



Università Politecnica delle Marche

Laurea Magistrale

Manutenzione preventiva per la robotica e l'automazione intelligente

PHM Asia Pacific 2023
Data challenge

Dediu Razvan Alexandru

A.A. 2023/24

Contents

1	Introduzione	2
1.1	Scenario	2
1.2	Task	3
2	Data exploration	4
3	Feature Engineering	7
4	Task 1	8
5	Task 2	10
6	Task 3	14
7	Task 4	18
8	Task 5	21
9	Risultati	23
9.1	Task 1	23
9.2	Task 2	24
9.3	Task 3	26
9.4	Task 4	26
9.5	Task 5	27

1 Introduzione

1.1 Scenario

Gli scenari sperimentali prevedono un sistema di propulsione in cui il fluido di lavoro è acqua pressurizzata a 2 MPa e viene espulsa attraverso quattro elettrovalvole (SV1 – SV4), simulando i propulsori. P1 – P8 rappresentano i sensori di pressione e i dati temporali sono acquisiti con una frequenza di campionamento di 1 kHz per un intervallo di tempo di 0-1200 ms.

L'apertura e la chiusura delle elettrovalvole causano fluttuazioni di pressione dovute al fenomeno dell'acqua martello, seguite da modalità acustiche all'interno del sistema di propulsione. Un esempio tipico di dati temporali è mostrato nella Fig. 2. La elettrovalvola si apre a 100 ms e si chiude a 300 ms. Per tenere conto delle differenze individuali delle elettrovalvole presenti nell'equipaggiamento reale, il movimento della elettrovalvola ha una variabilità di 1 ms.

I tempi di apertura e chiusura rimangono a 400 ms, anche con la variabilità (ad esempio, apertura per 99,7 ms e chiusura per 300,3 ms). Questa sequenza viene eseguita tre volte consecutivamente, risultando in una misura totale di 1200 ms.

Si considerano anomalie dovute alla contaminazione da bolle e guasti dovuti all'apertura

anomala delle elettrovalvole. Inoltre, è inclusa nei dati di test anche un'anomalia sconosciuta.

Anomalia da bolle: Durante l'operazione reale di una navetta spaziale, possono occasionalmente comparire bolle d'aria nei tubi. La presenza di bolle modifica la velocità del suono, causando leggere variazioni nelle fluttuazioni di pressione. È auspicabile rilevare la comparsa di bolle e la loro posizione. Ci sono otto posizioni possibili: BV1 e BP1 a BP7. La quantità di bolle contaminate nel sistema di propulsione è costante in tutti i casi per semplicità.

Guasti delle elettrovalvole: Questo è uno dei principali modi di guasto nei sistemi di propulsione spaziale. È necessario determinare quali elettrovalvole hanno subito il guasto e il loro rapporto di apertura. Le elettrovalvole si aprono e si chiudono con un rapporto di apertura del 100% e dello 0%, rispettivamente. In caso di guasti, le elettrovalvole si aprono a un grado compreso tra lo 0% e il 100%, il che comporta una riduzione del volume di fluido attraverso l'elettrovalvola.

Anomalia sconosciuta: Durante l'operazione pratica, possono verificarsi anomalie o guasti completamente imprevisti e sconosciuti. È necessario distinguere queste anomalie sconosciute senza confonderle con anomalie e guasti noti. Alcune anomalie o guasti sconosciuti sono mescolati nei dati di test. Identificarli è anch'esso parte del compito in questa competizione.

Differenze individuali nei veicoli spaziali: Poiché le elettrovalvole presentano differenze individuali, come i tempi di apertura e chiusura, i dati temporali acquisiti dal sistema di propulsione spaziale mostrano variazioni, che a loro volta portano a differenze tra i veicoli spaziali. In questa competizione, si considerano quattro veicoli spaziali (No.1 a No.4). I risultati per i veicoli No.1 a No.3 sono inclusi nei dati di addestramento, mentre i dati di test sono composti dai risultati per i veicoli No.1 e No.4.

1.2 Task

- Determinare se tutti i dati di test sono normali o anomali.
- Per i dati rilevati come anomali, determinare se l'anomalia è dovuta a contaminazione da bolle, guasto delle valvole, o guasto sconosciuto.
- Per i dati identificati come contaminazione da bolle, determinare la posizione della bolla tra otto posizioni: BV1, e da BP1 a BP7.
- Per i dati identificati come guasti alle elettrovalvole, determinare quale delle quattro elettrovalvole (SV1 a SV4) è guasta.
- Per le elettrovalvole identificate come guaste, prevedere il rapporto di apertura ($0\% \leq \text{Rapporto di apertura} < 100\%$).

2 Data exploration

Ogni caso presente nel dataset è stato analizzato sia nel dominio del tempo che in quello delle frequenze.

Nel dominio del tempo, sono stati esaminati i campioni provenienti da diversi scenari: normali, anomali e di guasto. Questa analisi ha permesso di osservare le variazioni nel comportamento temporale dei segnali, con particolare attenzione ai cambiamenti nell'apertura delle elettrovalvole.

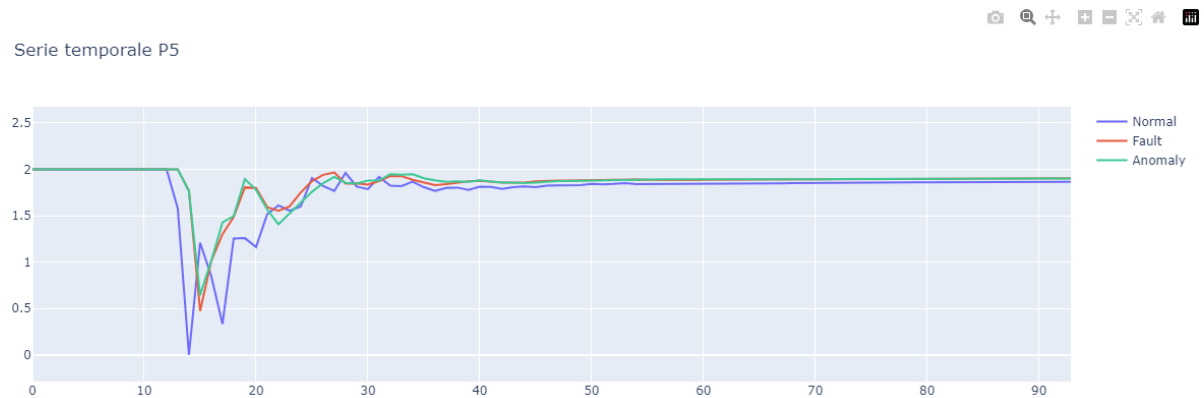


Figure 1: Fase di apertura dei 3 diversi casi nel dominio del tempo per sensore P5

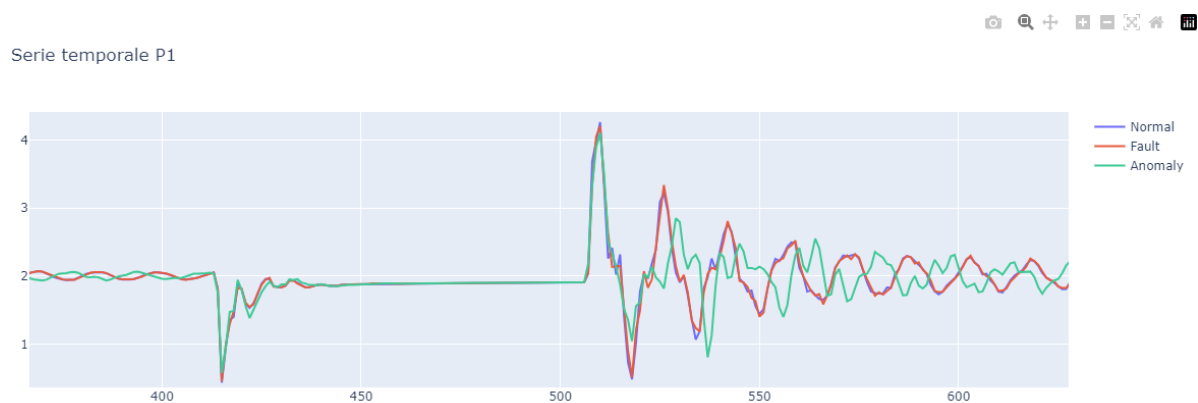


Figure 2: Fasi di apertura e chiusura per sensore P1

Per ogni scenario, sono stati calcolati due parametri sulla intera serie temporale (0-1200): la deviazione standard e l'area sottola curva che rappresenta il segnale (metodo dei trapezi). In particolare, l'analisi della deviazione standard ha rivelato un cambiamento significativo a

partire dal caso 104, che è l'ultimo caso normale presente nel dataset. Questo cambiamento è stato immediatamente evidente e ha fornito indicazioni utili sulla transizione dai casi normali ai casi anomali o di guasto. Di seguito la deviazione standard per i sensori P1 e P5.

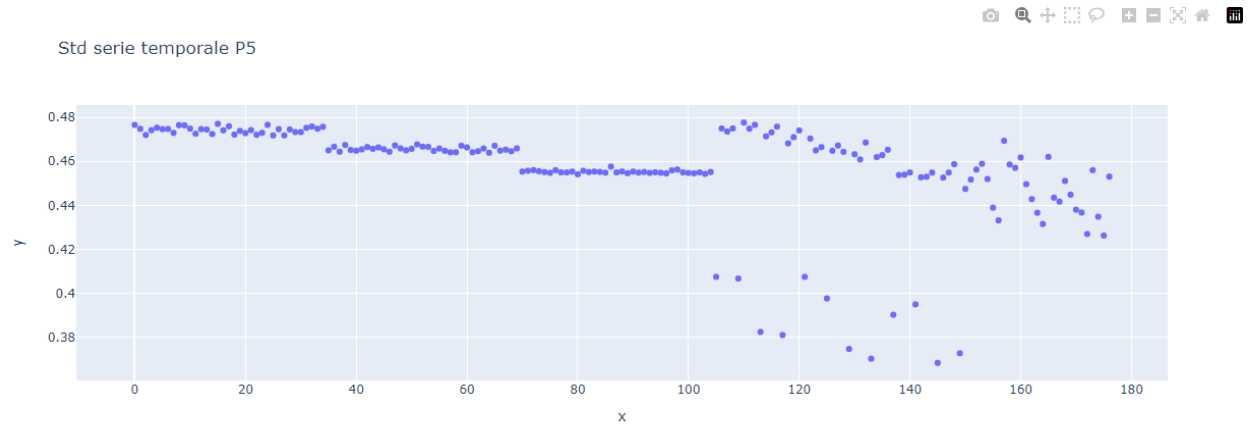


Figure 3: Varianza sensore P5

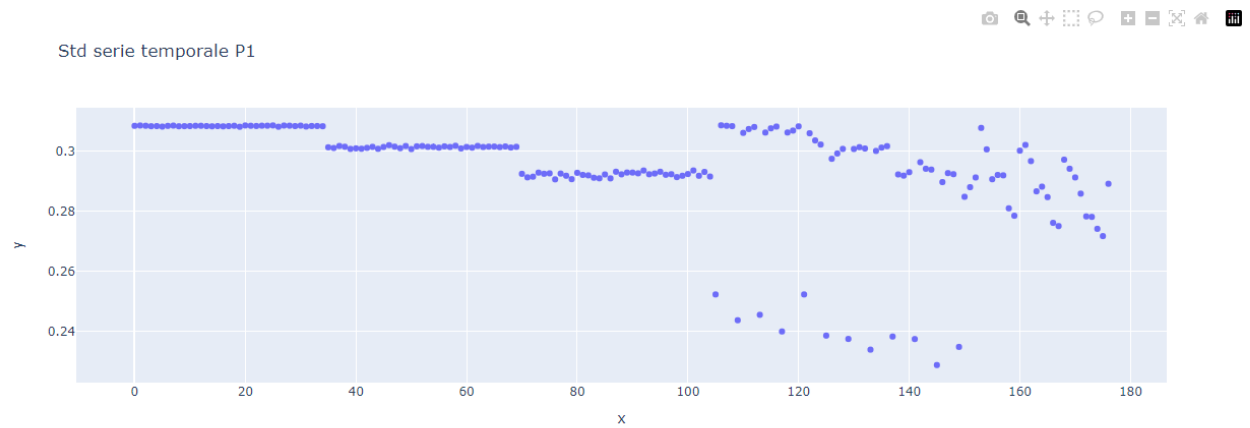


Figure 4: Varianza sensore P1

Nel dominio delle frequenze, durante una prima analisi esplorativa è stato possibile individuare diversi picchi massimi per i diversi casi normali e anormali. Diseguito le trasformate di fourier per il sensore P1 e P5 e lo spettro di potenza del sensore P5.

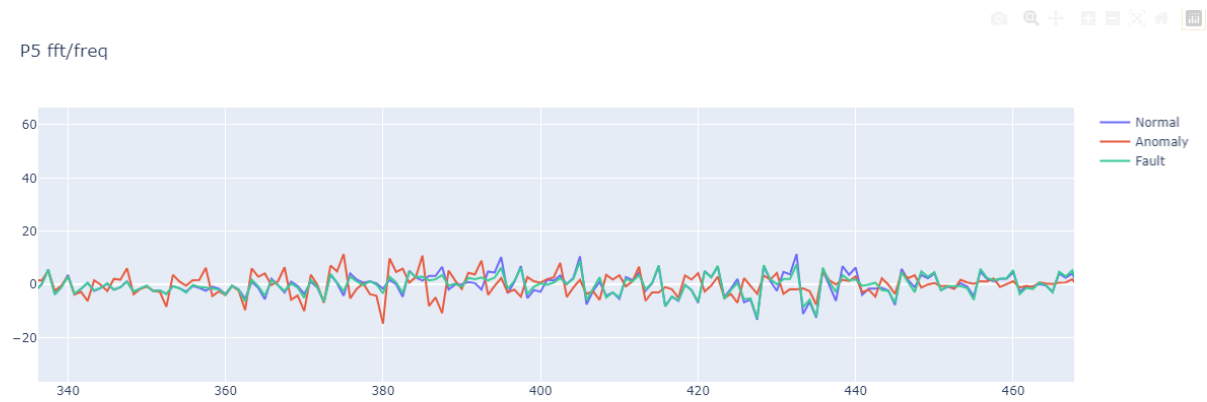


Figure 5: FFT sensore P5

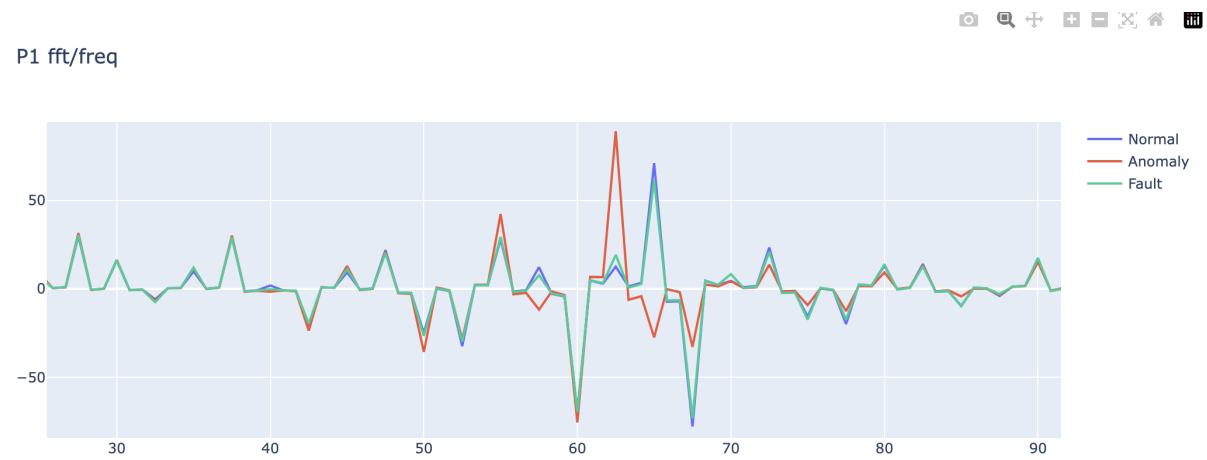


Figure 6: FFT sensore P1

Lo spettro di potenza con scala logaritmica è una rappresentazione che mostra la distribuzione dell'energia di un segnale in funzione della frequenza, ma con l'asse delle ampiezze espresso in decibel.

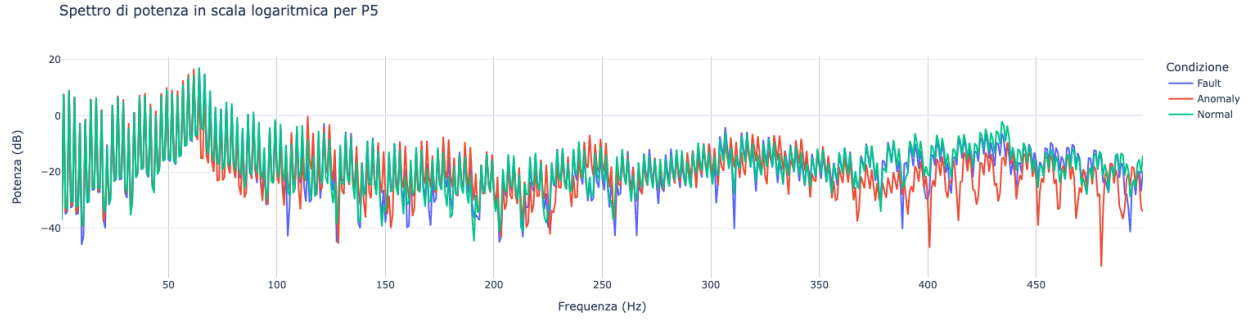


Figure 7: Spettro di potenza in scala log per sensore P5

3 Feature Engineering

Per feature engineering si intende il processo con il quale si selezionano e si estraggono features dai dati grezzi per creare un insieme di variabili di input da passare al modello di machine learning. In questo caso è stata effettuata sia un'analisi nel dominio del tempo che nel dominio delle frequenze.

Le statistiche sono state calcolate anche sulla fft di ogni serie temporale.

Nome	Formula
Media	$\mu = \frac{1}{n} \sum_{i=1}^n x_i$
Mediana	Corrisponde al 50° percentile
75° e 25° Percentile	i valori sotto i quali ricadono rispettivamente il 75% e 25% dei dati.
Varianza	$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$
Integrale di Linea	$\int_a^b f(x) dx \approx \frac{\Delta x}{2} (f(x_0) + 2 \sum_{i=1}^{n-1} f(x_i) + f(x_n))$
Minimo	$\min_{t \in \{1, \dots, T\}} x_t$
Massimo	$\max_{t \in \{1, \dots, T\}} x_t$

Table 1: Statistiche nel Dominio del Tempo

Nome	Formula
SINAD	$P_{\text{signal}} / \overline{P_{\text{noise}} + P_{\text{distortion}}}$
Entropia	$-\sum_{i=1}^N p_i \log(p_i)$
SNR	$P_{\text{signal}} / \overline{P_{\text{noise}}}$
Crest Factor	Peak Value $\overline{\text{RMS}}$
Impulse Factor	Peak Value $\overline{\mu}$
Band Power	$\sum_{f \in \text{band}} X(f) ^2$
Peak Value	$\max(x_i)$
Sum Power Spectrum	$\sum_{f=1}^N X(f) ^2$
Harmonic Amplitude Ratio	$\max(\text{spectrum}) / \overline{\text{mean}(\text{spectrum})}$
Clearance Factor	$\max(x_i) / \overline{\text{RMS}}$
RMS	$\sqrt{\frac{1}{N} \sum_{i=1}^N x_i^2}$
Sum power spectrum	$\sum_{f=1}^N X(f) ^2$

Table 2: Features nel Dominio delle Frequenze

Le caratteristiche (features) sono state selezionate per ogni task in base al modello adottato e ai risultati ottenuti. È stata eseguita un'analisi approfondita per identificare le features più rilevanti per ciascun compito.

Questo processo ha permesso di ottimizzare le performance del modello, migliorando la precisione delle previsioni e riducendo l'errore. I metodi adottati verranno meglio descritti nei paragrafi successivi.

4 Task 1

Il primo task consiste nel classificare i casi normali e anormali. Non essendo presente un dataset etichettato per effettuare i test, si è deciso di suddividere il dataset di training in due parti: una di addestramento e una di testing. Il dataset iniziale è stato suddiviso in 80% train e 20% test.

I dati sono stati etichettati con un valore binario che permette di distinguere i due casi, senza considerare il tipo di anomalia.

Tra i diversi modelli di classificazione provati, i risultati migliori sono stati ottenuti dal modello Random Forest.

Le feature sono state selezionate utilizzando il metodo integrato nel classificatore messo a disposizione dalla libreria sklearn. Questo metodo sfrutta i calcoli interni del modello per misurare l'importanza delle caratteristiche, come l'importanza di Gini e la diminuzione media della precisione. In sostanza, questo metodo misura quanto l'impurità (o la casualità) all'interno di un nodo di un albero decisionale diminuisce quando una specifica caratteristica

viene utilizzata per suddividere i dati.

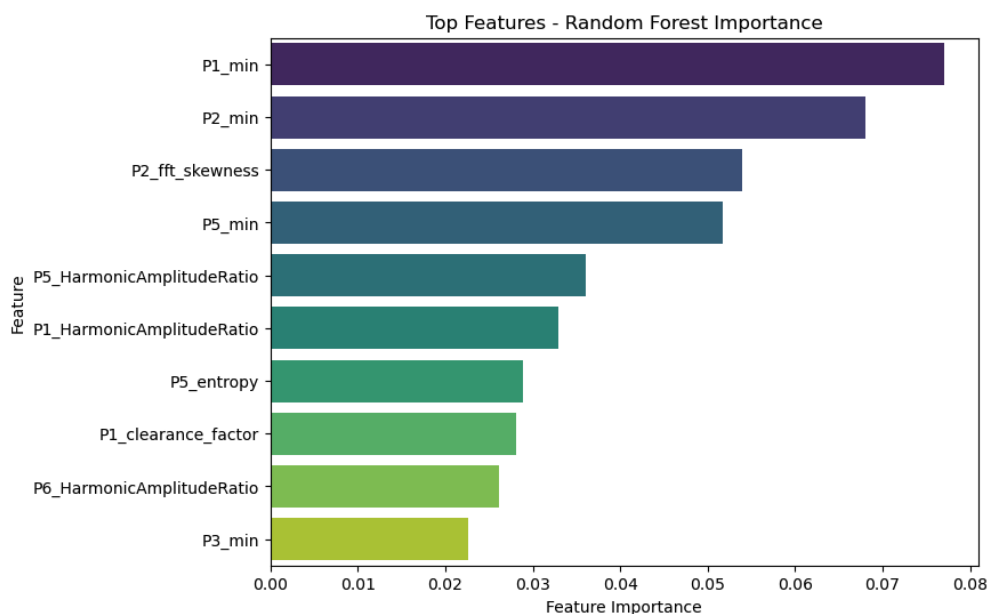


Figure 8: Features con metodo built-in random forest

Inoltre, è stata creata una Grid Search per valutare con quanti estimatori il modello riportasse i risultati migliori. La Grid Search esplora una serie di combinazioni di iperparametri per identificare la configurazione ottimale. Per verificare l'overfitting del modello, è stata eseguita una Cross Validation (con 5 folder). La Cross Validation suddivide il dataset di addestramento in diversi sottoinsiemi e addestra il modello su ciascuno di essi, valutandone le prestazioni sugli altri sottoinsiemi. Questo approccio aiuta a garantire che il modello sia in grado di generalizzare bene sui dati non visti.

Nell'immagine seguente si possono vedere i risultati ottenuti dalla classificazione con modello Random Forest, con 50 estimatori.

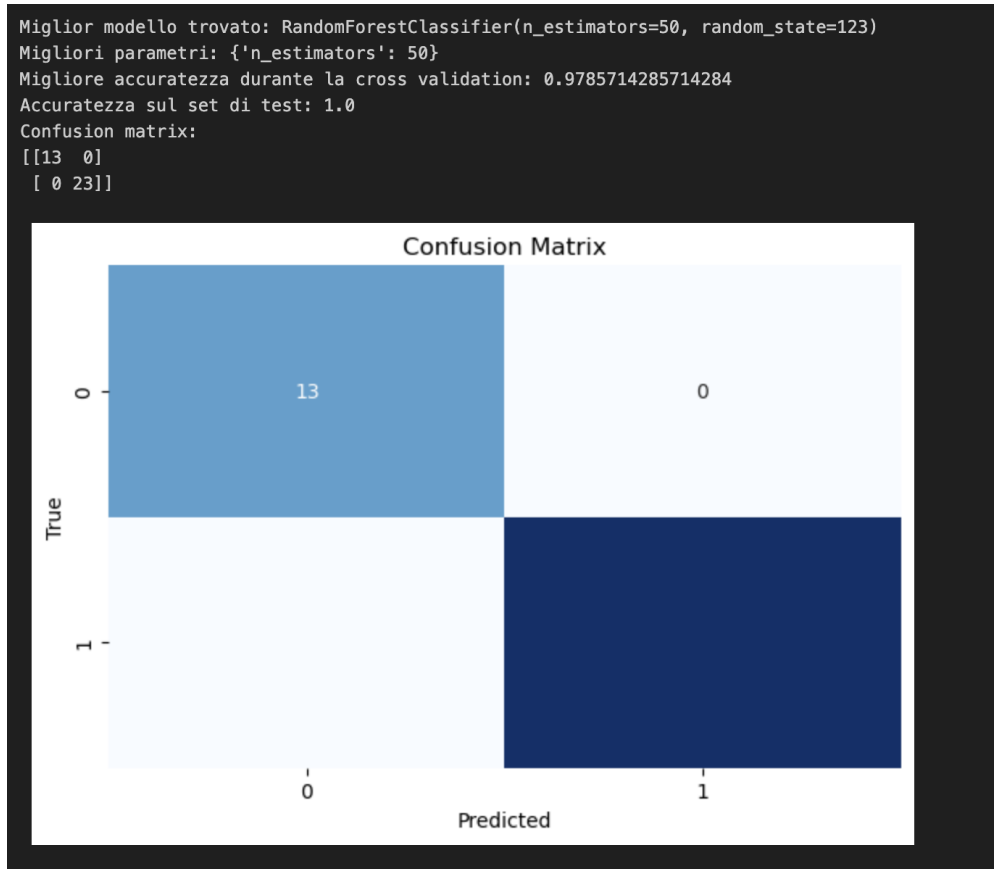


Figure 9: Risultati della classificazione con dataset di train

5 Task 2

Il secondo task consiste nel classificare i casi anormali del dataset di train e test, in anomalie per contaminazione da bolle d'aria, guasti alle elettrovalvole o guasti sconosciuti. Non essendo presenti anomalie sconosciute nei dati di train, la distinzione di questi casi particolari è stata fatta attraverso l'utilizzo di un approccio non supervisionato. Dopo aver classificato i dati di test in dati normali e anormali, questi ultimi sono stati utilizzati come input per un modello K-MEANS per distinguere le anomalie sconosciute da quelle conosciute. Per la classificazione dei casi anormali conosciuti sono stati provati diversi modelli, tra cui il Random Forest ed il classificatore XGBoost. Il miglior modello per distinguere i casi di contaminazione da bolle e guasti sulle valvole, è risultato essere XGBoost. Anche in questo caso, il miglior modello è stato scelto attraverso l'utilizzo di una grid search e di una cross-validation.

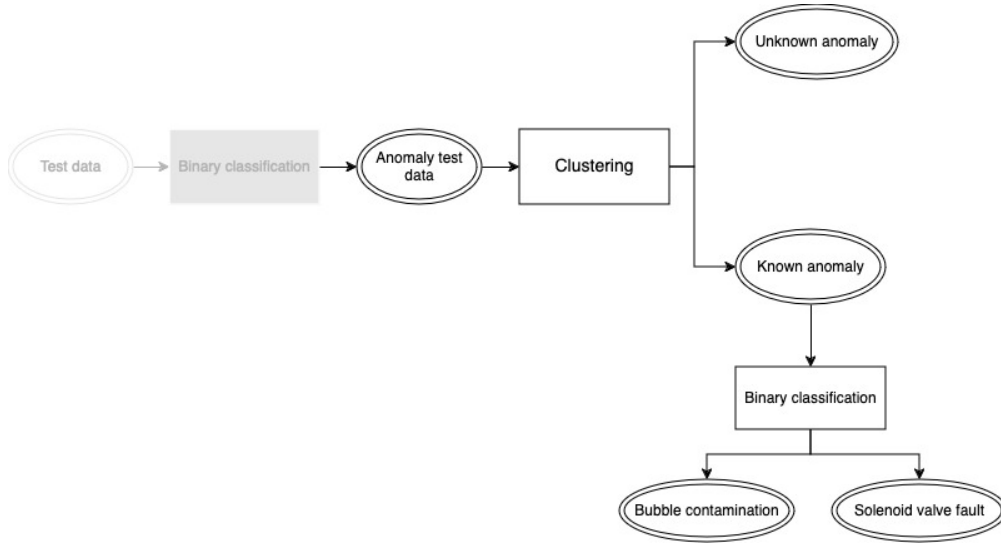


Figure 10: Schema task 2

Le features utilizzate per i diversi modelli sono state selezionate con il metodo ANOVA F-Value, implementato con la funzione `f_classif` di scikit-learn.

L'ANOVA F-value per la selezione delle feature funziona così: partiamo da un dataframe con N features e una colonna target categoriale. Per ogni feature, si calcola la media complessiva delle osservazioni e le medie delle osservazioni all'interno di ciascuna classe della variabile target. La variabilità tra gruppi misura la differenza tra le medie delle classi e la media complessiva della feature, ponderata per il numero di osservazioni in ogni classe. La variabilità all'interno dei gruppi misura la differenza tra ogni osservazione e la media del suo gruppo. Dopo aver calcolato i gradi di libertà, si determinano i valori medi dei quadrati per ottenere l'F-value come rapporto tra la variabilità tra gruppi e la variabilità all'interno dei gruppi. Un F-value alto indica che la feature è significativa per discriminare tra le classi della variabile target.

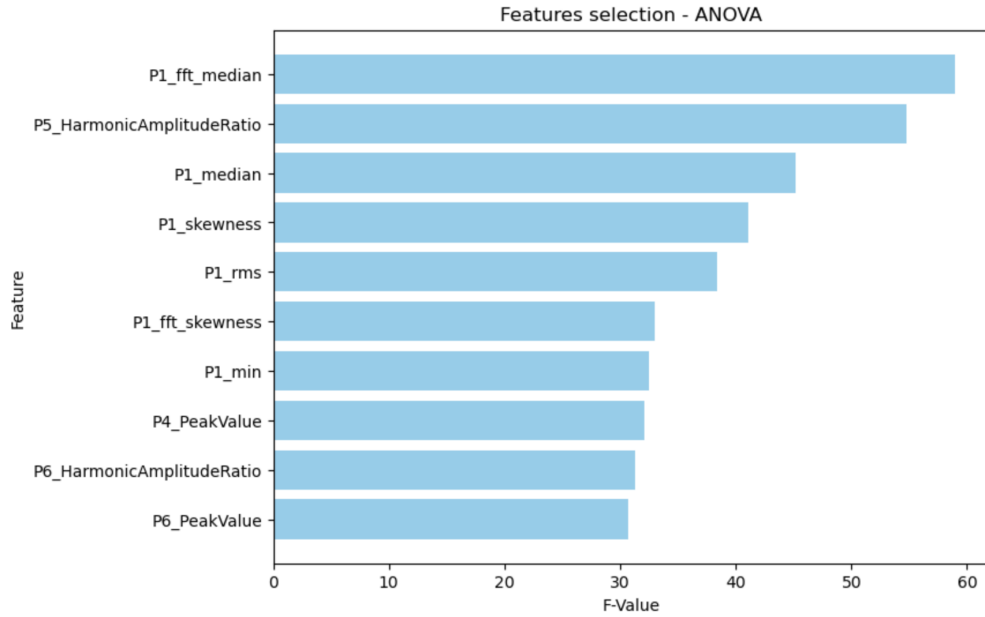


Figure 11: Selezione features task 2

I risultati del clustering mostrano una distinzione netta tra i casi conosciuti e sconosciuti. Nella figura si può notare tale distinzione per le features RMS e Minimo di P1 .

Clustering con K-means

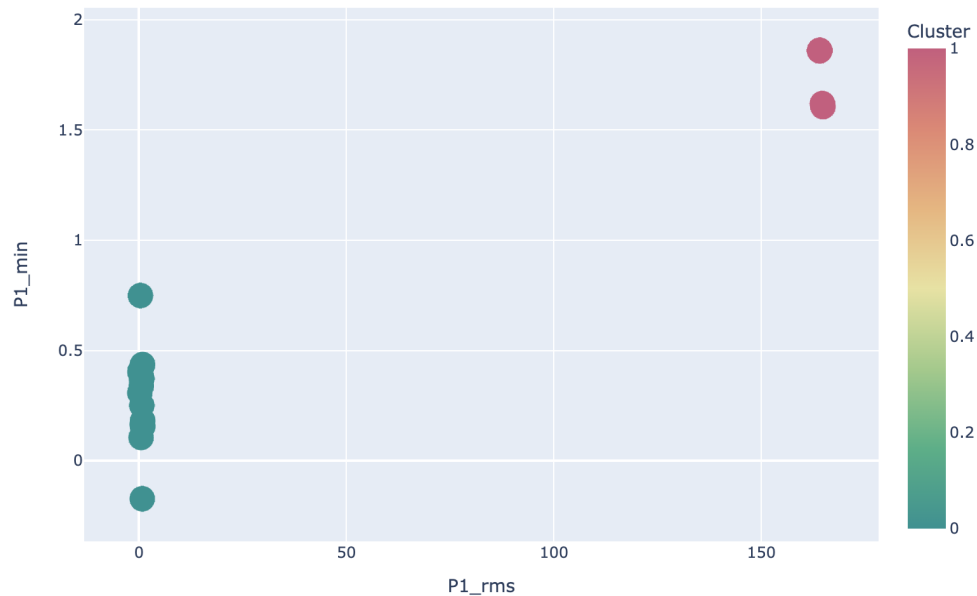


Figure 12: Clustering K-means task 2

Il modello che distingue i due diversi casi anormali è stato prima addestrato con i casi di

train. Come variabile target è stata utilizzata la condizione presente nel file *labels.xlsx*.

	Case	Spacecraft	Condition	SV1	SV2	SV3	SV4	BP1	BP2	BP3	BP4	BP5	BP6	BP7	BV1	binary_condition
index																
105	106	1	2	0	100	100	100	0	0	0	0	0	0	0	0	1
106	107	1	2	25	100	100	100	0	0	0	0	0	0	0	0	1
107	108	1	2	50	100	100	100	0	0	0	0	0	0	0	0	1
108	109	1	2	75	100	100	100	0	0	0	0	0	0	0	0	1
109	110	1	2	100	0	100	100	0	0	0	0	0	0	0	0	1
...
172	173	3	3	100	100	100	100	0	0	0	1	0	0	0	0	1
173	174	3	3	100	100	100	100	0	0	0	0	1	0	0	0	1
174	175	3	3	100	100	100	100	0	0	0	0	0	1	0	0	1
175	176	3	3	100	100	100	100	0	0	0	0	0	0	1	0	1
176	177	3	3	100	100	100	100	0	0	0	0	0	0	0	1	1

Figure 13: Dataframe per per file labels

Di seguito vengono riportati i risultati del modello XGBoostClassifier insieme con utilizzo di Grid search e del suo metodo interno di cross-validation.

```

Accuracy: 1.0
Confusion Matrix:
[[21  0]
 [ 0  8]]
Classification Report:
              precision    recall  f1-score   support

     0           1.00        1.00        1.00         21
     1           1.00        1.00        1.00         8

   accuracy              1.00          1.00          1.00         29
  macro avg           1.00          1.00          1.00         29
weighted avg           1.00          1.00          1.00         29

Best parameters found by grid search:
{'learning_rate': 0.01, 'max_depth': 3, 'n_estimators': 150}

```

Figure 14: Risultati train per classificare casi anomali

Lo stesso modello è stato utilizzato per classificare i casi nel dataset di test. La figura seguente riporta i casi distinti per le anomalie e fault.

```

Anomaly:  [178 186 193 196 197 204 209 216 219 221]
Fault:    [179 188 190 202 205 211 212 214]

```

Figure 15: Risultati test per classificare casi anomali

6 Task 3

Nel terzo task, l'obiettivo è individuare la posizione delle bolle d'aria all'interno dei casi precedentemente identificati. Questo problema è stato affrontato come una classificazione multi-classe, dove è stato creato un classificatore binario per ciascuna possibile posizione delle bolle.

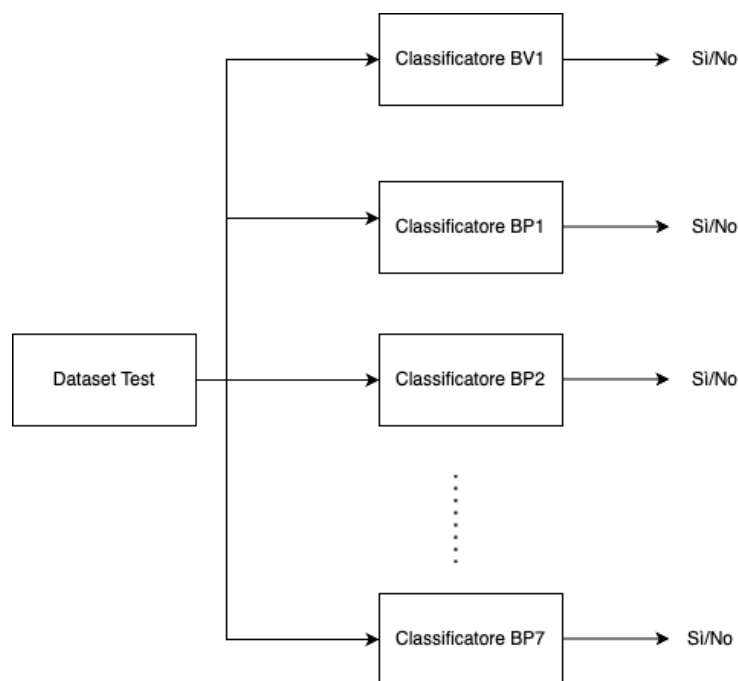


Figure 16: Classificatore multi-classe

Come input sono stati utilizzati sia il dataset delle caratteristiche dei dati originali, sia il dataset dei dati suddivisi in tre parti.

Nel secondo caso, ogni segnale è stato suddiviso in tre parti, triplicando così la quantità di dati a disposizione. Per selezionare le caratteristiche da utilizzare nei modelli, sono stati valutati tre approcci diversi: il test Chi-Square, il metodo interno del Random Forest e il metodo ANOVA. Dopo confronti dettagliati, è stato scelto di procedere con il metodo ANOVA, in quanto ha dimostrato di ottenere i migliori risultati con il modello multi-classe sviluppato.

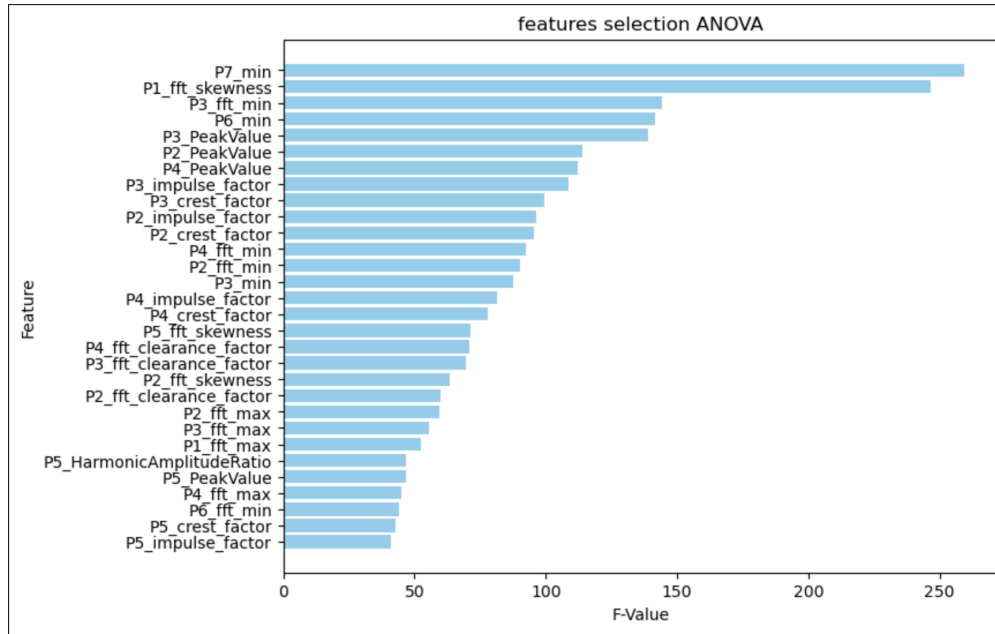


Figure 17: Features selection - task 3 - Anova

Anche in questo caso, il modello Random Forest è stato scelto come classificatore binario. Dopo aver addestrato il modello con i dati di train, sono stati utilizzati i casi di anomalia (contaminazione da bolle) del dataset di test ricavati nel precedente task. Di seguito sono riportati i risultati ottenuti.

	Case	Condition
0	178	BP2
1	186	BP6
2	193	BP1
3	196	BP4
4	197	BP7
5	204	BP3
6	209	BP7
7	216	NaN
8	219	BP5
9	221	BP2

Figure 18: Risultati task 3 - Dataset Test

Per ottenere risultati più affidabili dal modello, ogni segnale è stato suddiviso in tre parti. Questa suddivisione ci permette di ampliare il dataset, fornendo una base più ampia di dati per l'addestramento e il test. Con questa metodologia, siamo in grado di allenare il modello su tutte le possibili posizioni dei sensori, cosa che non sarebbe stata possibile con i dati originali. In questo modo, il modello può essere addestrato e testato su una gamma completa di posizioni, migliorando la sua capacità di generalizzare e di fornire risultati più attendibili e robusti.

La suddivisione dei dati è stata effettuata dividendo il segnale di 1200 elementi in 3 parti da 400 elementi ciascuna. Ogni segnale è composto da una prima parte di 100 istanti, che rappresenta l'apertura della valvola, e dai restanti 300 istanti che rappresentano la chiusura. Le caratteristiche sono state selezionate con lo stesso metodo utilizzato in precedenza, ovvero il metodo ANOVA, selezionando 20 caratteristiche.

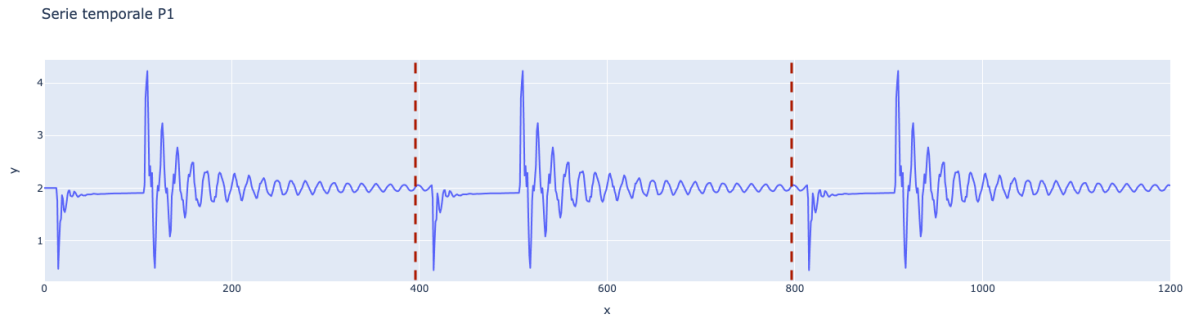


Figure 19: Split di ogni segnale

Le features sono state selezionate con lo stesso metodo utilizzato in precedenza, ovvero il metodo ANOVA. Anche in questo caso sono state selezionate 20 features.

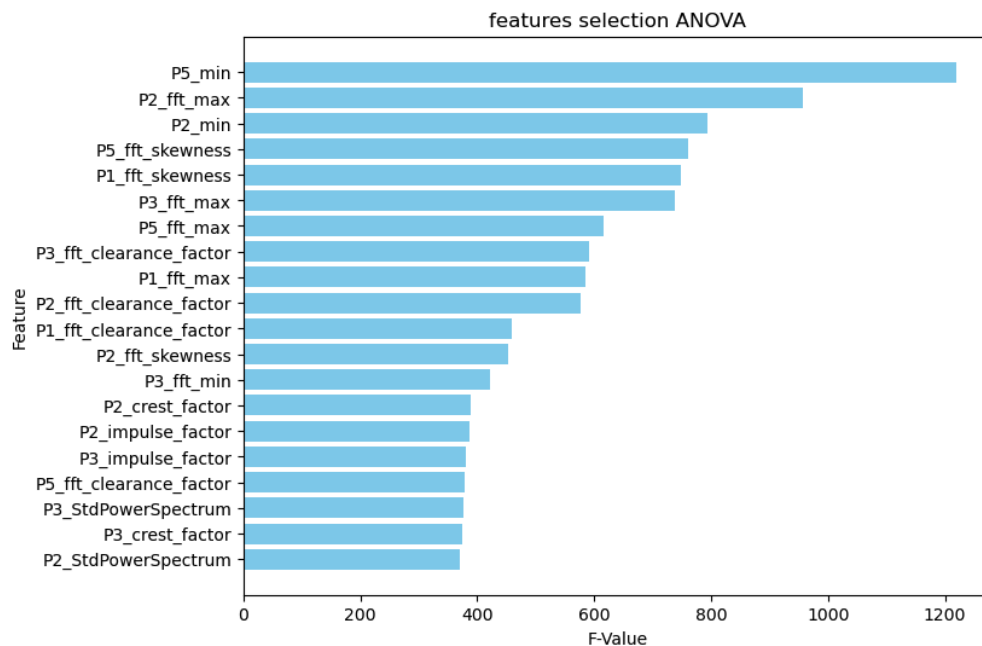


Figure 20: Split di ogni segnale

Per l'addestramento dei modelli, il dataset è stato suddiviso in 70% dati di train e 30% dati di test. I risultati ottenuti dall'addestramento sono riportati nelle seguenti heatmap.

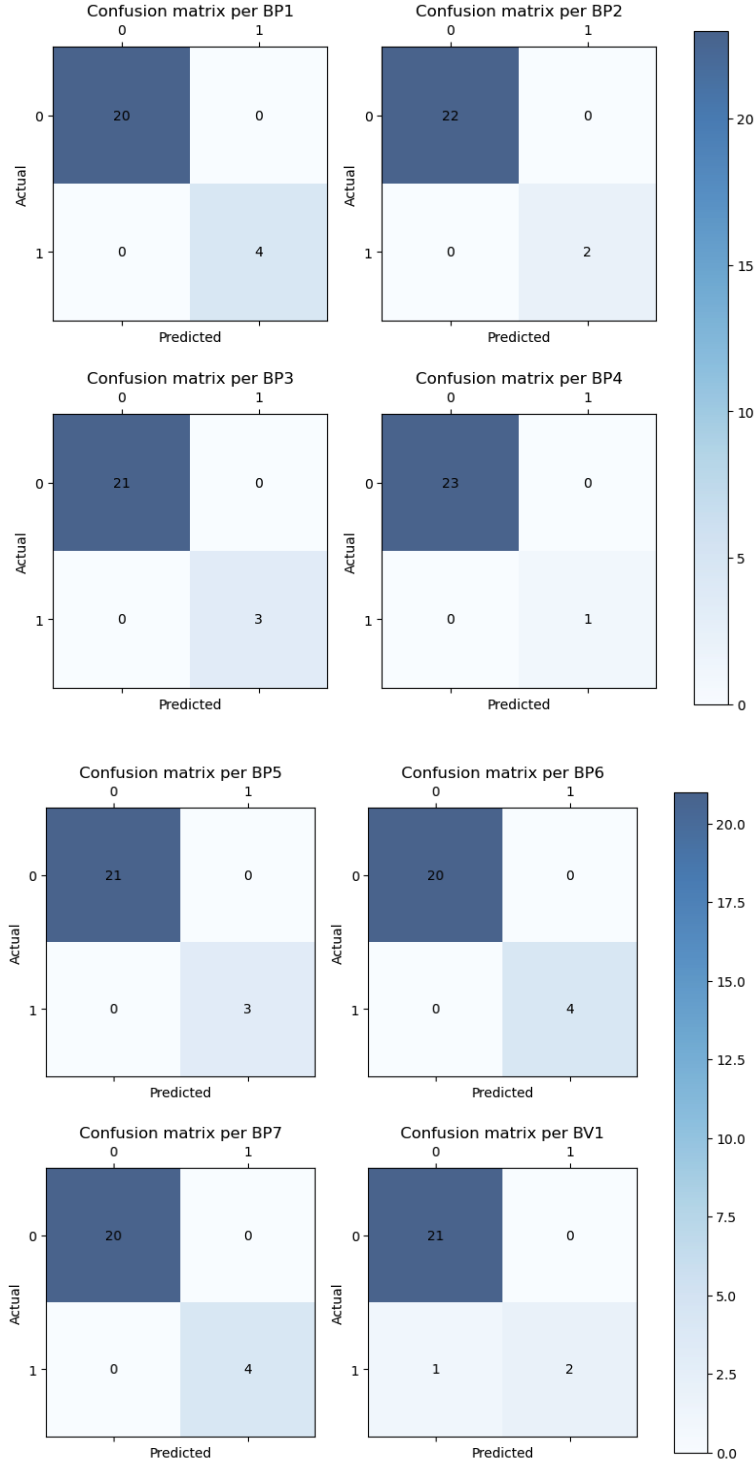


Figure 21: Matrice di confusione sensori BP1-BP7 e BV1

I modelli addestrati sono stati successivamente utilizzati per classificare i dati di test, e i risultati ottenuti sono presentati nella seguente tabella. Solo uno dei casi non è stato classificato correttamente, questo caso è indicato con *NaN*

	Case	Condition	real_case
0	0	BP2	178
1	1	BP2	178
2	2	BP2	178
3	24	BP6	186
4	25	BP6	186
5	26	BP6	186
6	45	BP1	193
7	46	BP1	193
8	47	BP1	193
9	54	BP4	196
10	55	BP4	196
11	56	BP4	196
12	78	BP3	204
13	79	BP3	204
14	80	BP3	204
15	114	nan	216
16	115	nan	216
17	116	nan	216
18	123	BP5	219
19	124	BP5	219
20	125	BP5	219
21	129	BP2	221
22	130	BP2	221
23	131	BP2	221

Figure 22: Risultati classificazione dataset di test (casi 178-223)

7 Task 4

Nel quarto task, l'obiettivo principale è quello di identificare con precisione la posizione in cui si verifica un guasto alle elettrovalvole. Per raggiungere questo obiettivo, i dati sono stati suddivisi in tre insiemi distinti, simili a quanto fatto nel task precedente. Questa suddivisione consente di avere una quantità sufficiente di dati sia per l'addestramento che per il test del modello, migliorando così la robustezza e l'affidabilità delle previsioni.

Per selezionare le variabili più rilevanti, è stato applicato il metodo interno del Random Forest, che ha permesso di individuare le 20 feature più significative per la classificazione. La figura ?? illustra chiaramente queste feature selezionate, offrendo una panoramica visiva delle variabili considerate cruciali per il processo di classificazione.

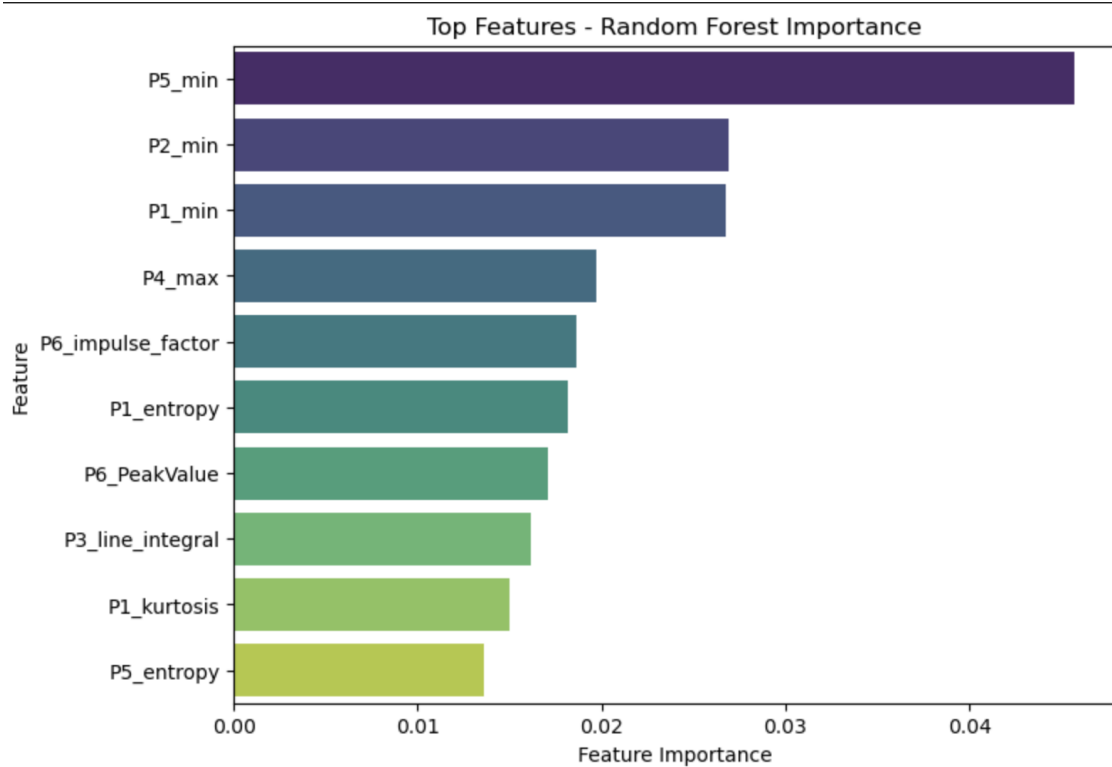


Figure 23: Features selezionate per il task 4

Per la classificazione dei dati, è stato utilizzato un approccio multi-classe, in cui ogni classificatore è basato su un modello XGBoost. Questo modello è stato ottimizzato attraverso una grid search, che ha permesso di trovare i migliori parametri per ottenere le stime più precise. La figura 23 successiva mostra le matrici di confusione ottenute durante l'addestramento del modello, fornendo una visione dettagliata delle prestazioni del modello sui dati di addestramento.

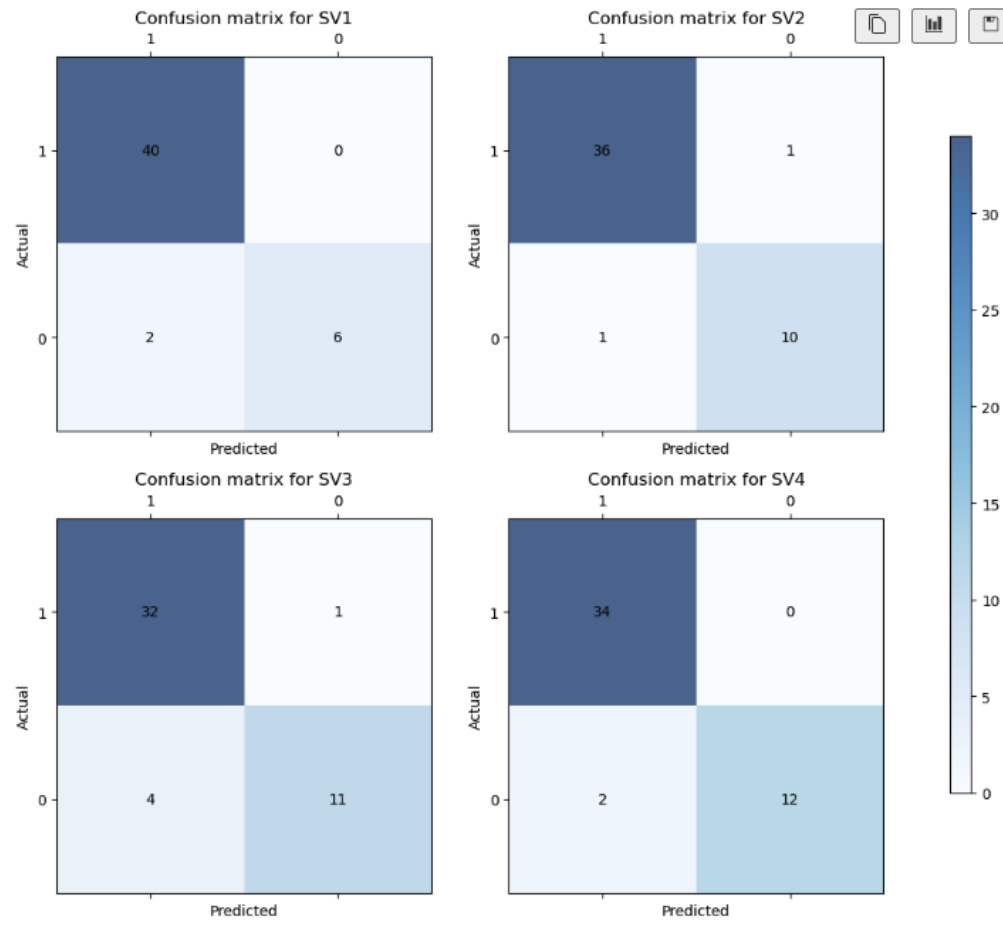


Figure 24: Matrici di confusione per l'addestramento del task 4

Durante la fase di classificazione dei dati di test, sono emersi alcuni casi in cui la classificazione non è stata corretta. Questi casi sono stati identificati e segnalati nella colonna *Condition* con il valore *NaN*. La figura 35 successiva presenta i risultati della classificazione dei dati di test per il task 4, evidenziando le aree in cui il modello ha mostrato difficoltà e suggerendo possibili aree di miglioramento.

	Case	Condition	real_case
0	3	SV2	179
1	4	SV2	179
2	5	SV2	179
3	30	SV1	188
4	31	SV1	188
5	32	SV1	188
6	36	nan	190
7	37	SV3	190
8	38	SV3	190
9	72	nan	202
10	73	SV3	202
11	74	SV3	202
12	81	SV2	205
13	82	SV2	205
14	83	SV2	205
15	99	SV2	211
16	100	nan	211
17	101	nan	211
18	102	SV2	212
19	103	SV2	212
20	104	SV2	212
21	108	SV4	214
22	109	SV4	214
23	110	SV4	214

Figure 25: Risultati della classificazione dei dati di test per il task 4

Questo scenario indica che, nonostante l'adozione di tecniche avanzate come XGBoost e la grid search, ci sono ancora dei margini di miglioramento nella classificazione, che potrebbero essere affrontati con ulteriori raffinamenti dei modelli o con l'inclusione di nuovi dati.

8 Task 5

Questo task consiste nel determinare, per i casi guasti, il rapporto di apertura della eltrovalvola. L'idea è quella di utilizzare un modello di regressione per individuare la percentuale di apertura, utilizzando come regressori le features selezionate con il metodo chi2 e come variabile dipendente la percentuale da predire.

Sono state calcolate le correlazioni tra l'apertura e le 15 features più importanti, in modo da indentificare eventuali relazioni lineari; in particolare è stato utilizzato il coefficiente di pearson per determinare la forza e la direzione della relazione lineare.

	Column	Correlation	P-value
8	P2_min	-0.573881	2.004876e-05
9	P2_line_integral	-0.671363	1.761682e-07
7	P1_min	-0.737157	2.304863e-09
4	P5_25th_percentile	-0.758784	4.138854e-10
3	P1_mean	-0.759074	4.039747e-10
2	P1_line_integral	-0.759104	4.029543e-10
1	P2_25th_percentile	-0.763973	2.668840e-10
6	P5_mean	-0.765671	2.306176e-10
5	P5_line_integral	-0.765767	2.287271e-10
0	P1_25th_percentile	-0.773289	1.179680e-10

Figure 26: Correlazioni tra features e open ratio

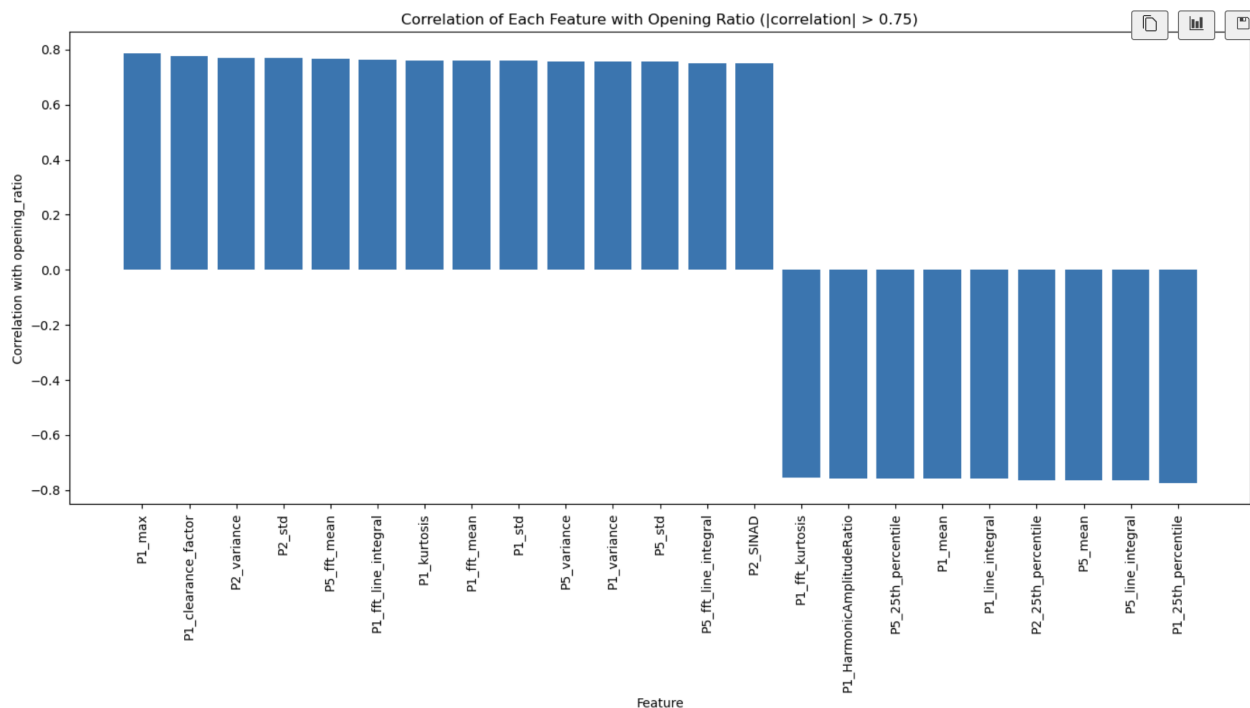


Figure 27: Plot correlazioni più tra features e open ratio

i risultati della regressione lineare per i dati di train è presente nella figura seguente. L'errore assoluto (MAE) è circa del 19%.

Il dataframe è stato suddiviso in 33% dati di test e 67% dati di train, con un random state di 42.

```

y_pred: [ 60.83357369  0.75011531 52.26182063 58.80237202  9.36844626
 63.43013389 -3.12568734 55.6041524  9.77588123 30.25210094
-29.34059039 41.2636179 57.64913094 42.09137655 74.22567841
55.50923475]
-----
y_test : [75  0 50 75  0 25  0 75  0 25  0 75 50 75 25 25]

```

Figure 28: Risultati regressione lineare per dati di train

Come feature aggiuntiva è stata aggiunta la distanza DTW (dynamic time warping) tra ogni caso di guasto ed i casi normali. La distanza considerata per ogni caso è la media delle distanze tra ogni caso guasto e tutti i casi normali. Questa feature è stata alla fine eliminata in quanto peggiorava notevolmente i risultati della regressione Lineare.

Dalla figura si può notare l'andamento lineare dei casi di guasto nei dati di train per diversi rapporti di apertura, la line rossa nella figura segna il campo di spacecraft.

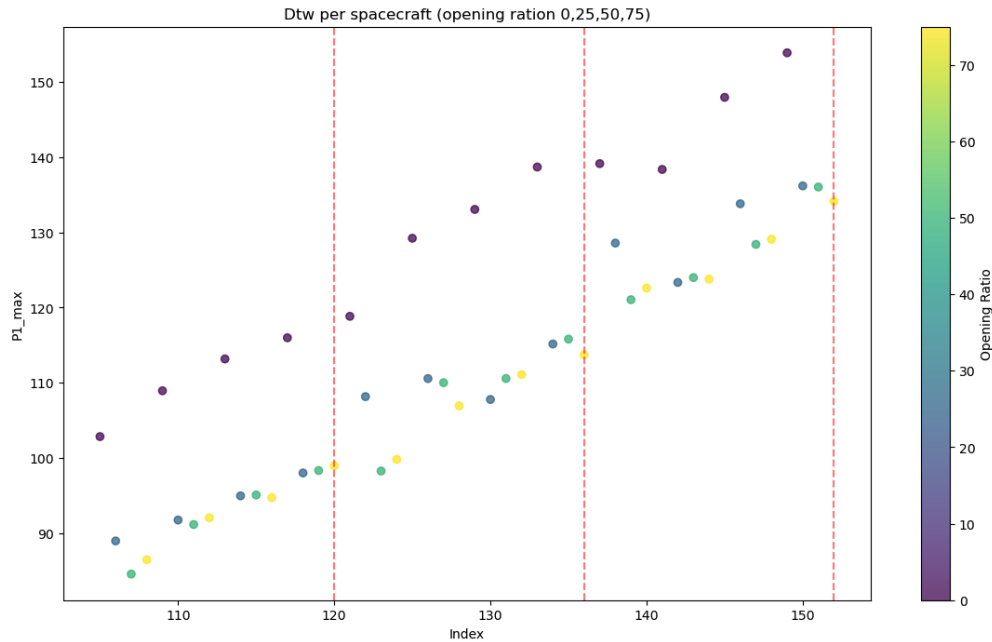


Figure 29: DTW per spacecraft

9 Risultati

In questa sezione vengono analizzati i risultati ottenuti nei vari compiti. Questi risultati sono stati confrontati con quelli riportati sul sito PHMAP.

9.1 Task 1

Per valutare l'accuratezza della classificazione dei casi normali e anormali, sono stati utilizzati i risultati disponibili sul sito web specificato. Il modello di classificazione ha dimostrato un'elevata efficacia.

In particolare, il modello ha riportato due casi di falso positivo, ossia situazioni in cui casi

normali sono stati erroneamente classificati come anormali. Nonostante questi falsi positivi, l'accuratezza complessiva del modello è del 96%, indicando che il modello fornisce previsioni corrette nella stragrande maggioranza dei casi.

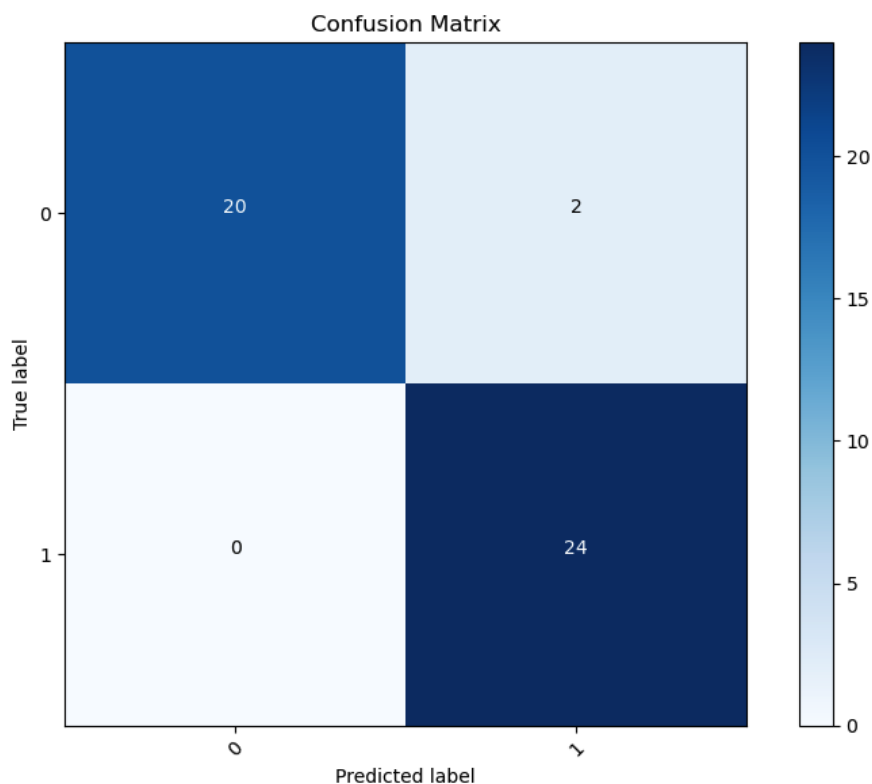


Figure 30: Matrice di confusione task 1

i due falsi positivi sono i casi 181 e 199 dello spacecraft numero 1.

9.2 Task 2

Con l'algoritmo di clustering utilizzato nel task 2, è stato possibile distinguere e classificare correttamente tutti i casi sconosciuti, come mostrato nella Figura 31. Questo risultato evidenzia l'efficacia del clustering nel rilevare e separare correttamente le diverse categorie all'interno dei dati sconosciuti.

	task2_real	task2_pred	Case
0	1	1	184
1	1	1	192
2	1	1	200
3	1	1	207
4	1	1	218
5	1	1	222

Figure 31: Risultati classificazione dei casi sconosciuti per task 2

Successivamente, i casi conosciuti sono stati impiegati come input per un modello di classificazione, il quale ha riportato un'accuratezza complessiva del 90%. Questo livello di accuratezza indica una buona performance del modello nel prevedere correttamente la categoria dei casi noti. Tuttavia, è importante notare che gli unici due casi classificati erroneamente sono stati i due falsi positivi già identificati nel task 1. Inoltre, sono stati classificati correttamente tutti i casi dello spacecraft 4. I risultati sono mostrati nella Figura 32, l'etichetta 0 indica i casi errati del task 1, l'etichetta 1 i casi anomali e l'etichetta 2 i casi di fault.

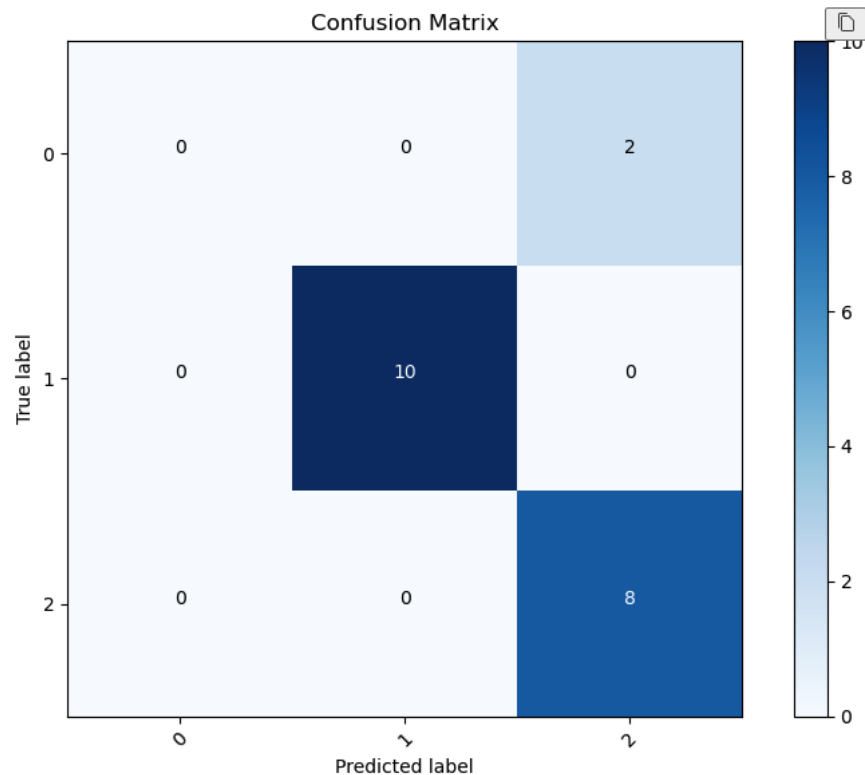


Figure 32: Matrice di confusione task 2

Tutti i casi di anomalia sono stati classificati correttamente ed i due casi classificati erroneamente sono due casi di fault.

9.3 Task 3

Il modello di classificazione del task 3 ha raggiunto un'accuratezza del 90%. L'unico errore di classificazione si è verificato con il caso 216, per il quale il modello non è riuscito ad assegnare una posizione, restituendo un valore NaN (classe -1).

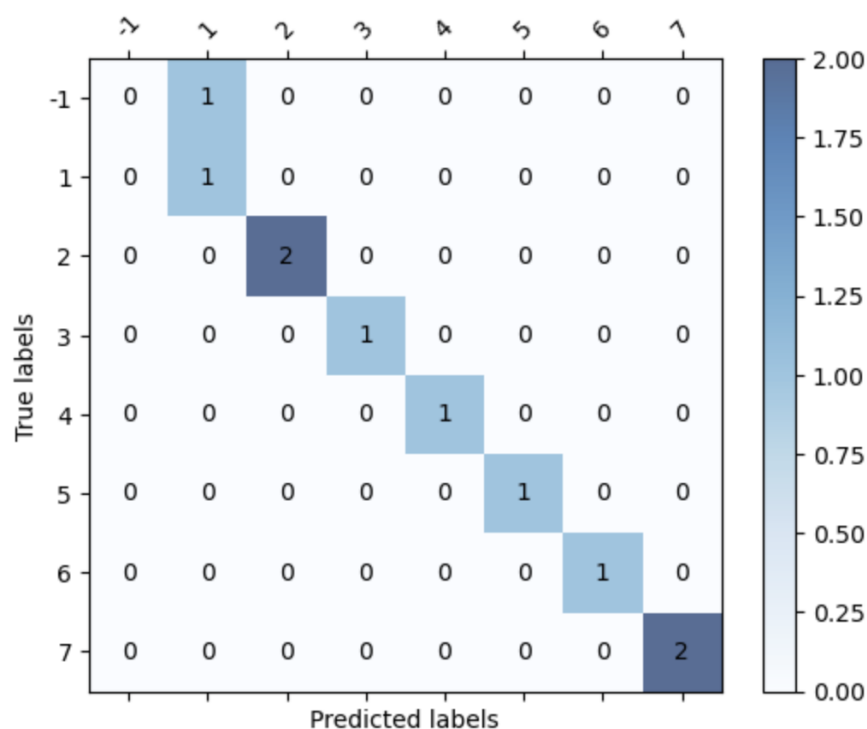


Figure 33: Matrice di confusione task 3

9.4 Task 4

Il modello di classificazione del task 4 ha raggiunto un'accuratezza del 70%. Dei 3 errori presenti, 2 sono errori provenienti dalla classificazione del task 1. Anche in questo caso la classe -1 indica eventuali campioni NaN.

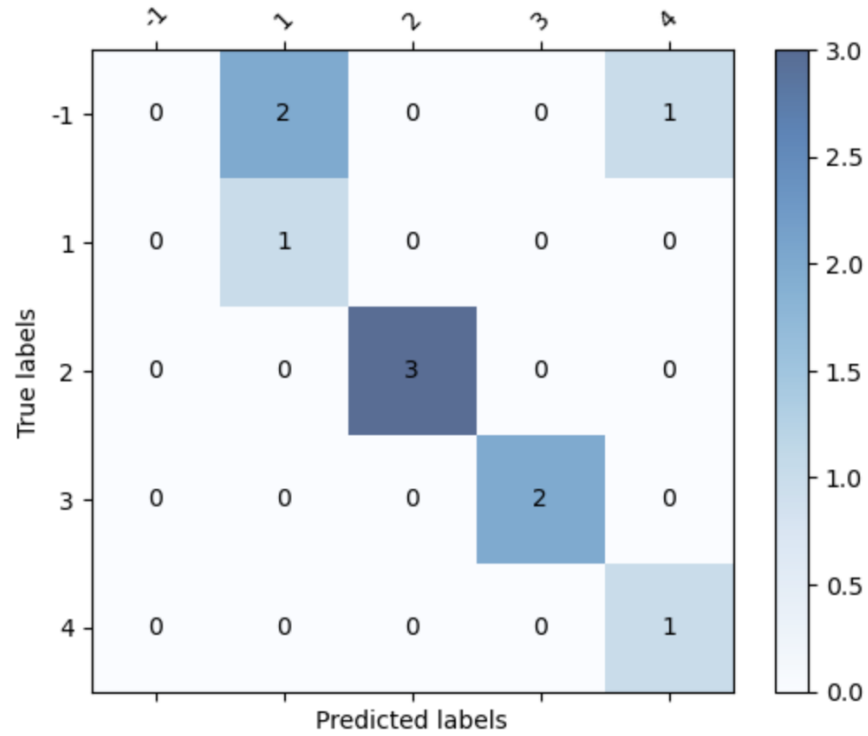


Figure 34: Matrice di confusione task 4

9.5 Task 5

Il modello di regressione lineare del task 5 riporta un errore assoluto (Mean Absolute Error) del circa 34%. Di seguito vengono riportati i risultati del modello.

	fault_loc_pred	Case
0	38.677162	179.0
1	0.623380	188.0
2	33.152843	190.0
3	50.999356	202.0
4	50.380887	205.0
5	46.528206	211.0
6	65.056425	212.0
7	53.765253	214.0
8	-1.000000	199.0
9	-1.000000	181.0

Figure 35: Risultati modello di regressione lineare per il task 5