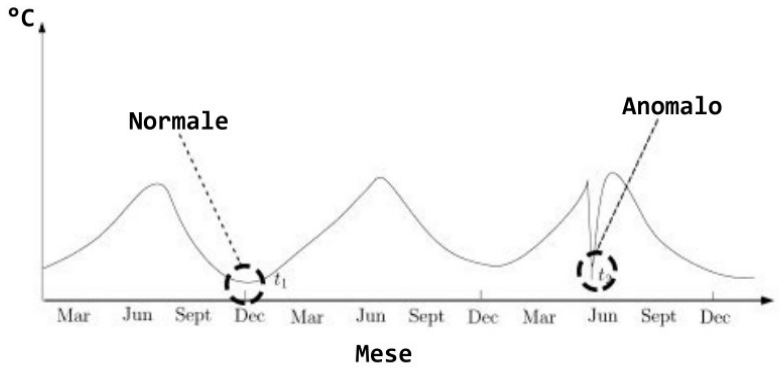


Figura 3.2:esempio di anomalia contestuale [4]

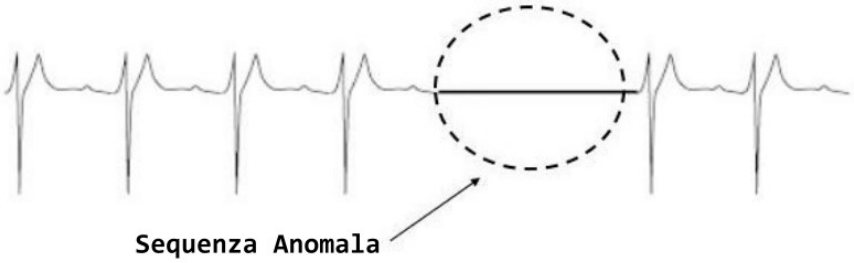


Figura 3.3:esempio di anomalia collettiva [4]

3.1 Tecniche di anomaly detection

In letteratura esistono diversi modelli di anomaly detection e in questa sezione vengono presentati due possibili modi di categorizzare queste tecniche. Una prima classificazione si basa sulla natura del dataset che si utilizza per la costruzione di un modello di anomaly detection. A seconda se i campioni di dati di un dataset vengono forniti con delle etichette assegnate da esperti in materia, le tecniche di anomaly detection vengono classificate in modelli con apprendimento supervisionato, semi-supervisionato e non supervisionato. Un altro modo di classificare questi modelli si basa sui metodi di separazione delle anomalie rispetto al resto del dataset. In questo caso possiamo classificarli in tre tipi: metodi statistici, metodi basati sulla prossimità e metodi basati sul clustering. Il processo di etichettatura di un dataset viene spesso eseguito manualmente da esperti del settore e richiede, solitamente, un notevole sforzo sia in termini di analisi che in termini di tempo. [5]

RILEVAMENTO ANOMALIE NON SUPERVISIONATO

In alcuni scenari applicativi, gli oggetti etichettati come "normali" o "anomali" non sono disponibili.

Pertanto, è necessario utilizzare un metodo di apprendimento senza supervisione.

I metodi di rilevamento dei valori anomali non supervisionati presuppongono implicitamente che gli oggetti normali sono in qualche modo "raggruppati". In altre parole, un metodo di rilevamento dei valori anomali non supervisionato prevede che gli oggetti normali seguano uno schema molto più frequentemente dei valori anomali.

Gli oggetti normali non devono necessariamente rientrare in un gruppo che condivide un'elevata somiglianza. Infatti essi possono formare più gruppi, in cui ogni gruppo ha caratteristiche distinte. Tuttavia, si prevede che un valore anomalo si verifichi molto lontano nello spazio delle caratteristiche da uno qualsiasi di quei gruppi di oggetti normali.

I valori anomali collettivi condividono un'elevata somiglianza in una piccola area. I metodi senza supervisione non possono rilevare efficacemente tali valori anomali. In alcune applicazioni, gli oggetti normali sono distribuiti in modo diverso e molti di questi oggetti non seguono schemi forti. Ad esempio, in alcuni problemi di rilevamento delle intrusioni e di virus informatici, le normali attività sono molto diverse e molte non rientrano in cluster di alta qualità. In tali scenari, i metodi senza supervisione possono avere un alto tasso di falsi positivi, possono etichettare erroneamente molti oggetti normali come valori anomali (intrusioni o virus in queste applicazioni) e lasciare che molti valori anomali effettivi non vengano rilevati. A causa dell'elevata somiglianza tra intrusioni e virus (ovvero, devono attaccare le risorse chiave nei sistemi di destinazione), la modellazione dei valori anomali utilizzando metodi supervisionati può essere molto più efficace. [5]

RILEVAMENTO ANOMALIE SUPERVISIONATO

I modelli di anomaly detection supervisionati sono utilizzati quando i campioni di dati, che costituiscono il dataset, sono stati etichettati come “normali” o “anomali” da esperti in

materia. Tipico approccio in questi casi è trattare l'anomaly detection come un problema di classificazione. Sebbene in letteratura esistono molti modelli di classificazione da applicare, le principali sfide per l'anomaly detection supervisionato sono le seguenti:

• Le due classi (cioè normale vs anomala) sono tipicamente sbilanciate. Ovvero, la popolazione di dati etichettati come anomali è generalmente molto minore rispetto ai dati normali. In questi casi devono essere implementati metodi per la gestione dello sbilanciamento delle classi. Un tipico metodo è l'oversampling delle anomalie il quale, ha il compito di aumentare la loro distribuzione nel training-set utilizzato per la costruzione del classificatore. La mancanza di campioni anomali può limitare l'efficienza dei classificatori.

• In molte applicazioni di anomaly detection, identificare il maggior numero di anomalie è molto più importante che classificare erroneamente un comportamento normale come anomalia. Di conseguenza, in fase di valutazione, quando un modello di classificazione è utilizzato come anomaly detection supervisionato bisognerebbe dare più peso al parametro Recall per minimizzare la perdita di identificazione delle anomalie. Nell'ambito della classificazione, il "Recall", è definito come il numero dei veri positivi diviso il numero totale di elementi che appartengono alla classe su cui si sta calcolando questa metrica.

RILEVAMENTO ANOMALIE SEMI-SUPERVISIONATO

In molte applicazioni, sebbene sia possibile ottenere alcuni esempi etichettati, il numero di tali esempi etichettati è spesso piccolo. È possibile incontrare casi in cui solo un piccolo insieme di oggetti normali e/o anomali è etichettato, ma la maggior parte dei dati non lo è. Per affrontare tali scenari sono stati sviluppati metodi di rilevamento dei valori anomali semi-supervisionati. Questi metodi possono essere considerati applicazioni di metodi di

apprendimento semi-supervisionati. Ad esempio, quando sono disponibili alcuni oggetti normali etichettati, possiamo usarli, insieme a oggetti senza etichetta che si trovano nelle vicinanze, per addestrare un modello per oggetti normali. Il modello composto da istanze normali può quindi essere utilizzato per rilevare valori anomali, quegli oggetti che non si adattano al modello di istanze normali sono classificati come valori anomali.

Se sono disponibili solo alcuni valori anomali etichettati, il rilevamento dei valori anomali semi-supervisionato è più complicato. È improbabile che un piccolo numero di valori anomali etichettati rappresenti tutti i possibili valori anomali. Pertanto, è improbabile che la creazione di un modello per valori anomali basato su pochi valori anomali etichettati sia efficace.

Secondo le ipotesi fatte, possiamo classificare i valori anomali in tre tipi: metodi statistici, metodi basati sulla prossimità e metodi basati sul clustering. [5]

METODI STATISTICI

I metodi statistici (noti anche come metodi basati su modelli) fanno ipotesi di normalità dei dati. Presuppongono che gli oggetti normali siano generati da un modello statistico (stocastico) e che i dati che non seguono il modello siano valori anomali. Di conseguenza, gli oggetti normali si trovano in regioni ad alta probabilità per il modello stocastico e gli oggetti nelle regioni a bassa probabilità sono valori anomali. Tuttavia, ci sono molti modi diversi per apprendere i modelli generativi. In generale, i metodi statistici per il rilevamento dei valori anomali possono essere suddivisi in due categorie principali, metodi parametrici e metodi non parametrici, a seconda di come i modelli vengono specificati e appresi. Un metodo parametrico presuppone che gli oggetti normali siano generati da una distribuzione parametrica con parametro θ. La funzione di densità di probabilità della

distribuzione parametrica f (x, θ) fornisce la probabilità che l'oggetto x sia generato dalla distribuzione.

Più piccolo è questo valore, più è probabile che x sia un valore anomalo.

Un metodo non parametrico non presuppone un modello statistico a priori. Esso tenta di determinare il modello dai dati di input. Si noti che la maggior parte dei metodi non parametrici non presuppone che il modello sia completamente privo di parametri (una tale ipotesi renderebbe quasi impossibile l'apprendimento del modello dai dati). Generalmente, i metodi non parametrici spesso assumono la posizione che il numero e la natura dei parametri sono flessibili e non fissati in anticipo.

I metodi non parametrici spesso fanno meno ipotesi sui dati e quindi possono essere applicabili in più scenari. [5]

APPROCCI BASATI SULLA PROSSIMITÀ

Dato un insieme di oggetti nello spazio delle caratteristiche, è possibile utilizzare una misura della distanza per quantificare la somiglianza tra gli oggetti. Intuitivamente, gli oggetti che sono lontani dagli altri possono essere considerati valori anomali. Gli approcci basati sulla prossimità presuppongono che la vicinanza di un oggetto anomalo ai suoi vicini devii significativamente dalla vicinanza dell'oggetto alla maggior parte degli altri oggetti nel set di dati.

Esistono due tipi di metodi di rilevamento dei valori anomali basati sulla prossimità, metodi basati sulla distanza e metodi basati sulla densità. Un metodo di rilevamento dei valori anomali basato sulla distanza consulta le vicinanze che ha un oggetto, le quali sono definite da un determinato raggio. Un oggetto è quindi considerato un valore anomalo se il suo vicinato non ha abbastanza punti vicini. Un metodo di rilevamento delle anomalie

basato sulla densità indaga la densità di un oggetto e quella dei suoi vicini. Quindi un oggetto è identificato come un valore anomalo se la sua densità è molto inferiore a quella dei suoi vicini. [5]

APPROCCI BASATI SUL CLUSTERING

La nozione di valori anomali è strettamente correlata a quella di cluster. Gli approcci basati sul clustering rilevano i valori anomali esaminando la relazione tra oggetti e cluster. Intuitivamente, un outlier è un oggetto che appartiene a un cluster piccolo e remoto o non appartiene a nessun cluster.

È possibile fare delle considerazioni. Se consideriamo un oggetto, questo appartiene a qualche cluster? In caso contrario, viene identificato come un valore anomalo. C'è una grande distanza tra l'oggetto e il cluster a cui è più vicino? Se si, è un valore anomalo. L'oggetto fa parte di un cluster piccolo o sparso? Se sì, tutti gli oggetti in quel cluster sono valori anomali.

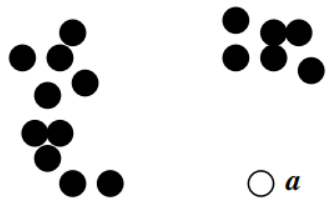


Figura 3.4: tipologia anomalia

L'oggetto a è un valore anomalo perché non appartiene a nessun cluster

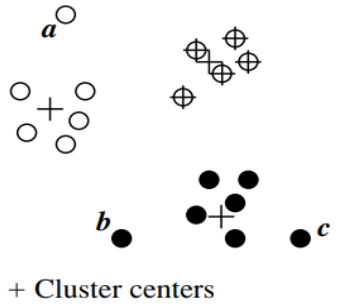


Figura 3.5: tipologia di anomalia

Gli oggetti a, b e c sono valori anomali perché sono lontani dai cluster a cui sono più vicini [5]

3.2 Metodi di valutazione

ACCURATEZZA

L'accuratezza è definita come la frazione dell'insieme di dati di test su di cui il modello fornisce una previsione corretta. È per definizione un numero compreso tra 0 ed 1. A volte è espressa in percentuale.

$$\text{ACC} = \frac{n.\ di\ previsioni\ corrette}{n.\ di\ campioni\ nell'insieme\ di\ test}$$

Un'accuratezza pari a 0.98 vuol dire che, su 100 campioni, il nostro modello fornisce previsioni sbagliate soltanto su 2 campioni.

L'accuratezza è sicuramente una prima metrica da prendere in considerazione ma non distingue tra falsi positivi e falsi negativi. In tutta una serie di situazioni è importante distinguerli e quantificarli e, pertanto, abbiamo a disposizione altre metriche.

CONFUSION MATRIX

La matrice di confusione , nota anche come matrice di errore, è rappresentata da una matrice che descrive le prestazioni di un modello di classificazione su un insieme di dati di test.

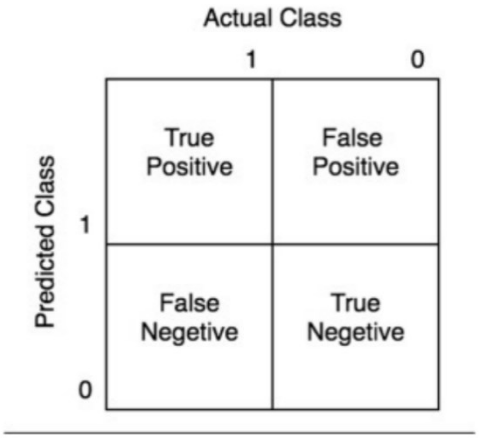


Figura 3.6: confusion matrix

Vero positivo (TP): l'osservazione è prevista positiva ed è effettivamente positiva.

Falso positivo (FP): l'osservazione è prevista positiva ed è effettivamente negativa.

Vero negativo (TN): l'osservazione è prevista negativa ed è effettivamente negativa.

Falso negativo (FN): l'osservazione è prevista negativa ed è effettivamente positiva.

F1-SCORE

Per spiegare questa metrica occorre spiegarne altre due: precision e recall. Queste altre due metriche prestazionali vogliono quantificare il tasso di Veri Positivi (TP) e di Veri Negativi (TN) in modo da poter stabilire come massimizzarli.

Queste sono le definizioni:

$$\text{PREC} = \frac{TP}{TP+FP}$$

$$\text{REC} = \frac{TP}{P}$$

dove:

- TP = numero di veri positivi
- FP: numero di falsi positivi
- P = numero totale di campioni positivi nel SET

La Precision è la "frazione di casi identificati come positivi che sono correttamente positivi". Ad esempio, nel caso di un modello rivolto alla diagnosi di una malattia, la precision è la frazione di pazienti identificati come malati che sono correttamente malati.

Il Recall invece è la "frazione di positivi che sono identificati dal modello come positivi".

La Precision peggiora se vi sono tanti falsi positivi. Il Recall peggiora se vi sono tanti falsi negativi.

L'F1-score, calcolato a partire dalla Precision e Recall, è:

$$\text{F1-score} = \frac{(2*precision*recall)}{precision+recall}$$

COEFFICIENTE DI SILHOUETTE

Viene utilizzato per valutare le prestazioni di un algoritmo di clustering. Si riferisce a un metodo di interpretazione e convalida della coerenza all'interno di cluster. Il coefficiente di silhouette permette di valutare il cluster utilizzando il cluster stesso.

È definito per ogni campione ed è composto da due punteggi:

- **a** : La distanza media tra un campione e tutti gli altri punti della stessa classe.
- **b** : La distanza media tra un campione e tutti gli altri punti nel cluster successivo più vicino.

Il coefficiente di silhouette *s* per un singolo campione è quindi dato come:

$$s = \frac{b - a}{\max(a, b)}$$

Vantaggi:

- Il punteggio è compreso tra -1 per il clustering errato e +1 per il clustering altamente denso. Punteggi intorno a zero indicano cluster sovrapposti.
- Il punteggio è più alto quando i cluster sono densi e ben separati, il che si riferisce a un concetto standard di cluster.

4. CASO STUDIO

4.1 L'impianto sperimentale e il sistema di acquisizione

L'impianto sperimentale si trova presso il Dipartimento di Ingegneria Industriale e Scienze Matematiche (DIISM) dell'Università Politecnica delle Marche (Ancona, Italia). Il modello 3D dell'impianto è mostrato nella Figura 4.1. Riproduce l'estrazione di petrolio e gas naturale da pozzi esauriti. In particolare, la vita utile dei giacimenti di idrocarburi è correlata al loro potenziale e ai costi operativi. Un pozzo si esaurisce se l'acqua al suo interno è in quantità tali da non poter essere estratta, oppure se i volumi di idrocarburi prodotti diventano antieconomici visti gli altissimi costi operativi per produrli, insostenibili se la produzione è scarsa o nulla. In tali circostanze, la soluzione più ovvia sarebbe l'installazione di apposite pompe poste in superficie e alla base del pozzo petrolifero con un costo molto elevato rispetto ai volumi di idrocarburi prodotti. Il sistema in esame rappresenta una soluzione indubbiamente più efficiente. L'estrazione da un giacimento esaurito viene effettuata sfruttando la pressione di un ipotetico giacimento al culmine della sua vita utile. Per le sue caratteristiche fisiche, la pressione di quest'ultimo è superiore alla pressione di trasporto e, quindi, in grado di creare aspirazione sul giacimento esaurito, che, al contrario, non ha una pressione sufficiente per il trasporto sulla linea. Gli eiettori gas-liquido possono miscelare due fasi a pressioni diverse (esaurimento e pozzi buoni) e impartire l'energia di trasporto necessaria. Mentre in una situazione realistica, i fluidi trattati sono petrolio greggio e gas naturale, per motivi di sicurezza vengono utilizzati acqua e aria ambiente nel caso dell'impianto sperimentale. In particolare, l'acqua in un serbatoio e una pompa volumetrica modellano il comportamento di un pozzo pressurizzato, mentre l'aria ambiente simula il gas naturale di un giacimento esaurito. L'acqua in pressione (“INLET WATER) entra nell'eiettore, creando un vuoto che aspira una certa

quantità di aria dall'ambiente (“INLET AIR”), creando così una miscela bifasica (“INLET MIXTURE”). La miscela così ottenuta viene convogliata in un serbatoio verticale che funge da raccoglitore di lumache per separare la fase liquida (“OUTLET WATER”) e gas (“OUTLET AIR”). L'impianto è dotato di tre valvole pneumatiche: per controllare la pressione dell'acqua in ingresso (V1), per regolare la pressione all'interno del serbatoio (V2) e il livello dell'acqua (V3). Tutte le “VM” in Figura 4.1 rappresentano valvole di intercettazione utilizzate per riprodurre anomalie nell'impianto. [6]

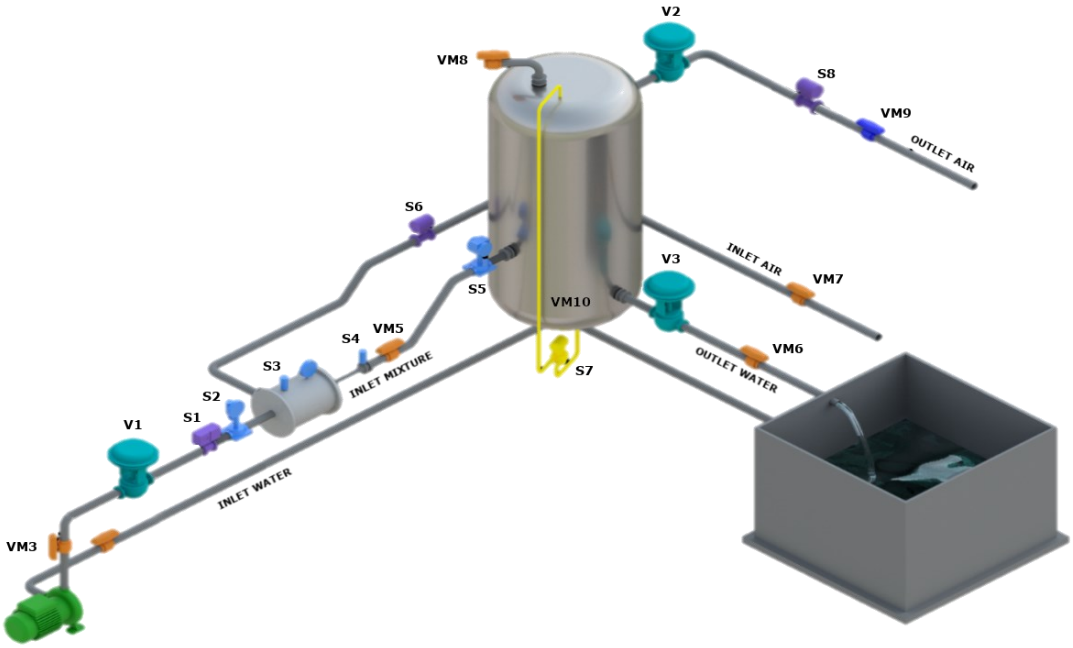


Figura 4.1: il modello 3D dell'impianto sperimentale.

Il dataset utilizzato, chiamato “DatasetAnomalies.xlsx”, è composto da 21 anomalie ottenute agendo sulle valvole di intercettazione degli impianti.

Tutte le anomalie sono state riprodotte agendo sulle chiusure delle valvole di intercettazione (VM), come descritto in Tabella 1.

Nome	Data e tempo	S1_rif [bar]	S7_rif [mm]	S5_rif [bar]	Descrizione
Regime	29/4/2022 10:41:45	5.5	300	1.3	La pressione dell'acqua in ingresso (S1) è fissata a 5,5 bar, la pressione del serbatoio (S5) a 1,3 bar e il livello del serbatoio (S7) a 300 mm.
V10L1	29/4/2022 12:46:49	5.5	300	1.3	Descrive una piccola perdita d'acqua nel serbatoio ottenuta dalla chiusura della valvola VM10 del 30%
V10L2	29/4/2022 12:52:28	5.5	300	1.3	Descrive una perdita d'acqua media da un serbatoio ottenuta dalla chiusura della valvola VM10 del 60%
V10L3	29/4/2022 12:59:44	5.5	300	1.3	Descrive una seria perdita d'acqua da un serbatoio ottenuta dalla chiusura della valvola VM10 del 100%
V3L1	29/4/2022 13:50:11	5.5	300	1.3	Descrive una piccola ostruzione nel sistema di tubazioni di ingresso dell'acqua ottenuta dopo la chiusura della valvola VM3 del 30%
V3L2	29/4/2022 13:52:28	5.5	300	1.3	Descrive una media ostruzione nel sistema di tubazioni di ingresso dell'acqua ottenuta dopo la chiusura della valvola VM3 del 60%
V3L3	29/4/2022 13:54:51	5.5	300	1.3	Descrive una seria ostruzione nel sistema di tubazioni di ingresso dell'acqua ottenuta dopo la chiusura della valvola VM3 del 100%
V5L1	29/4/2022 10:56:33	5.5	300	1.3	Descrive una piccola ostruzione nel sistema di tubazioni di ingresso miscela ottenuta dopo la chiusura della valvola VM5 del 30%
V5L2	29/4/2022 10:59:24	5.5	300	1.3	Descrive un'ostruzione media nel sistema di tubazioni di ingresso miscela ottenuta dalla chiusura della valvola VM5 del 60%
V5L3	29/4/2022 11:4:9	5.5	300	1.3	Descrive una grave ostruzione nel sistema di

					tubazioni di ingresso miscela ottenuta dopo la chiusura della valvola VM5 del 100%
V6L1	29/4/2022 10:44:17	5.5	300	1.3	Descrive una piccola ostruzione nel sistema di tubazioni di uscita dell'acqua ottenuta dopo la chiusura della valvola VM6 del 30%
V6L2	29/4/2022 10:46:22	5.5	300	1.3	Descrive un'ostruzione media nel sistema di tubazioni di uscita dell'acqua ottenuta dopo la chiusura della valvola VM6 del 60%
V6L3	29/4/2022 10:50:56	5.5	300	1.3	Descrive una grave ostruzione nel sistema di tubazioni di uscita dell'acqua ottenuta dopo la chiusura della valvola VM6 del 100%
V7L1	29/4/2022 11:9:54	5.5	300	1.3	Descrive una piccola ostruzione nel sistema di tubazioni di ingresso dell'aria ottenuta dopo la chiusura della valvola VM7 del 30%
V7L2	29/4/2022 11:12:8	5.5	300	1.3	Descrive un'ostruzione media nel sistema di tubazioni di ingresso dell'aria ottenuta dalla chiusura della valvola VM7 del 60%
V7L3	29/4/2022 11:14:7	5.5	300	1.3	Descrive una grave ostruzione nel sistema di tubazioni di ingresso dell'aria ottenuta dopo la chiusura della valvola VM7 del 100%
V8L1	29/4/2022 13:11:46	5.5	300	1.3	Descrive una piccola perdita d'aria nel serbatoio ottenuta dalla chiusura della valvola VM8 del 30%
V8L2	29/4/2022 13:14:49	5.5	300	1.3	Descrive una media perdita d'aria nel serbatoio ottenuta dalla chiusura della valvola VM8 del 60%
V8L3	29/4/2022 13:5:33	5.5	300	1.3	Descrive una grave perdita d'aria nel serbatoio ottenuta dalla chiusura della valvola VM8 del 100%
V9L1	29/4/2022 13:38:48	5.5	300	1.3	Descrive una piccola ostruzione nel sistema di tubazioni di uscita dell'aria ottenuta dopo la chiusura della valvola VM9 del 30%

V9L2	29/4/2022 13:41:53	5.5	300	1.3	Descrive una media ostruzione nel sistema di tubazioni di uscita dell'aria ottenuta dopo la chiusura della valvola VM9 del 60%
V9L3	29/4/2022 13:44:48	5.5	300	1.3	Descrive una grave ostruzione nel sistema di tubazioni di uscita dell'aria ottenuta dopo la chiusura della valvola VM9 del 100%

Tabella 1: descrizione anomalie.

Le variabili presenti nel dataset sono le seguenti: [6]

ID	**DESCRIZIONE**	**Unità di misura**
S1	Pressione dell'acqua in ingresso	[bar]
S2	Portata dell'acqua in ingresso	[m^3/h]
S3	Pressione dell'eiettore	[bar]
S4	Pressione tubo miscelatore	[bar]
S5	Pressione del serbatoio	[bar]
S6	Portata d'aria in entrata	[m^3/h]
S7	Livello dell'acqua del serbatoio	[mm]
S8	% Portata d'aria in uscita	[%]
S9	% Portata d'acqua in uscita	[%]
S10	%Pompa acqua in ingresso	[%]
S1_rif	Punto di riferimento della pressione dell'acqua in ingresso	[bar]
S5_rif	Punto di riferimento della pressione del serbatoio	[bar]

S7_rif	Punto di riferimento del livello dell'acqua del serbatoio	[mm]
PID_S1	Stato del regolatore della pressione dell'acqua in ingresso	[on/off]
PID_S5	Stato del controller della pressione del serbatoio	[on/off]
PID_S7	Stato del regolatore del livello dell'acqua del serbatoio	[on/off]
Da VM1 a VM10	Il valore 1 (0) identifica l'istante di tempo durante il quale si è (non) riprodotta un'anomalia chiudendo l'apposita valvola di intercettazione	[on/off]

Tabella 2: descrizione variabili.

4.2 INTRODUZIONE

Il caso di studio prende in esame l'impianto descritto nel capitolo 4. L'obiettivo dello studio è quello di sviluppare un algoritmo che sia in grado, date le letture sensoristiche dell'impianto, di identificare eventuali anomalie. Il problema dell'identificazione delle anomalie può essere scomposto in due sotto problemi:

1. Valutare se il sistema si trova o meno in una condizione di anomalia.
2. Se il sistema si trova in una condizione di anomalia classificare il tipo di anomalia.

Per affrontare il primo problema è stato utilizzato il linguaggio di programmazione Python ed in particolare la libreria PyCaret che contiene diversi algoritmi utilizzati per l'identificazione delle anomalie. Per quanto riguarda il secondo problema è stato utilizzato

anche in questo caso il linguaggio di programmazione Python e, attraverso la libreria Scikit-learn, è stato possibile utilizzare l'algoritmo DBSCAN per la costruzione di cluster. Questo perché si vuol sapere, qualora il sistema di anomaly detection rilevasse una anomalia, il tipo di anomalia e le sue caratteristiche in relazione al cluster a cui verrà assegnato.

4.3 LIBRERIA PYCARET

PyCaret è una libreria di apprendimento automatico open source a basso codice in Python che mira a ridurre l'ipotesi a informazioni dettagliate sul tempo di ciclo in un esperimento ML. Consente di eseguire esperimenti end-to-end in modo rapido ed efficiente.

Il modulo di rilevamento delle anomalie di PyCaret è un modulo di apprendimento automatico non supervisionato utilizzato per identificare elementi , eventi o osservazioni rari che sollevano sospetti differendo in modo significativo dalla maggior parte dei dati. In genere, gli elementi anomali si tradurranno in un qualche tipo di problema come frode bancaria, un difetto strutturale, problemi medici o errori. Fornisce diverse funzionalità di pre-elaborazione che preparano i dati per la modellazione tramite la funzione di setup. Dispone di oltre 10 algoritmi pronti per l'uso e diversi grafici per analizzare le prestazioni dei modelli addestrati.

Le funzioni principali di questa libreria sono:

- Funzione setup.
- Funzione create_model.
- Funzione assign_model.
- Funzione plot_model.
- Funzione predict_model.

La funzione setup inizializza l'ambiente e prepara i dati di modellazione. Dopo averla eseguita viene dedotto il tipo di ciascun attributo (numerico o categoriale) e sui dati vengono eseguite diverse attività di pre-elaborazione. Come unico parametro obbligatorio viene richiesto il dataframe da analizzare. Ulteriori parametri della funzione setup che sono stati utilizzati per il caso studio sono [7]:

- original_data: visualizza la forma originale del set di dati.
- numeric_features: numero di funzionalità dedotte come numeriche.
- trasformed_data: visualizza la forma del set di dati trasformato.
- preprocess: se impostato su False, non vengono applicate trasformazioni. I dati devono essere pronti per la modellazione quando il parametro preprocess è impostato su False.
- normalize: quando è impostato su True, trasforma le funzionalità ridimensionandole a un determinato intervallo. Il tipo di ridimensionamento è definito dal parametro normalize_method.
- normalize_method: definisce il metodo per il ridimensionamento. Per impostazione predefinita, il metodo di normalizzazione è impostato su 'zscore' che coincide con la standardizzazione del dato. In statistica si definisce standardizzazione un processo attraverso il quale si riconduce una variabile aleatoria distribuita secondo una media μ e una deviazione standard σ ad una variabile aleatoria con distribuzione normale con media 0 e varianza 1.

$$z_{ij} = \frac{x_{ij-\mu_i}}{\vartheta_i}$$

$$\forall j = 1, \dots, m \;\; n.campioni$$

$$\forall i = 1, \dots, p \quad n.variabili$$

I valori di μ_i e ϑ_i sono calcolati dai dati raccolti e, nel caso di studio in esame, fanno riferimento ai valori di varianza campionaria e media campionaria valutati sui dati a regime.

Il parametro normalize_method viene ignorato quando il parametro normalize è impostato su False.

Per il rilevamento delle anomalie dell'impianto sono stati standardizzati e trasformati i dati presenti nel dataset Regime. Come si vede nella Figura 9 non sono stati considerati gli attributi da v_man_1 a v_man_9, presenti nel dataset originale. Questi attributi indicano se una valvola è stata chiusa o meno e ciò equivale a dire se c'è una anomalia o meno, siccome stiamo trattando anomaly detection non supervisionata il modello non deve apprendere da questa tipologia di dato. Questi valori serviranno successivamente per valutare le prestazioni del sistema ma non per allenare il modello. Inoltre sono presenti 3131 campioni e 11 attributi (set di dati trasformato).

```
s=setup(df, normalize=True, ignore_features=('v_man_1','v_man_2','v_man_3','v_man_4','v_man_5','v_man_6','v_man_7','v_man_8','v_man_9'))
```

	Description	Value
0	Session id	7157
1	Original data shape	(3131, 17)
2	Transformed data shape	(3131, 11)
3	Ignore features	9
4	Numeric features	17
5	Preprocess	True
6	Imputation type	simple
7	Numeric imputation	mean
8	Categorical imputation	constant
9	Low variance threshold	0
10	Normalize	True
11	Normalize method	zscore
12	CPU Jobs	-1
13	Use GPU	False
14	Log Experiment	False
15	Experiment Name	anomaly-default-name
16	USI	0244

Figura 4.2: risultati funzione setup inerente al caso studio

Il modello di rilevamento delle anomalie viene creato utilizzando la funzione create_model che accetta un parametro obbligatorio, ovvero il nome dell'algoritmo. Questa funzione restituisce un modello addestrato. Gli algoritmi che la libreria mette a disposizione sono:

- Angle Based Outlier Detection: questa tecnica si basa sull'idea di controllare l'angolo formato da un insieme di tre punti qualsiasi nello spazio delle features. La varianza nella grandezza di tale spazio risulta essere diversa per valori anomali rispetto a quelli considerati normali. Di solito la varianza osservata è maggiore per i punti inlier che per i valori anomali, quindi tale misura ci aiuta a raggruppare i punti normali e quelli anomali in modo diverso.
- Cluster Based Outlier Detection: i metodi di rilevamento dei valori anomali basati sul cluster presuppongono che i normali oggetti di dati appartengano a cluster grandi e densi, mentre i valori anomali appartengano a cluster piccoli o sparsi o non appartengano a nessun cluster.

- Histogram Based Outlier Detection: possiamo utilizzare l'istogramma come strumento statistico non parametrico per acquisire i valori anomali. La procedura è composta da 2 steps, costruzione dell'istogramma e rilevamento dell'anomalia. L'istogramma viene costruito sulla base del training set e potrebbe essere richiesto l'inserimento di alcuni parametri come ad esempio il numero di contenitori. Per determinare se un oggetto è un valore anomalo, possiamo confrontarlo con l'istogramma. Se l'oggetto cade in uno dei contenitore dell'istogramma, l'oggetto è considerato normale. In caso contrario è considerato un valore anomalo. [5]
- Isolation Forest: le foreste di isolamento (IF) sono costruite sulla base di alberi decisionali. Le foreste di isolamento sono state costruite sulla base del fatto che le anomalie sono le istanze che sono "poche e diverse". In una foresta di isolamento, i dati sotto-campionati casualmente vengono elaborati in una struttura ad albero basata su features selezionate casualmente. I campioni che viaggiano più in profondità nell'albero hanno meno probabilità di essere anomalie poiché hanno richiesto più tagli per isolarli. Allo stesso modo, i campioni che finiscono in rami più corti indicano anomalie in quanto è stato più facile per l'albero separarli da altre osservazioni.
- k-Nearest Neighbors Detector: l'algoritmo di classificazione più semplice è l'algoritmo k-Nearest Neighbours (o k-NN in breve) è un metodo non parametrico utilizzato per la classificazione. L'input è costituito dai k esempi di addestramento più vicini nello spazio delle caratteristiche. L'output è un'appartenenza a una classe.
- Local Outlier Factor: il punteggio di anomalia di ogni campione è chiamato Local Outlier Factor. Misura la deviazione locale della densità di un dato campione

rispetto ai suoi vicini. È locale in quanto il punteggio dell'anomalia dipende da quanto è isolato l'oggetto rispetto al quartiere circostante. Confrontando la densità locale di un campione con le densità locali dei suoi vicini, si possono identificare campioni che hanno una densità sostanzialmente inferiore rispetto ai loro vicini. Questi sono considerati valori anomali.

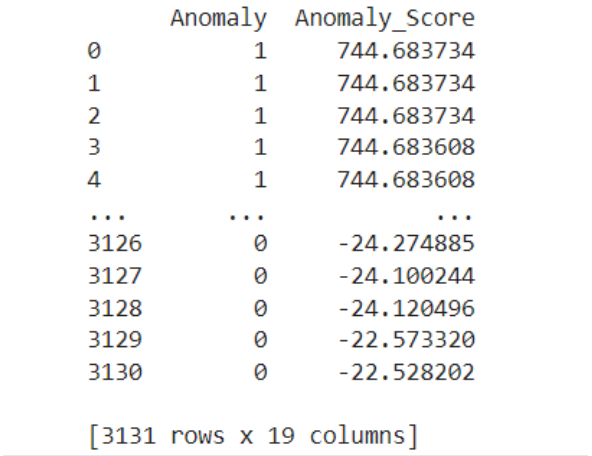
- One-class SVM detector: One-Class Support Vector Machine (SVM) è un modello non supervisionato per il rilevamento di anomalie o valori anomali. A differenza del normale SVM supervisionato, l'SVM a una classe non ha etichette target per il processo di addestramento del modello. One-class SVM apprende il confine dalle istanze "normali" e identifica i dati al di fuori del confine come anomalie.
- Principal Component Analysis: l'analisi delle componenti principali è principalmente una tecnica di riduzione della dimensionalità. Funziona identificando i componenti principali. I componenti principali sono vettori di caratteristiche indipendenti chiamati anche autovettori di un oggetto dato che spiegano la varianza massima delle istanze. Ogni Principal Component è una combinazione lineare di caratteristiche correlate esistenti e giace ortogonale ad altri autovettori. Il rilevamento delle anomalie si basa sull'errore di ricostruzione. Una volta identificati i Principal Component, scegliendo tutti i componenti principali possiamo ricostruire i dati originali dai dati trasformati senza perdita di dati. Allo stesso modo, scegliendo solo Principal Component che spiegano la maggior parte della varianza dovremmo essere in grado di ricreare un'approssimazione dei dati originali. L'errore generato durante la ricostruzione durante la generazione dei dati originali è chiamato errore di ricostruzione. Per anomalie nei dati, l'errore di ricostruzione è elevato.

- Minimum Covariance Determinant: il metodo del minimo determinante di covarianza (MCD) è uno stimatore altamente robusto della localizzazione multivariata e della dispersione. Poiché la stima della matrice di covarianza è la pietra angolare di molti metodi statistici multivariati, l'MCD è un importante elemento costitutivo nello sviluppo di robuste tecniche multivariate.
- Class Outlier Factor: il COF utilizza il rapporto tra la distanza media di concatenamento dell'istanza e la media delle distanze medie di concatenamento dei k-nearest neighbor dell' istanza come punteggio anomalo per le osservazioni.
- Stochastic Outlier Selection: utilizza il concetto di affinità per quantificare la relazione da un punto a un altro punto. L'affinità è proporzionale alla somiglianza tra due punti. Quindi, un dato può avere molta o poca affinità con un dato diverso. Un dato viene selezionato come valore anomalo quando tutti gli altri punti hanno un'affinità insufficiente con esso.

Un altro parametro importante della funzione create_model è il parametro *fraction* cioè la quantità di contaminazione del set di dati, ovvero la proporzione di valori anomali nel set di dati. Questo parametro viene utilizzato durante l'addestramento per definire la soglia sulla funzione di decisione.

Una volta creato il modello, vorremmo assegnare le etichette di anomalia al nostro set di dati per analizzare i risultati. Otterremo questo utilizzando la funzione assign_model.

Si noti dalla figura 4.3 che sono state aggiunte due colonne Anomaly e Anomaly_Score. Il valore 0 sta per inlier e 1 per outlier/anomalia. Gli Anomaly_Score sono i valori calcolati dall'algoritmo. I valori più anomali vengono assegnati con punteggi di anomalia maggiori.



	Anomaly	Anomaly_Score
0	1	744.683734
1	1	744.683734
2	1	744.683734
3	1	744.683608
4	1	744.683608
...	...	...
3126	0	-24.274885
3127	0	-24.100244
3128	0	-24.120496
3129	0	-22.573320
3130	0	-22.528202

[3131 rows x 19 columns]

Figura 4.3: anteprima dei valori non anomali e anomali del dataset di regime

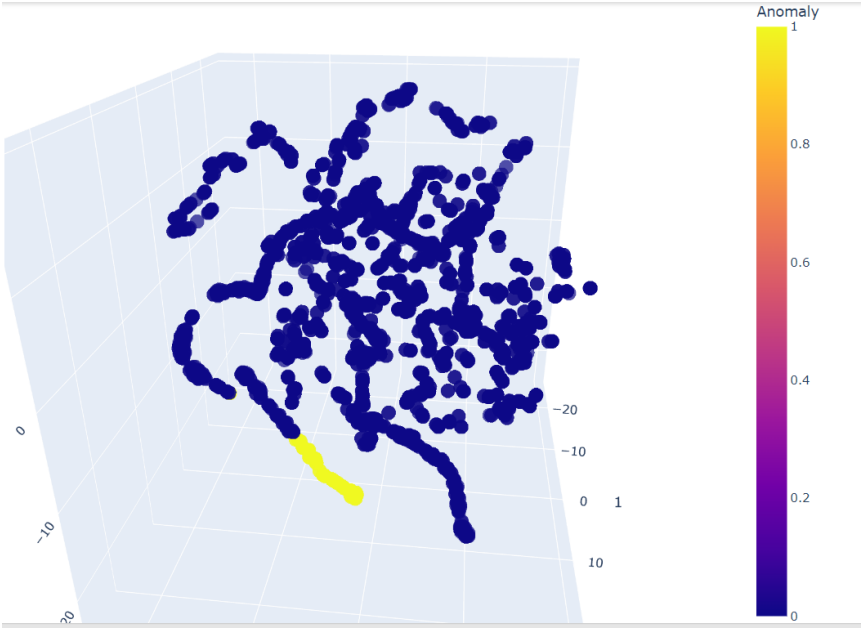


Figura 4.4: grafico relativo al modello addestrato con i dati di regime

La funzione plot_model può essere utilizzata per analizzare il modello di rilevamento delle anomalie su diversi aspetti. Questa funzione prende in input un modello addestrato e restituisce un grafico.

La funzione predict_model viene utilizzata per assegnare etichette di anomalia a un nuovo set di dati invisibile. Utilizzeremo il modello addestrato per prevedere i dati archiviati in una determinata variabile che non sono mai stati esposti al modello [8].

4.4 Classificazione delle anomalie

Per raggruppare le anomalie è stato utilizzato utilizzata la libreria Scikitlearn ed in particolare l'algoritmo DBSCAN il quale richiede due parametri obbligatori:

- eps: la distanza minima che devono avere due dati in ingresso per essere considerati "vicini".

- min_samples: numero minimo di dati in ingresso, “vicini” tra loro, richiesto per formare un cluster.

I valori di questi due parametri che hanno dato risultati migliori secondo i valori del coefficiente di silhouette sono eps=0.5 e min_samples=30. Il dataset utilizzato per affrontare questa problematica è composto dai dataset di tutte le anomalie presenti nella tabella 1 escludendo quelli con chiusura delle valvole del 30% e i dataset V7L1, V7L2,

V7L3, V10L1, V10L2, V10L3. Il dataset così composto è stato anche in questo caso standardizzato e suddiviso in training set (75%) e test set (15%). Il training set è stato suddiviso in cluster attraverso l’algoritmo DBSCAN con parametri eps=0.5 e min_samples=30, questi valori hanno fornito il valore del coefficiente di silhouette di 0.5223 e 17 cluster diversi.

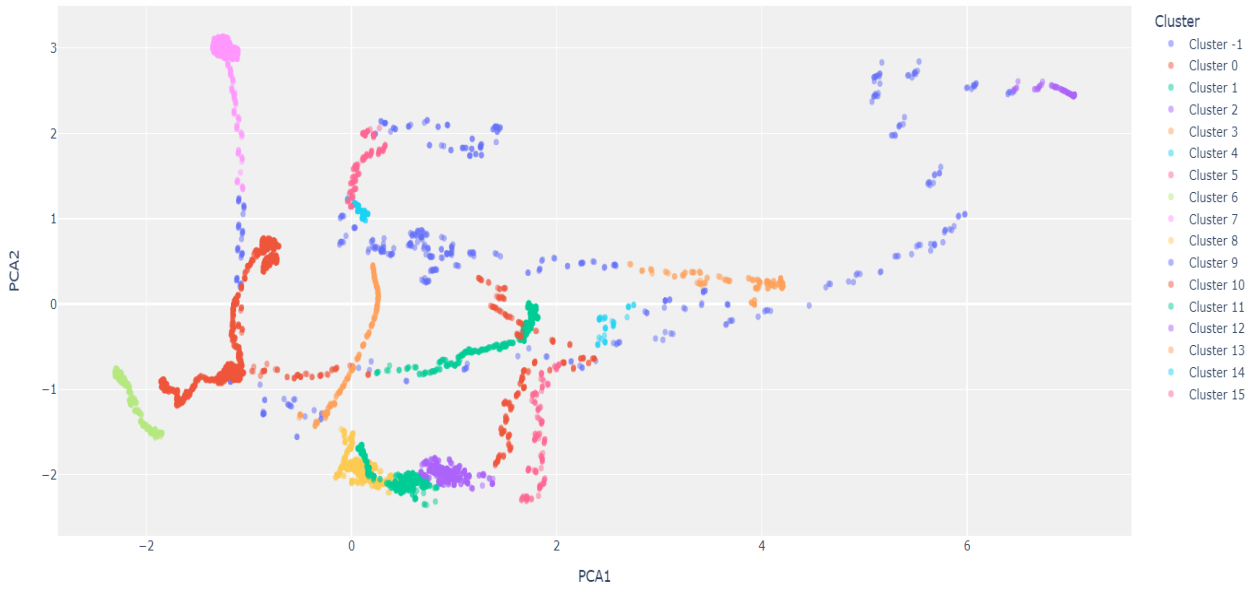


Figura 4.5: grafico dei cluster

La figura 4.5 mostra i cluster in relazione a 2 variabili pca1 e pca2. La principal component analysis è una tecnica popolare per l'analisi di grandi set di dati contenenti un numero elevato di dimensioni/caratteristiche per osservazione che aumenta l'interpretabilità dei dati preservando la quantità massima di informazioni e consentendo la visualizzazione di dati multidimensionali.

È stata fatta una descrizione dei cluster ottenuti in modo tale da vedere quali e quante anomalie erano contenute in ogni cluster e il range di valori che ogni attributo di ogni cluster poteva assumere. Sulla base di queste valutazioni sono state prese le letture del dataset di test ed è stata calcolata la distanza euclidea della lettura dai centroidi di ogni cluster in modo tale da poter dire a quale cluster appartiene ogni singola lettura. Sulla base delle valutazioni corrette ed errate sul test set si è valutato il sistema tramite il parametro dell'accuratezza.

5. RISULTATI OTTENUTI

algoritmi	fraction	V3L2	V3L3	V5L2	V5L3	V6L2	V6L3	V7L2	V7L3	V8L2	V8L3	V9L2	V9L3	V10L2	V10L3
svm	0.8	0.971	0.93	0.993	0.991	0.907	0.89	0.978	0.98	0.986	0.987	0.962	0.988	0.982	0.995
abod	0.015	0.974	0.99	0.979	0.901	0.958	0.981	0.987	0.984	0.985	0.99	0.781	0.989	0.986	0.993
iforest	0.5	0.991	0.942	0.975	0.895	0.937	0.949	0.978	0.99	0.984	0.994	0.376	0.985	0.982	0.995
cluster	0.7	0.971	1	0.962	0.907	0.928	0.93	0.98	0.987	0.986	0.987	0.919	0.988	0.982	0.996
knn	0.3	0.971	1	0.932	0.9	0.893	0.963	0.978	0.943	0.986	0.987	0.982	0.988	0.982	0.986
lof	0.5	0.971	1	0.993	0.9	0.89	0.931	0.978	0.92	0.986	0.987	0.982	0.988	0.982	0.988
cof	0.8	0.8	0.823	0.829	0.769	0.792	0.836	0.832	0.864	0.865	0.865	0.848	0.826	0.868	0.874
mcd	0.4	0.971	1	0.993	0.936	0.968	0.969	0.994	0.991	0.983	0.994	0.979	0.977	0.982	0.996
histogram	0.8	0.991	0.983	0.918	0.9	0.918	0.937	0.978	0.963	0.987	0.986	0.601	0.989	0.982	0.996
pca	0.8	0.971	1	0.957	0.895	0.912	0.905	0.978	0.994	0.986	0.987	0.897	0.99	0.982	0.994
sos	0.9	0.941	0.901	0.859	0.189	0.897	0.906	0.933	0.144	0.953	0.976	0.955	0.944	0.977	0.943

Tabella 3: risultati f1-score di ogni algoritmo per ogni anomalia

algoritmo	media
svm	0.967143
abod	0.962714
iforest	0.926643
cluster	0.965929
knn	0.963643
lof	0.964
cof	0.835071
mcd	0.980929
histogram	0.937786
pca	0.960571
sos	0.822714

Tabella 4: risultato valutazione f1-score media ottenuta da ogni algoritmo.

anomalia	f1-score medio
V3L2	0.876926667
V3L3	0.88075
V5L2	0.865833333
V5L3	0.76525
V6L2	0.833333333
V6L3	0.84975
V7L2	0.882833333
V7L3	0.813333333
V8L2	0.890583333
V8L3	0.895
V9L2	0.7735
V9L3	0.887666667
V10L2	0.890583333
V10L3	0.896333333

Tabella 5:risultato valutazione f1-score media per ogni anomalia

Come si vede dalla Tabella 4 l'algoritmo che ha ottenuto i risultati migliori è il Minimum Covariance Determinant (mcd) il quale ha in media un punteggio di f1-score pari a 0.9809, seguito da One-class SVM detector (svm) e Cluster Based Outlier Detection (cluster), mentre gli algoritmi che hanno fornito i risultati peggiori sono Class Outlier Factors (cof) e

Stochastic Outlier Selection (sos). In particolare l'algoritmo "sos" non è riuscito a identificare correttamente le anomalie V5L3 e V7L3 mentre l'algoritmo "cof" ha ottenuto, in generale, risultati inferiori rispetto alla media.

Nella tabella 5 è stata riportata la valutazione f1-score media per ogni anomalia e, gli algoritmi, hanno ottenuto risultati significativamente inferiori rispetto alla media sulle anomalie V5L3, V7L3, V9L2.

5.1 CLASSIFICAZIONE ANOMALIE: CARATTERIZZAZIONE DEI CLUSTER

	n. istanze
cluster -1	391
cluster 0	2414
cluster 1	589
cluster 2	212
cluster 3	232
cluster 4	117
cluster 5	144
cluster 6	426
cluster 7	1315
cluster 8	685
cluster 9	80
cluster 10	220
cluster 11	583
cluster12	405
cluster 13	141
cluster 14	33
cluster 15	123

Tabella 6: numero di istanze di ogni cluster

anomalia 0 =V3L2
anomalia 1=V3L3
anomalia 2=V5L2
anomalia 3=V5L3
anomalia 4=V6L2
anomalia 5=V6L3
anomalia 6=V8L2
anomalia 7=V8L3
anomalia 8=V9L2
anomalia 9=V9L3

Tabella 7:legenda anomalie

Per ogni cluster che conteneva al suo interno più di una anomalia sono stati riportati i grafici contenenti i valori che assume ogni features in relazione a ogni anomalia presente nel cluster ad eccezione del cluster -1.

5.1.1 CLUSTER -1

	valore max	*valore min*
portata_aria OUT	21.519	0.055
s7_pressione ingresso acqua	5.545	1.079
portata_aria IN	44.031	-0.044
Portata_acqua ingresso L6	10.006	0.083
pressione_eiettore	1.209	0.971
pressione_tubo miscelatore	2.635	1.061
pressione_serbatoio	1.393	1.038
livello_acqua	308.353	110.339
%Aria_out	100	73.87
%Acqua_out	100	0
%Pompa_acqua_in	88.95	0

Tabella 8: valore max e min di ogni attributo del cluster -1

Il cluster -1 contiene gli outliers cioè punti che si discostano particolarmente dalle altre osservazioni e quindi anche dai cluster. Non bisogna confondere gli outliers con le anomalie. Tutto il dataset utilizzato è composto da valori anomali, gli outliers sono valori che si discostano particolarmente dai valori del dataset.

5.1.2 CLUSTER 0

istanze anomalia 0	122
istanze anomalia 1	79
istanze anomalia 2	174
istanze anomalia 3	62
istanze anomalia 4	580
istanze anomalia 5	351
istanze anomalia 6	900
istanze anomalia 8	71
istanze anomalia 9	75

Tabella 9: numero di istanze per ogni anomalia nel cluster 0

	valore max	*valore min*
portata_aria OUT	32.286	0.063
s7_pressione ingresso acqua	5.552	4.647
portata_aria IN	28.084	12.084
Portata_acqua ingresso L6	10.031	9.168
pressione_eiettore	0.974	0.97
pressione_tubo miscelatore	1.365	1.215
pressione_serbatoio	1.334	1.206
livello_acqua	411.326	286.939
%Aria_out	89.17	68.95
%Acqua_out	31.83	0
%Pompa_acqua_in	89.11875	73.83125

Tabella 10: valore max e min di ogni attributo del cluster 0

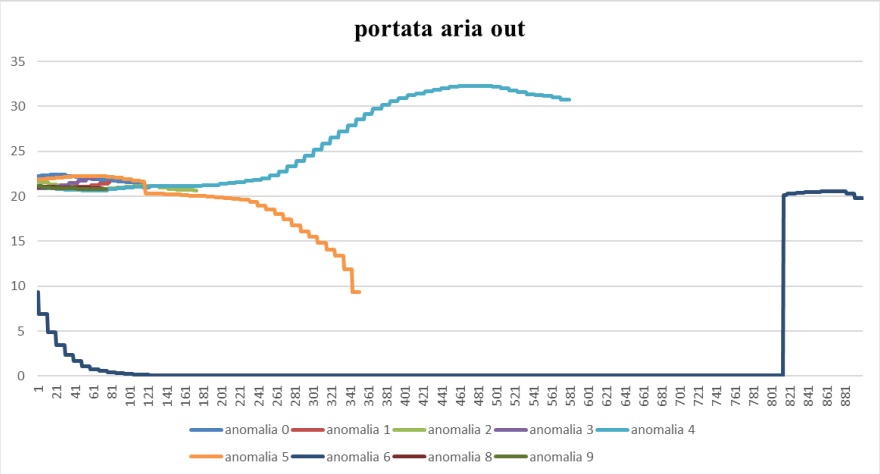


Figura 5.1: grafico attributo portata aria out delle anomalie presenti nel cluster 0

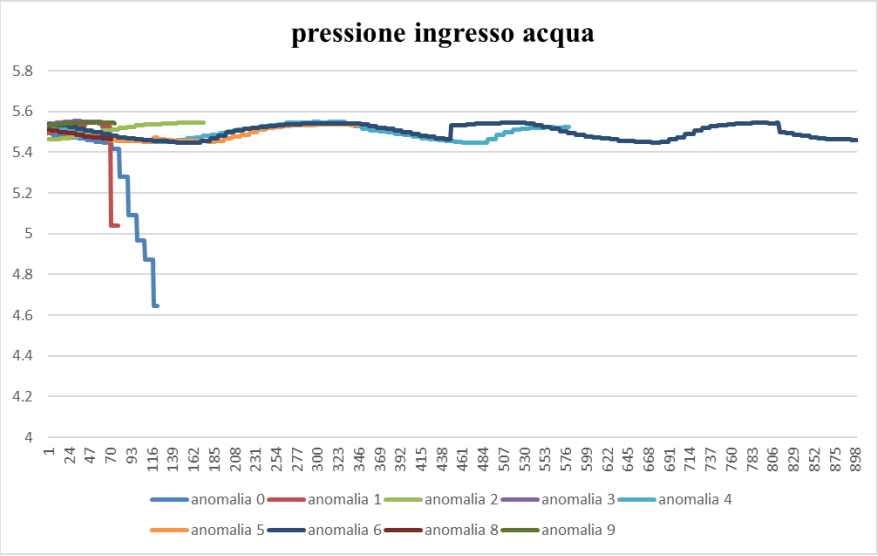


Figura 5.2: grafico attributo pressione ingresso acqua delle anomalie presenti nel cluster 0

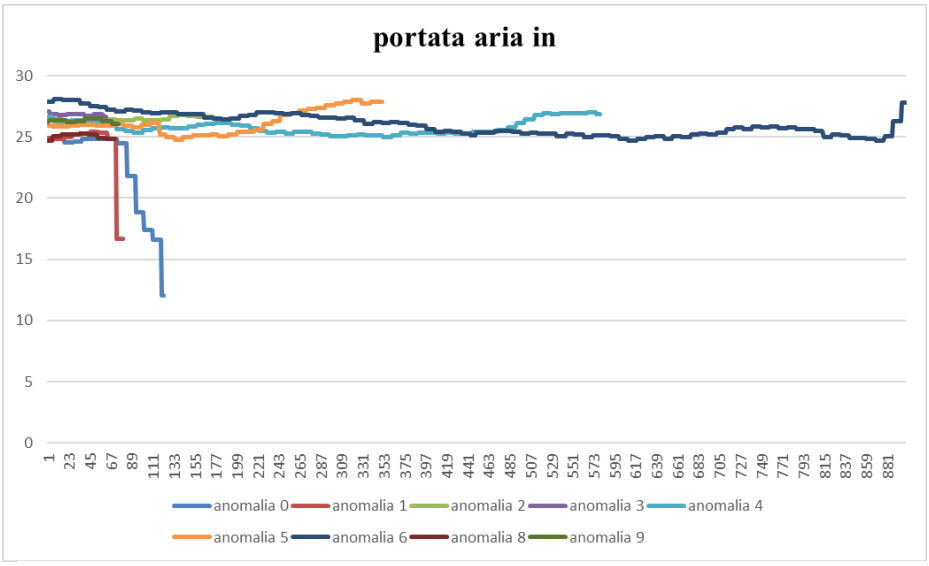


Figura 5.3: grafico attributo portata aria in delle anomalie presenti nel cluster 0

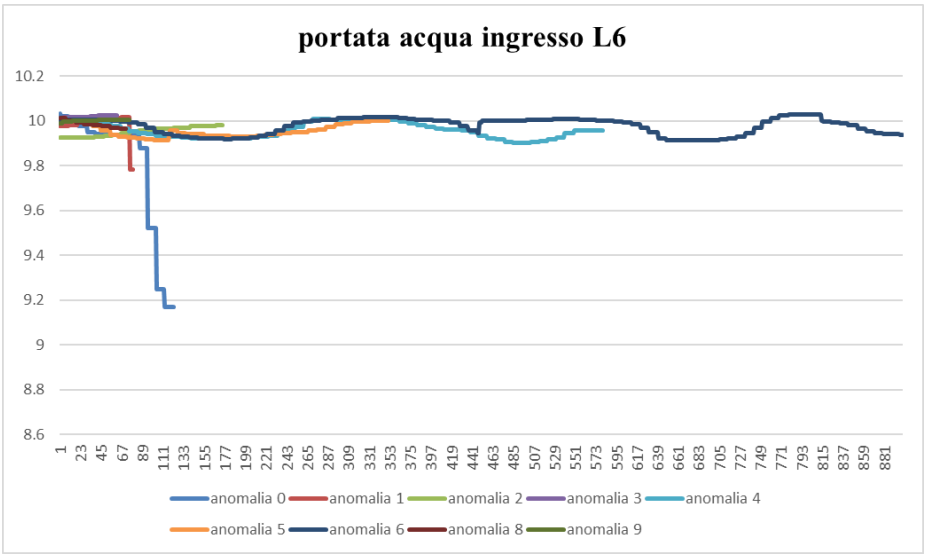


Figura 5.4: grafico attributo portata acqua ingresso L6 delle anomalie presenti nel cluster 0

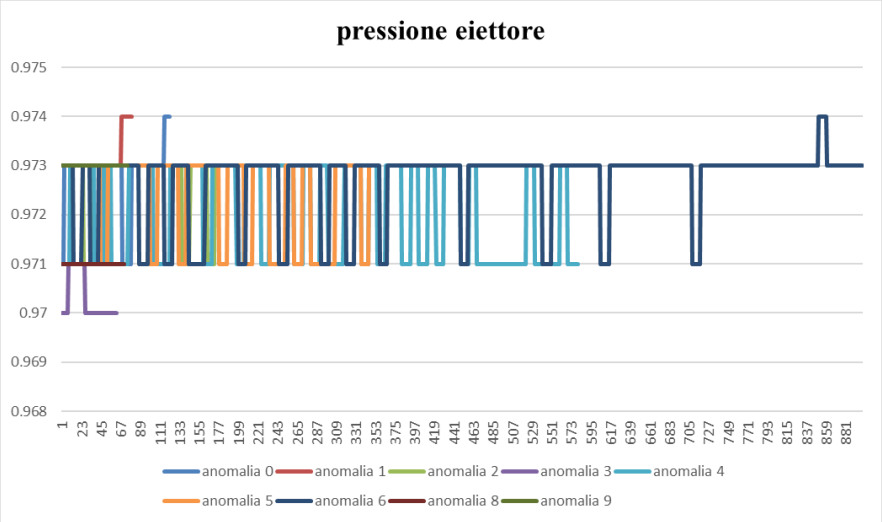


Figura 5.5: grafico attributo pressione eiettore delle anomalie presenti nel cluster 0

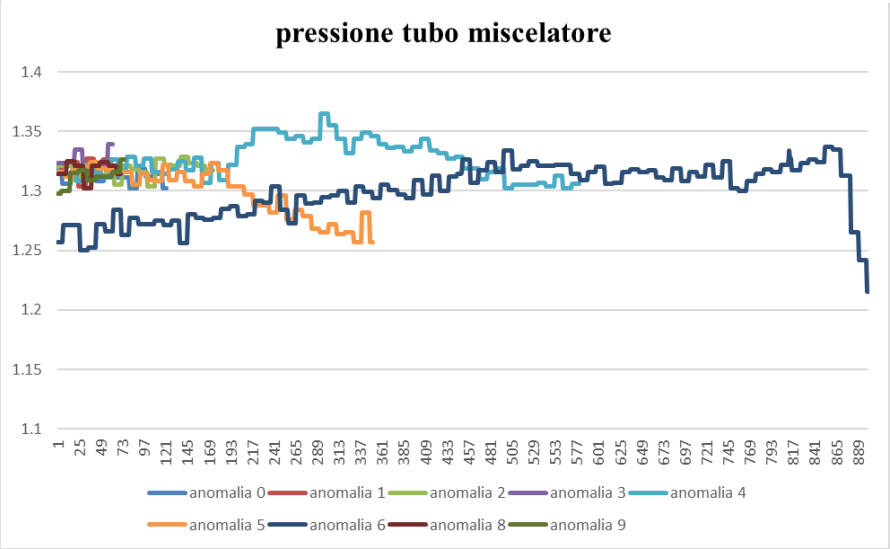


Figura 5.6: grafico attributo pressione tubo miscelatore delle anomalie presenti nel cluster 0

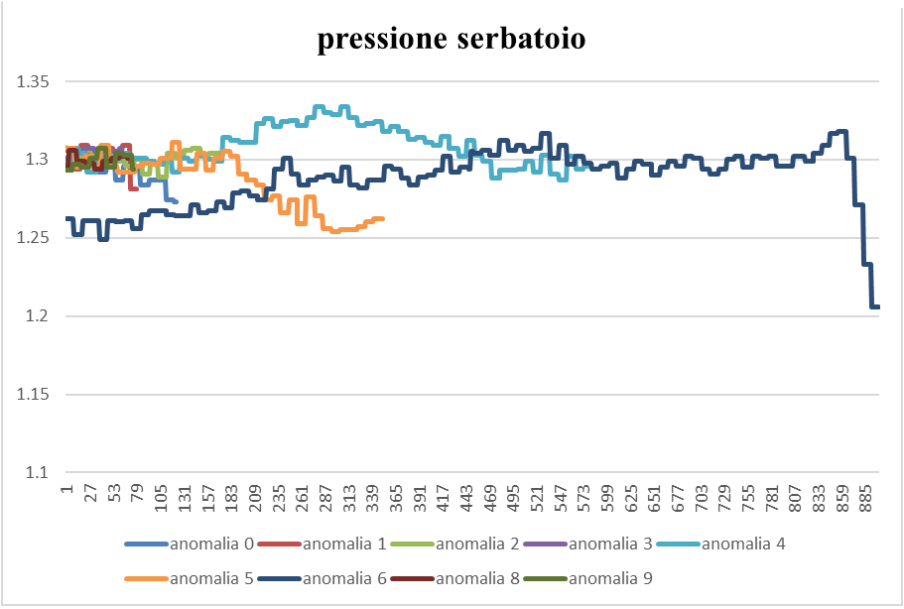


Figura 5.7: grafico attributo pressione serbatoio delle anomalie presenti nel cluster 0

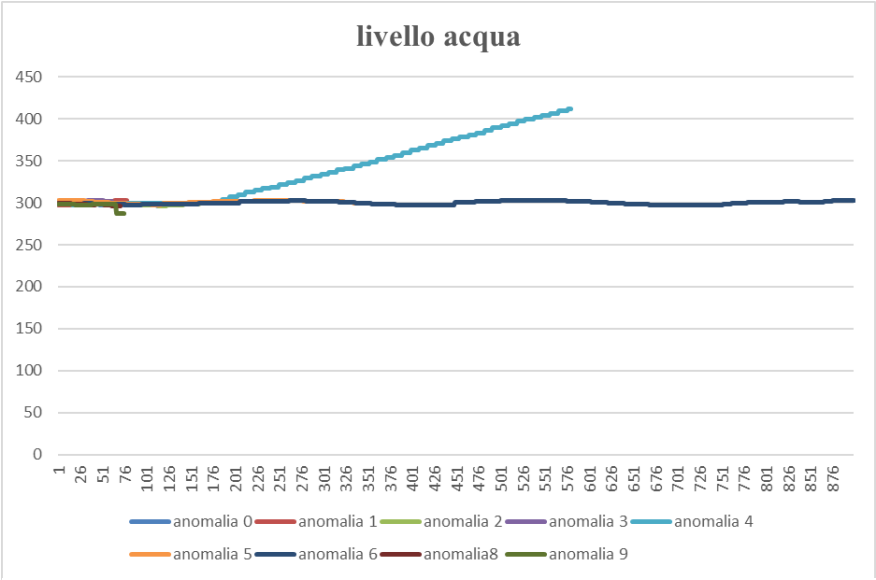


Figura 5.8: grafico attributo livello acqua delle anomalie presenti nel cluster 0

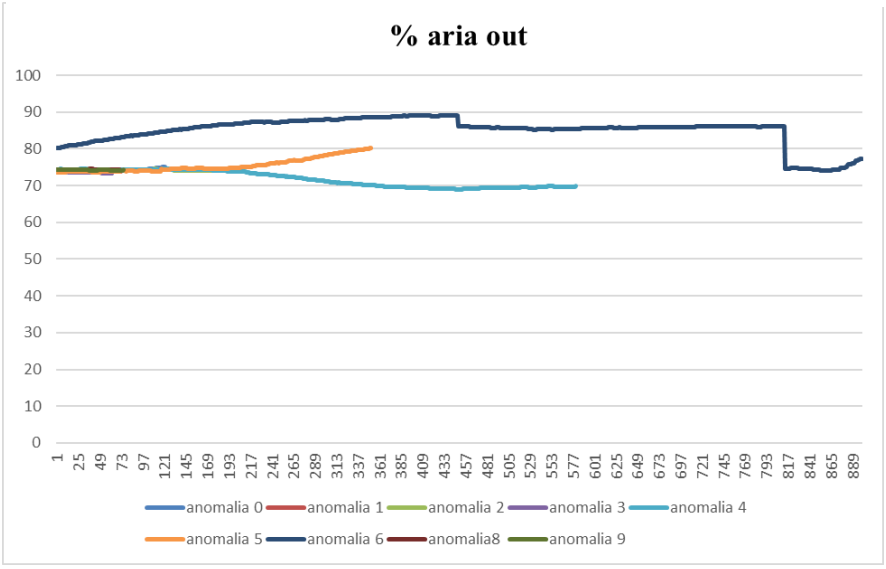


Figura 5.9: grafico attributo % aria out delle anomalie presenti nel cluster 0

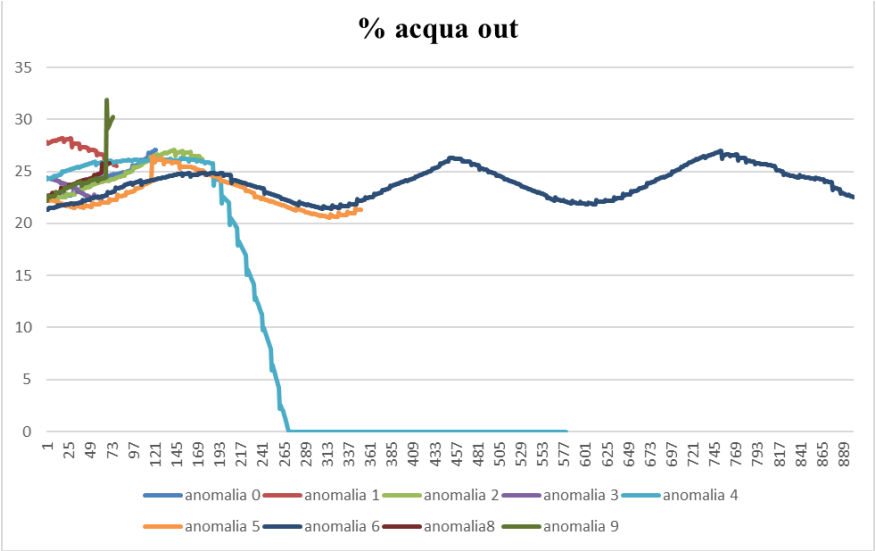


Figura 5.10: grafico attributo % acqua out delle anomalie presenti nel cluster 0

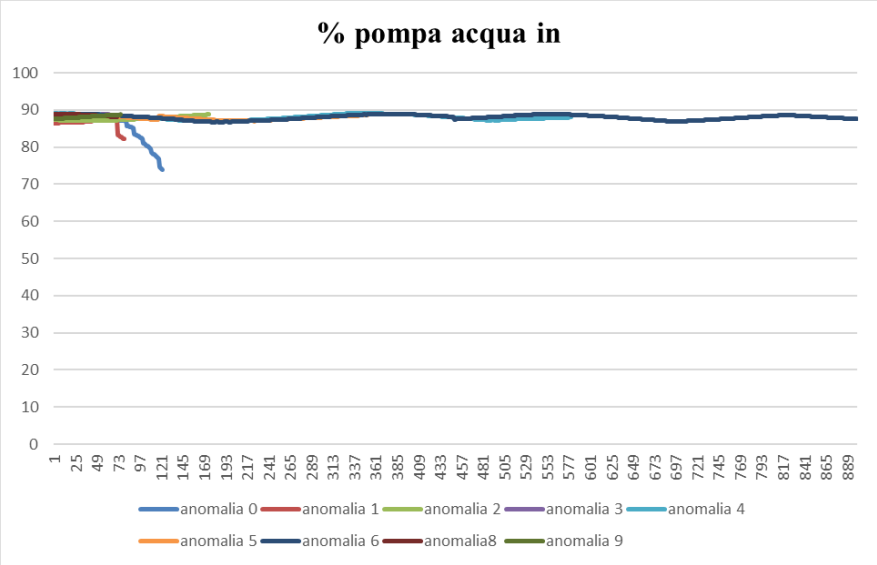


Figura 5.11: grafico attributo % pompa acqua in delle anomalie presenti nel cluster 0

5.1.3 CLUSTER 1

	valore max	*valore min*
portata_aria OUT	21.002	0.055
s7_pressione ingresso acqua	4.647	4.283
portata_aria IN	12.084	8.206
Portata_acqua ingresso L6	8.88	8.496
pressione_eiettore	0.976	0.974
pressione_tubo miscelatore	1.341	1.285
pressione_serbatoio	1.31	1.258
livello_acqua	304.168	294.65
%Aria_out	82.22	75.08
%Acqua_out	31.42	24.62
%Pompa_acqua_in	73.55625	0

Tabella 11: valore max e min di ogni attributo del cluster 0

Il cluster 1 contiene solamente anomalie di tipo 0.

5.1.4 CLUSTER 2

	valore max	*valore min*
portata_aria OUT	1.064	0.047
s7_pressione ingresso acqua	1.079	0.998
portata_aria IN	8.184	0.209
Portata_acqua ingresso L6	0.083	0.004
pressione_eiettore	0.982	0.976
pressione_tubo miscelatore	1.061	1.04
pressione_serbatoio	1.038	1.019
livello_acqua	263.939	259.401
%Aria_out	100	96.1
%Acqua_out	100	72.93
%Pompa_acqua_in	0	0

Tabella 12: valore max e min di ogni attributo del cluster 2

Il cluster 2 contiene solamente anomalie di tipo 1.

5.1.5 CLUSTER 3

	valore max	*valore min*
portata_aria OUT	20.744	0.117
s7_pressione ingresso acqua	5.531	5.447
portata_aria IN	5.347	-0.044
Portata_acqua ingresso L6	10.087	9.995
pressione_eiettore	0.982	0.974
pressione_tubo miscelatore	2.115	1.926
pressione_serbatoio	1.27	1.202
livello_acqua	306.943	288.261
%Aria_out	93.33	73.86
%Acqua_out	45.37	26.38
%Pompa_acqua_in	89	86.79375

Tabella 13: valore max e min di ogni attributo del cluster 3

Il cluster 3 contiene solamente anomalie di tipo 2.

5.1.6 CLUSTER 4

	valore max	*valore min*
portata_aria OUT	0.063	0.055
s7_pressione ingresso acqua	5.536	5.45
portata_aria IN	10.913	5.884
Portata_acqua ingresso L6	10.038	10.002
pressione_eiettore	0.994	0.991
pressione_tubo miscelatore	2.127	2.113
pressione_serbatoio	1.195	1.185
livello_acqua	295.884	290.156
%Aria_out	100	100
%Acqua_out	37.43	26.6
%Pompa_acqua_in	87.58125	86.83125

Tabella 14: valore max e min di ogni attributo del cluster 4

Il cluster 4 contiene solamente anomalie di tipo 2.

5.1.7 CLUSTER 5

	valore max	*valore min*
portata_aria OUT	3.177	0.07
s7_pressione ingresso acqua	5.531	5.367
portata_aria IN	44.859	33.297
Portata_acqua ingresso L6	9.73	9.606
pressione_eiettore	1.209	1.175
pressione_tubo miscelatore	2.681	2.635
pressione_serbatoio	1.168	1.157
livello_acqua	265.349	260.855
%Aria_out	100	85.84
%Acqua_out	100	65.39
%Pompa_acqua_in	88.51875	87.99375

Tabella 15: valore max e min di ogni attributo del cluster 5

Il cluster 5 contiene solamente anomalie di tipo 3.

5.1.8 CLUSTER 6

	valore max	*valore min*
portata_aria OUT	41.668	32.951
s7_pressione ingresso acqua	5.543	5.441
portata_aria IN	27.65	24.366
Portata_acqua ingresso L6	10	9.905
pressione_eiettore	0.973	0.971
pressione_tubo miscelatore	1.363	1.284
pressione_serbatoio	1.334	1.274
livello_acqua	599.471	399.43
%Aria_out	68.59	66.05
%Acqua_out	0	0
%Pompa_acqua_in	88.975	86.93125

Tabella 16: valore max e min di ogni attributo del cluster 6

Il cluster 6 contiene solamente anomalie di tipo 5.

5.1.9 CLUSTER 7

Il cluster 7 contiene 2 anomalie, la anomalia 6 e 7.

istanze anomalia 6	561
istanze anomalia 7	754

	valore max	*valore min*
portata_aria OUT	7.911	0.063
s7_pressione ingresso acqua	5.552	5.446
portata_aria IN	43.672	33.691
Portata_acqua ingresso L6	10.042	9.919
pressione_eiettore	0.971	0.969
pressione_tubo miscelatore	1.137	1.083
pressione_serbatoio	1.128	1.056
livello_acqua	310.38	293.945
%Aria_out	100	84.05
%Acqua_out	21.48	0
%Pompa_acqua_in	88.8875	86.71875

Tabella 17: valore max e min di ogni attributo del cluster 7

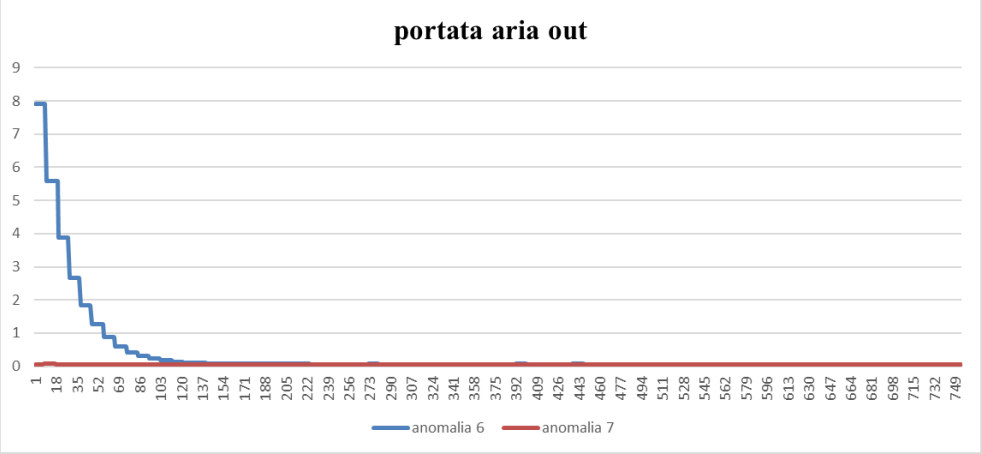


Figura 5.14: grafico attributo portata aria out delle anomalie presenti nel cluster 7

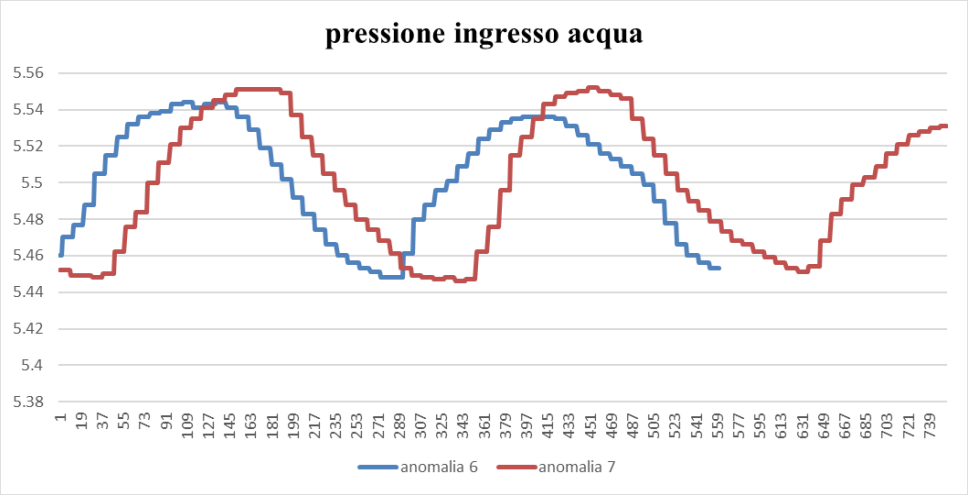


Figura 5.15: grafico attributo pressione ingresso acqua delle anomalie presenti nel cluster 7

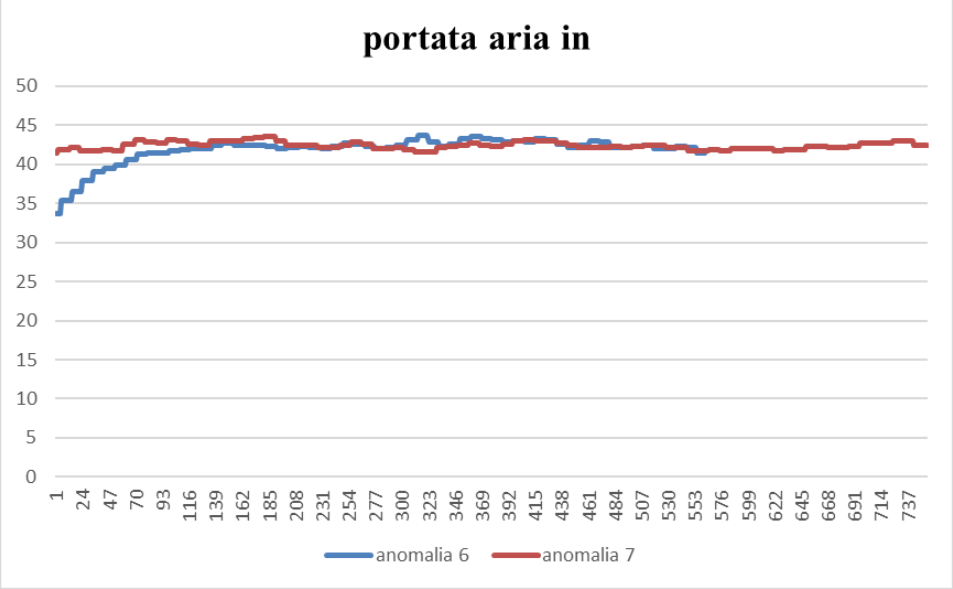


Figura 5.16: grafico attributo portata aria in delle anomalie presenti nel cluster 7

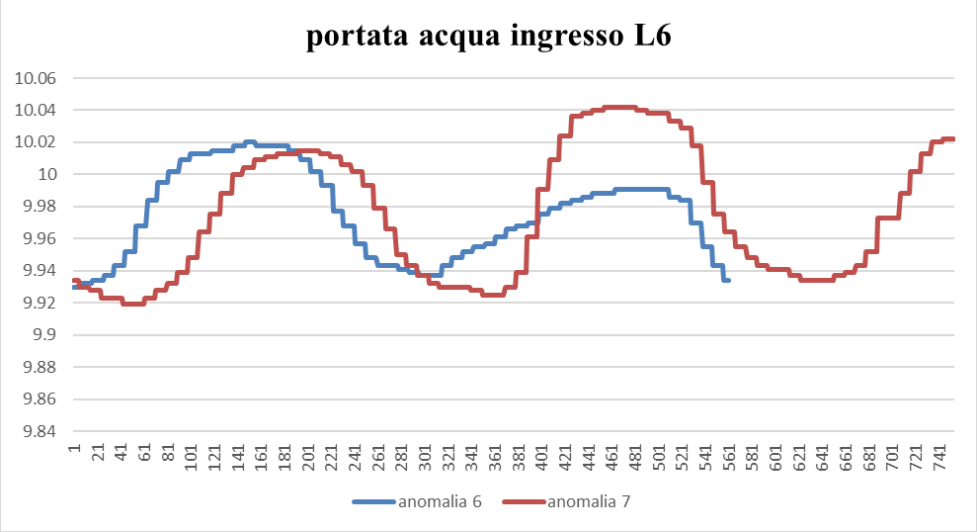


Figura 5.17: grafico attributo portata acqua ingresso L6 delle anomalie presenti nel cluster 7

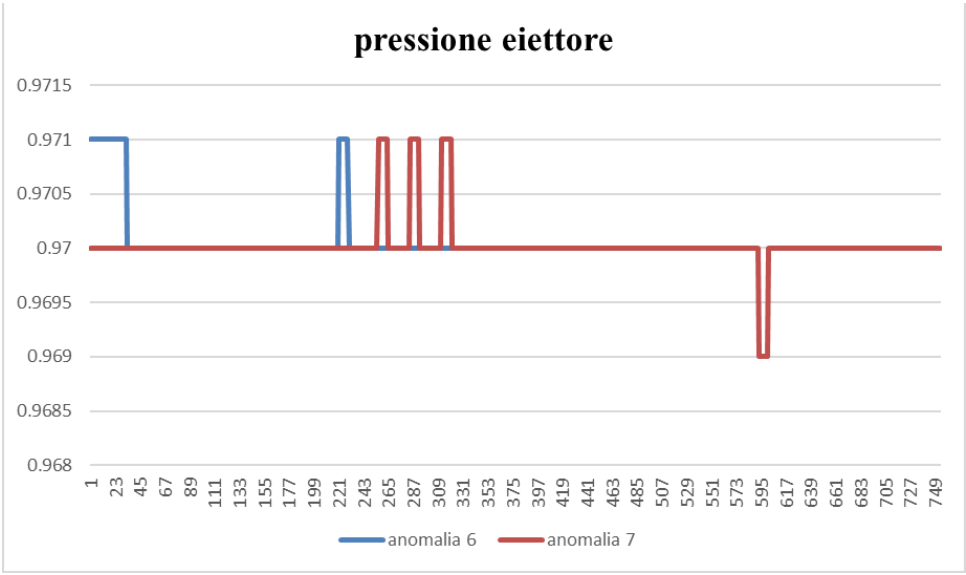


Figura 5.18: grafico attributo pressione eiettore delle anomalie presenti nel cluster 7

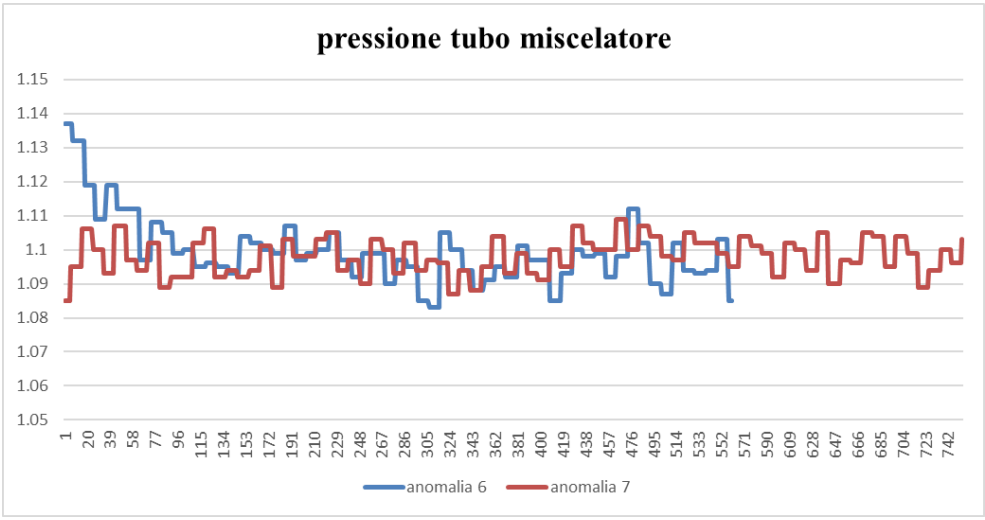


Figura 5.19: grafico attributo pressione tubo miscelatore delle anomalie presenti nel cluster 7

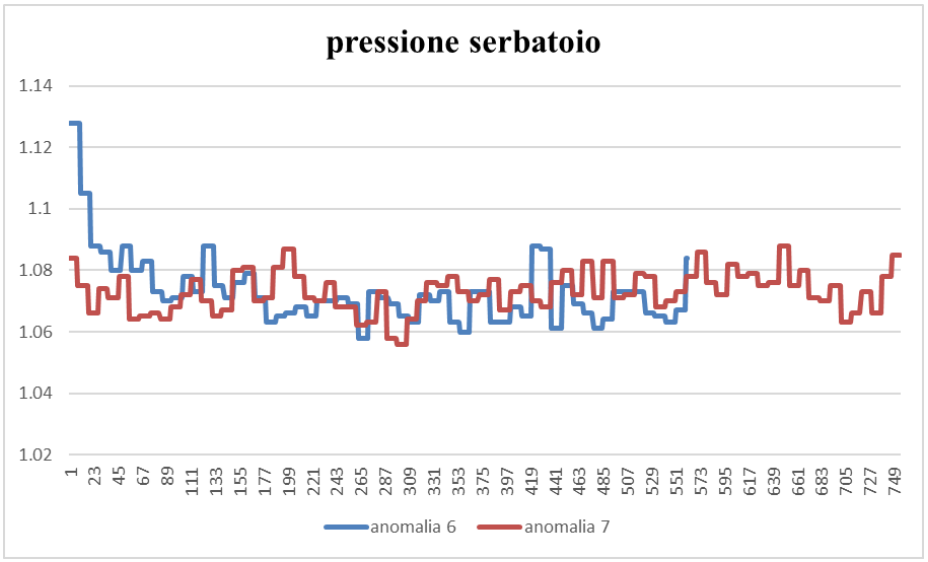


Figura 5.20: grafico attributo pressione serbatoio delle anomalie presenti nel cluster 7

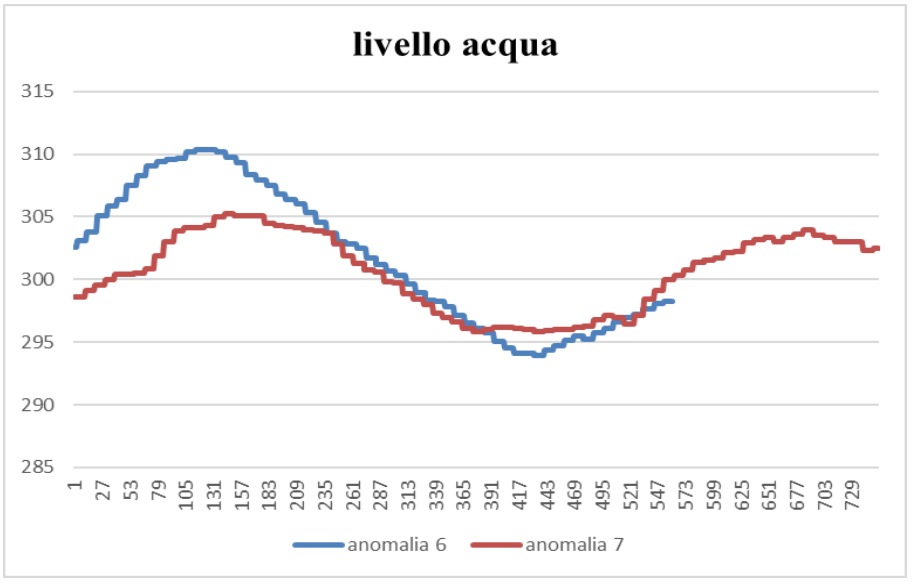


Figura 5.21: grafico attributo livello acqua delle anomalie presenti nel cluster 7

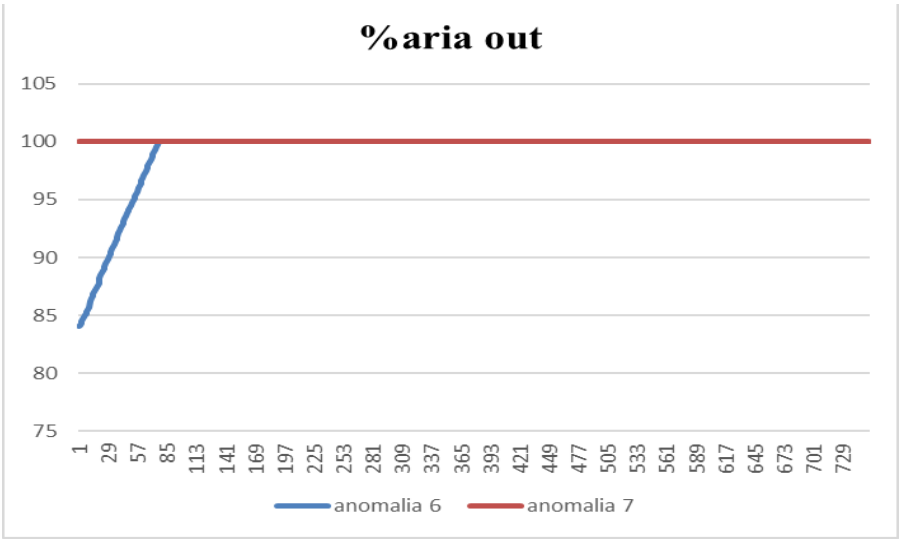


Figura 5.22: grafico attributo % aria out delle anomalie presenti nel cluster 7

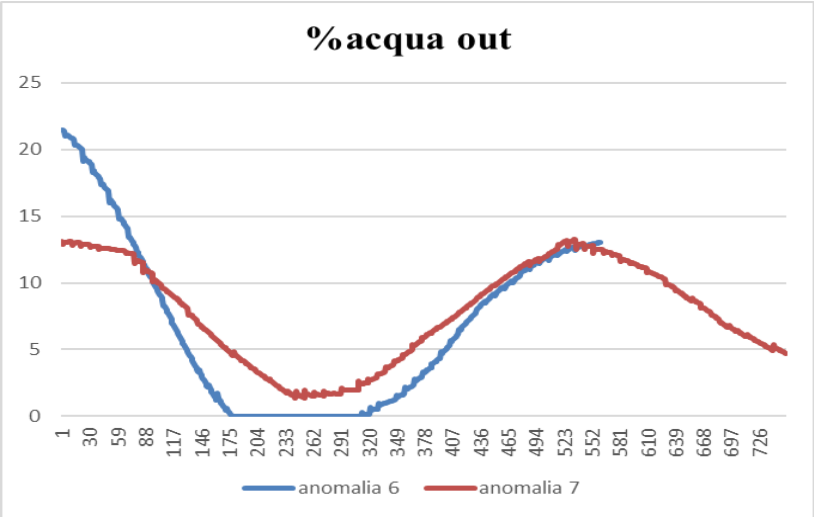


Figura 5.23: grafico attributo % acqua out delle anomalie presenti nel cluster 7

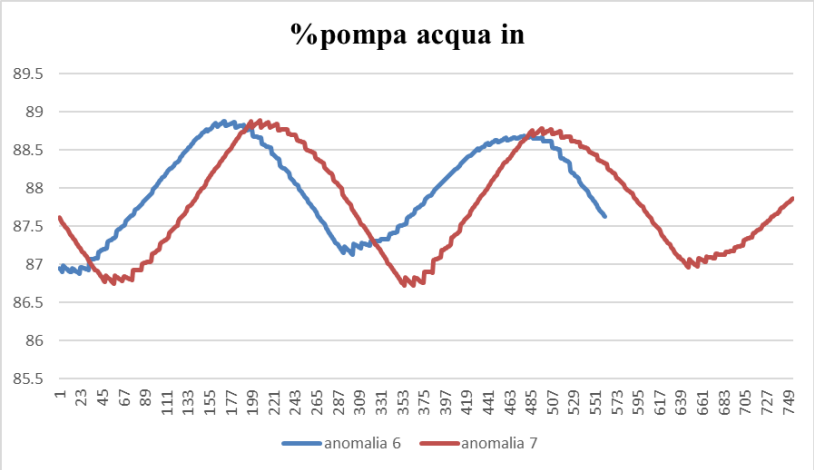


Figura 5.24: grafico attributo % pompa acqua in delle anomalie presenti nel cluster 7

5.1.10 CLUSTER 8

	valore max	*valore min*
portata_aria OUT	28.976	20.189
s7_pressione ingresso acqua	5.541	5.04
portata_aria IN	27.113	18.666
Portata_acqua ingresso L6	10.006	9.453
pressione_eiettore	0.973	0.971
pressione_tubo miscelatore	1.361	1.299
pressione_serbatoio	1.331	1.285
livello_acqua	283.723	110.339
%Aria_out	74.99	70.11
%Acqua_out	100	96.54
%Pompa_acqua_in	88.83125	44.6

Tabella 18: valore max e min di ogni attributo del cluster 8

Il cluster 8 contiene solamente l'anomalia di tipo 8.

5.1.11 CLUSTER 9

	valore max	*valore min*
portata_aria OUT	2.363	0.282
s7_pressione ingresso acqua	5.274	4.674
portata_aria IN	28.759	23.275
Portata_acqua ingresso L6	9.674	9.148
pressione_eiettore	0.976	0.973
pressione_tubo miscelatore	1.298	1.213
pressione_serbatoio	1.27	1.209
livello_acqua	289.892	257.242
%Aria_out	86.51	83.31
%Acqua_out	40.62	5.57
%Pompa_acqua_in	8.18125	0

Tabella 19: valore max e min di ogni attributo del cluster 9

Il cluster 9 contiene solamente l'anomalia 9.

5.1.12 CLUSTER 10

	valore max	*valore min*
portata_aria OUT	0.282	0.063
s7_pressione ingresso acqua	5.274	4.389
portata_aria IN	22.413	9.756
Portata_acqua ingresso L6	9.609	9.078
pressione_eiettore	0.976	0.973
pressione_tubo miscelatore	1.445	1.259
pressione_serbatoio	1.43	1.244
livello_acqua	254.113	203.442
%Aria_out	86.68	73.72
%Acqua_out	100	45.51
%Pompa_acqua_in	0	0

Tabella 20: valore max e min di ogni attributo del cluster 10

Il cluster 10 contiene solamente l'anomalia di tipo 8.

5.1.13 CLUSTER 11, 12, 13, 14, 15

I cluster 11, 12, 13, 14, 15 contengono tutti l'anomalia 9.

	valore max	*valore min*
portata_aria OUT	28.342	20.431
s7_pressione ingresso acqua	5.535	5.004
portata_aria IN	26.591	18.531
Portata_acqua ingresso L6	10.006	9.366
pressione_eiettore	0.974	0.971
pressione_tubo miscelatore	1.364	1.306
pressione_serbatoio	1.324	1.287
livello_acqua	160.966	48.828
%Aria_out	74.67	70.14
%Acqua_out	100	100
%Pompa_acqua_in	88.90625	38.15

Tabella 21: valore max e min di ogni attributo del cluster 11

	valore max	*valore min*
portata_aria OUT	27.043	25.392
s7_pressione ingresso acqua	5.271	4.892
portata_aria IN	25.928	16.663
Portata_acqua ingresso L6	9.615	9.501
pressione_eiettore	0.976	0.971
pressione_tubo miscelatore	1.351	1.306
pressione_serbatoio	1.322	1.287
livello_acqua	306.899	225.781
%Aria_out	71.67	70.79
%Acqua_out	100	95.1
%Pompa_acqua_in	2.45	0

Tabella 22: valore max e min di ogni attributo del cluster 12

	valore max	*valore min*
portata_aria OUT	0.133	0.055
s7_pressione ingresso acqua	3.716	2.761
portata_aria IN	5.566	0.428
Portata_acqua ingresso L6	8.235	6.752
pressione_eiettore	0.977	0.976
pressione_tubo miscelatore	1.339	1.219
pressione_serbatoio	1.294	1.217
livello_acqua	254.289	222.917
%Aria_out	96.42	89.38
%Acqua_out	100	49.33
%Pompa_acqua_in	0	0

Tabella 23: valore max e min di ogni attributo del cluster 13

	valore max	*valore min*
portata_aria OUT	0.063	0.063
s7_pressione ingresso acqua	4.728	4.4
portata_aria IN	21.538	17.069
Portata_acqua ingresso L6	8.089	7.087
pressione_eiettore	0.974	0.974
pressione_tubo miscelatore	1.396	1.351
pressione_serbatoio	1.359	1.326
livello_acqua	205.998	182.777
%Aria_out	95.46	93.5
%Acqua_out	100	100
%Pompa_acqua_in	0	0

Tabella 24: valore max e min di ogni attributo del cluster 14

	valore max	*valore min*
portata_aria OUT	0.063	0.055
s7_pressione ingresso acqua	5.26	5.001
portata_aria IN	20.544	15.2
Portata_acqua ingresso L6	9.532	8.575
pressione_eiettore	0.976	0.973
pressione_tubo miscelatore	1.506	1.396
pressione_serbatoio	1.484	1.392
livello_acqua	173.568	118.314
%Aria_out	92.36	74.32
%Acqua_out	100	100
%Pompa_acqua_in	0	0

Tabella 25: valore max e min di ogni attributo del cluster 15

5.2 DESCRIZIONE ANOMALIE

anomalia	Cluster		Portata aria out	Pressione ingresso acqua	Portata aria in	Portata acqua ingresso L6	Pressione eiettore	Pressione tubo miscelatore	Pressione serbatoio	Livello acqua	% aria out	% acqua out	% pompa acqua in
0	0	Valore max	22.387	5.492	24.853	10.031	0.974	1.319	1.3	300.202	75.06	27.08	88.66
		Valore min	21.002	4.647	12.084	9.168	0.971	1.302	1.273	296.016	73.67	22.33	73.83
	1	Valore max	21.002	4.647	12.084	8.88	0.976	1.341	1.31	304.168	82.22	31.42	73.56
		Valore min	0.055	4.283	8.206	8.496	0.974	1.285	1.258	294.65	75.08	24.62	0
1	0	Valore max	21.519	5.55	25.409	10.015	0.974	1.327	1.309	302.846	74.32	28.22	87.41
		Valore min	20.901	5.041	16.644	9.782	0.973	1.304	1.281	297.558	73.73	25.56	82.2
	2	Valore max	1.064	1.079	8.184	0.083	0.982	1.061	1.038	263.939	100	100	0
		Valore min	0.047	0.998	0.209	0.004	0.976	1.04	1.019	259.401	96.1	72.93	0
2	0	Valore max	21.628	5.55	26.775	9.982	0.973	1.339	1.307	300.9	74.48	27.09	88.76
		Valore min	20.65	5.041	25.944	9.925	0.971	1.316	1.3	296.5	73.97	22.25	82.2

anomalia	Cluster		Portata aria out	Pressione ingresso acqua	Portata aria in	Portata acqua ingresso L6	Pressione eiettore	Pressione tubo miscelatore	Pressione serbatoio	Livello acqua	% aria out	% acqua out	% pompa acqua in
2	3	Valore min	20.744	5.531	5.347	10.087	0.982	2.115	1.27	306.943	93.33	43.37	89
		Valore max	0.117	5.447	0.05	9.995	0.974	1.926	1.202	288.261	73.86	26.38	86.8
	4	Valore min	0.063	5.536	10.913	10.038	0.994	2.127	1.195	295.884	100	37.43	87.6
		Valore max	0.055	5.45	5.884	10.002	0.991	2.113	1.185	290.156	100	26.6	86.8
3	0	Valore min	21.988	5.552	27.1	10.024	0.971	1.339	1.307	302.8	73.74	24.38	88.9
		Valore max	20.908	5.539	26.669	9.993	0.97	1.316	1.3	301.876	73.35	22.34	87.9
	5	Valore min	3.177	5.531	44.86	9.73	1.2	2.681	1.168	265.349	100	100	88.52
		Valore max	0.07	5.367	33.3	9.6	1.175	2.635	1.157	260.855	85.84	65.39	88
5	0	Valore min	22.239	5.536	27.988	10.004	0.973	1.324	1.311	303.15	80.15	26.4	88.931
		Valore max	9.304	5.45	24.747	9.912	0.971	1.257	1.254	297.43	73.54	20.56	87.01

anomalia	Cluster		Portata aria out	Pressione ingresso acqua	Portata aria in	Portata acqua ingresso L6	Pressione eiettore	Pressione tubo miscelatore	Pressione serbatoio	Livello acqua	% aria out	% acqua out	% pompa acqua in
5	6	Valore min	41.67	5.543	27.65	10	0.973	1.363	1.334	599.47	68.6	0	88.975
		Valore max	32.95	5.441	24.366	9.905	0.971	1.284	1.274	399.4	66.05	0	86.931
6	0	Valore min	20.533	5.545	28.084	10.029	0.974	1.337	1.318	302.8	89.17	26.98	88.87
		Valore max	0.063	5.445	24.68	9.912	0.971	1.215	1.206	296.8	73.96	21.33	86.7
	7	Valore min	7.911	5.54	43.67	10.02	0.971	1.137	1.128	310.4	100	21.48	88.88
		Valore max	0.063	5.45	33.69	9.93	0.97	1.083	1.058	293.9	84.05	0	86.88
8	8	Valore min	21.073	5.511	25.278	10.011	0.971	1.325	1.306	299.409	74.47	25.82	88.88
		Valore max	20.893	5.462	24.725	9.966	0.971	1.302	1.294	296.192	74.2	22.64	88.125
	10	Valore min	28.976	5.537	27.113	9.997	0.973	1.361	1.331	283.723	74.99	100	88.15
		Valore max	20.189	5.04	18.666	9.453	0.971	1.255	1.285	125.805	70.11	96.54	44.6

anomalia	Cluster		Portata aria out	Pressione ingresso acqua	Portata aria in	Portata acqua ingresso L6	Pressione eiettore	Pressione tubo miscelatore	Pressione serbatoio	Livello acqua	% aria out	% acqua out	% pompa acqua in
9	9	Valore max	1.784	5.052	27.294	9.523	0.974	1.278	1.234	289.892	86.51	22.09	0
		Valore min	0.423	4.674	23.275	9.148	0.973	1.213	1.209	264.776	83.74	5.57	0
	11	Valore max	28.342	5.535	26.591	10.006	0.974	1.364	1.324	160.966	74.67	100	88.906
		Valore min	20.431	5.004	18.531	9.366	0.971	1.306	1.287	48.828	70.14	100	38.15
	12	Valore max	27.043	5.271	25.928	9.615	0.976	1.339	1.322	306.899	71.67	100	2.45
		Valore min	25.392	4.892	16.663	9.501	0.971	1.219	1.287	225.781	70.79	95.1	0
	13	Valore max	0.133	3.716	5.566	8.235	0.977	1.396	1.294	254.289	96.42	100	0
		Valore min	0.055	2.761	0.428	6.752	0.976	1.351	1.217	222.917	89.38	49.33	0
	14	Valore max	0.063	4.728	21.538	8.089	0.974	1.506	1.359	205.998	95.42	100	0
		Valore min	0.063	4.4	17.069	7.087	0.974	1.396	1.326	182.777	93.5	100	0
	15	Valore max	0.063	5.26	20.544	9.532	0.976	1.351	1.484	173.568	92.36	100	0
		Valore min	0.055	5.001	15.2	8.575	0.973	1.306	1.392	118.314	74.32	100	0

Tabella 26: descrizione di ogni anomalia che è presente in più cluster

Nella tabella 26 sono state descritte solamente le anomalie che erano suddivise su più di un cluster. È necessario notare che l'anomalia 4 è totalmente contenuta nel cluster 0 e l'anomalia 7 nel cluster 7 quindi per avere una caratterizzazione della anomalia 4 e 7 basta vedere il cluster 4 e 7.

5.3 ASSEGNAZIONE CLUSTER AL DATASET DI TEST

Come già spiegato nel paragrafo 4.4, il dataset originale è stato suddiviso il training set e test set (75% training e 15% test). Una volta suddiviso il dataset di training in cluster e sulla base delle considerazioni fatte sono stati calcolati i centroidi di tutti i cluster e, per ogni lettura:

- È stata calcolata la distanza euclidea della lettura da ogni cluster;
- È stato assegnato il cluster con distanza euclidea minima;
- È stato verificato se quello che già era noto dalla descrizione dei cluster e delle anomalie del training set era conforme con il cluster assegnato alla lettura del test set.

I risultati che sono stati ottenuti sono:

- su 1758 letture totali, le assegnazioni corrette delle letture ai cluster sono state 1197;
- l'accuratezza ottenuta è pari a 0.681.

6. CONCLUSIONI

Tutti i dati che sono stati utilizzati sono provenienti da una macchina reale e non da una simulazione. I sensori posizionati sulla macchina hanno raccolto grandezze fisiche durante il funzionamento della macchina, che è stata effettivamente portata in stati di guasto differenti.

I risultati ottenuti dalla identificazione delle anomalie sono piuttosto soddisfacenti in quanto gli algoritmi utilizzati riescono ad avere prestazioni elevate riuscendo a distinguere le varie tipologie di anomalie. Ciò consente di poter effettuare operazioni di diagnostica dello stato dell'impianto. Un possibile sviluppo futuro potrebbe essere quello di individuare con buona precisione anche il grado di avanzamento di uno stesso guasto in modo tale da poter effettuare anche operazioni di prognostica, cioè aggiungere la possibilità di associare a ogni input anche una stima del tempo rimanente prima di entrare in uno stato di guasto.

Per quanto riguarda la classificazione delle anomalie, l'algoritmo DBSCAN ha ottenuto discreti risultati. Sicuramente un possibile scenario futuro potrebbe essere quello di utilizzare algoritmi diversi per la classificazione delle anomalie facendo un set opportuno dei parametri sul dataset dell'impianto.

Potrebbe anche essere interessante raccogliere dati in tempo reale in modo tale da poter identificare e classificare eventuali anomalie e riportare i dati in maniera aggregata implementando anche strumenti di visualizzazione e analisi dei dati.

Bibliografia

[1] https://it.mathworks.com/discovery/machine-learning.html.

[2] «What is machine learning and types of machine learning — Part-1,» https://towardsdatascience.com/what-is-machine-learning-and-types-of-machine-learning-andrews-machine-learning-part-1-9cd9755bc647.

[3] https://towardsdatascience.com/workflow-of-a-machine-learning-project-ec1dba419b94.

[4] A. Sari, «A Review of Anomaly Detection Systems in Cloud Networks and Survey of Cloud Security Measures in Cloud Storage Applications».

[5] M. K. J. P. Jiawei Han, Data Mining Concepts and techniques.

[6] E. C. M. B. Giovanni Mazzuto, «Dataset of an experimental multiphase water-air plant».

[7] https://pycaret.readthedocs.io/en/latest/api/anomaly.html.

[8] https://towardsdatascience.com/introduction-to-anomaly-detection-in-python-with-pycaret-2fecd7144f87.