Università Politecnica delle Marche

Facoltà di Ingegneria

Dipartimento di Ingegneria dell'Informazione

Corso di Laurea Magistrale in Ingegneria Informatica e dell'Automazione

---

Progetto di Manutenzione Preventiva per la Robotica e l'Automazione Intelligente

PHM North America 2023 Data Challenge

Docente

Freddi Alessandro

A cura di

D'Anna Alessandra,
Di Sabatino Walter,
Visi Andrea

Anno Accademico 2024-2025

Indice

Elenco delle Figure

1 Introduzione

La seguente relazione si pone l'obiettivo di andare a descrivere il lavoro svolto per lo svolgimento della Data Challenge del 2023 proposta dalla *PHM North America Conference*, evento di riferimento nel settore del Prognostic and Health Management.

Il tema principale riguarda la stima del degrado di un riduttore (gearbox) attraverso l'analisi di segnali di vibrazione raccolti in condizioni operative variabili. La sfida consiste nello sviluppare un modello in grado di stimare il livello di guasto (*fault severity*) del sistema, fornendo anche una misura di confidenza della previsione.

Un aspetto fondamentale è la capacità di generalizzazione del modello a condizioni e stati di salute che non sono riscontrabili nel dataset di addestramento, riflettendo scenari realistici.

Per affrontare il problema sono stati forniti due dataset principali:

- **Training Set**: organizzato in cartelle separate in base all'*Health Level* di appartenenza;
- **Testing Set**: contenente i file destinati alla valutazione del modello.

I file seguono una convenzione di nomenclatura che esplicita le condizioni operative corrispondenti.

La nomenclatura citata presenta la seguente forma:

`NNN_Vxxxx_yyyN_z.txt`

dove:

- NNN : identificativo progressivo del campione;
- Vxxxx: velocità di rotazione dell'albero di ingresso in rpm;
- yyyN: coppia applicata sull'albero di uscita in Newton metro (Nm);
- z: numero della ripetizione della misura.

Riportiamo un esempio di utilizzo di tale convenzione in Figura 1. È importante sottolineare che tale convenzione, con identificativo progressivo iniziale (`NNN`), è utilizzata nei file del *Testing Set*, mentre nel *Training Set* i file sono suddivisi in cartelle in base al livello di degrado e non riportano tale numerazione progressiva.

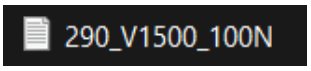

Figure 1: Esempio di nome di un file.

2 Data Understanding & Exploratory Data Analysis

L'analisi esplorativa dei dati rappresenta una fase cruciale per comprendere la struttura e le caratteristiche generali del dataset a disposizione, oltre che per valutarne qualità e coerenza. Questa sezione introduce i dati oggetto di studio e descrive il processo preliminare di organizzazione e preparazione che ha permesso di impostare correttamente le analisi successive.

2.1 Dataset overview e parsing intelligente

Il dataset PHM North America 2023 è composto da misurazioni di vibrazione triassiale raccolte su un riduttore a singolo stadio sotto diverse condizioni operative. I dati includono:

- **Health levels**: 0-10 (dove 0 = sano, 10 = massimo degrado);
- **Condizioni operative**: 15 velocità $\times$ 6 livelli di coppia nel training;
- **Segnali**: accelerazioni orizzontale, assiale, verticale + tachimetro;
- **Frequenza campionamento**: 20.480 Hz.

Il processo di parsing del dataset ha richiesto lo sviluppo di una procedura robusta, in grado di interpretare correttamente la nomenclatura dei file e di estrarre i relativi metadati operativi.

In particolare:

- Dai nomi dei file è stato possibile ricavare informazioni relative ai principali parametri operativi, come il numero di giri del riduttore (RPM), il livello di coppia applicata (Torque) e l'indice di ripetizione della misura.
- Dai nomi delle cartelle è stato invece estratto il livello di degrado associato a ciascun esperimento, indicato con il termine Health Level.

La procedura di parsing è stata progettata per gestire in maniera robusta variazioni o irregolarità nella nomenclatura, garantendo così l'assegnazione corretta dei metadati anche in presenza di formati non perfettamente standardizzati.

Grazie a questo approccio è stato possibile organizzare in maniera sistematica l'intero dataset, che comprende un totale di 2016 file di vibrazione triassiale. Le statistiche principali relative al parsing sono riportate in Tabella 1.

Table 1: Statistiche del parsing del dataset PHM North America 2023

Metrica	**Valore**
File totali processati	2016
Parsing riusciti	2016
Parsing falliti	0
Tasso di successo	100%
Condizioni operative uniche	78
Repetitions medie per condizione	3.69

2.2 Analisi della distribuzione degli health levels

L'analisi della distribuzione degli health levels mostrata in Figura 2 rivela caratteristiche critiche del dataset che influenzano significativamente l'approccio di modellazione.

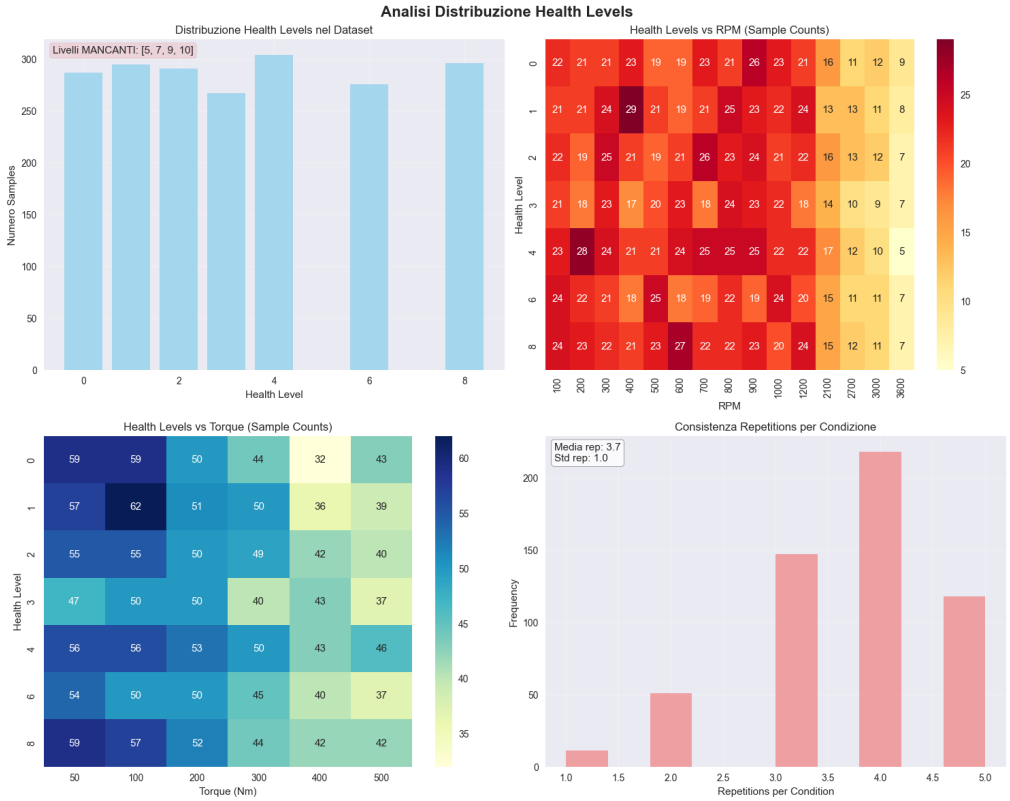

Figure 2: Distribuzione degli health levels nel dataset. (a) Distribuzione generale, (b) Health levels vs RPM, (c) Health levels vs Torque, (d) Consistenza repetitions per condizione

L'analisi rivela un aspetto fondamentale ai fini della comprensione della challenge: gli health levels 5, 7, 9 e 10 sono completamente assenti dal dataset di training, rappresentando una sfida significativa per la generalizzazione del modello. Questa caratteristica riflette condizioni realistiche dove non tutti i livelli di degrado sono disponibili durante la fase di training.

2.3 Analisi delle condizioni operative

Le condizioni operative rappresentano un fattore critico per la robustezza del modello. L'analisi della coverage operativa è mostrata in Figura 3.

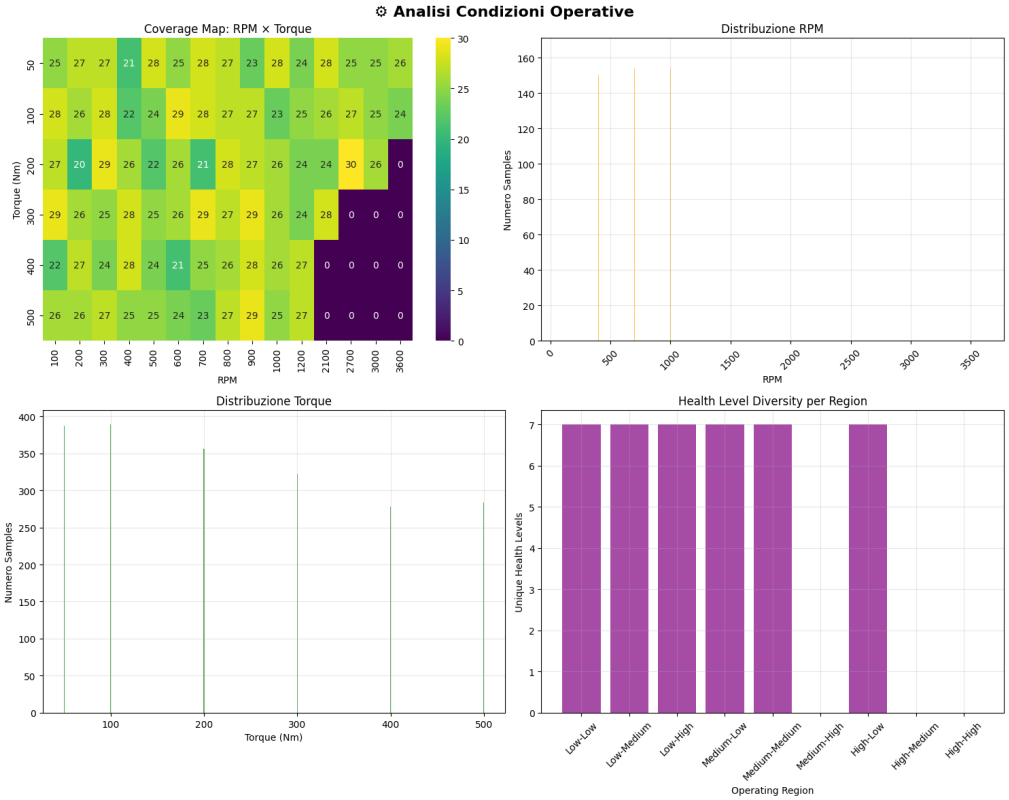

Figure 3: Analisi condizioni operative. (a) Coverage map RPM × Torque, (b) Distribuzione RPM, (c) Distribuzione Torque, (d) Diversità health levels per regione operativa

L'analisi rivela:

- **Coverage completa**: tutte le combinazioni RPM-Torque sono rappresentate nel training set.
- **Range operativo**: RPM da 100 a 3600, Torque da 50 a 500 Nm.
- **Distribuzione bimodale**: concentrazione su velocità basse (100-1000 RPM) e alte (2400-3600 RPM).

2.4 Feature Analysis vs Health Levels

L'analisi delle relazioni tra features e health levels (Figura 4) fornisce insights fondamentali sulla discriminabilità delle caratteristiche estratte. I passaggi specifici all'ingegneria delle feature verranno approfonditi nel Capitolo 3.

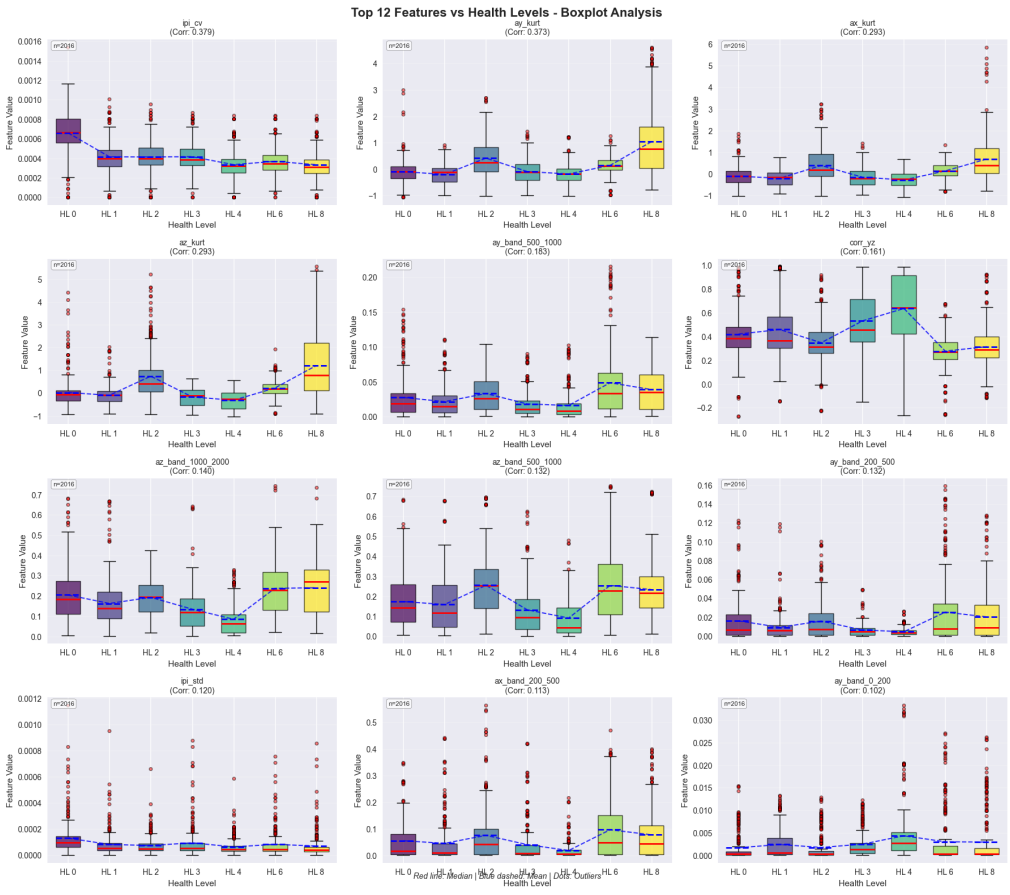

Figure 4: Boxplot delle top 12 features vs health levels. Le features mostrano chiari trend monotoni con l'evoluzione del degrado.

Le features più discriminanti presentano caratteristiche ottimali:

1. Trend Monotoni: le features RMS degli assi mostrano incrementi consistenti con l'aumento dell'health level, coerenti con la fisica del degrado.

2. Separabilità: le distribuzioni delle features tra diversi health levels mostrano overlap limitato, indicando buona capacità discriminativa.

3. Robustezza: le features mantengono trend consistenti tra diverse condizioni operative.

Nota sulla correlazione tra feature L'analisi preliminare delle correlazioni ha messo in evidenza la presenza di dipendenze significative tra diverse variabili, in particolare tra feature derivate dallo stesso dominio di misura. Questo fenomeno, noto come

multicollinearità, può influenzare negativamente le fasi di training dei modelli, riducendo l'interpretabilità e introducendo ridondanza informativa.

Per garantire un set di variabili più compatto ed efficiente, la riduzione della multicollinearità è stata oggetto di un'analisi mirata, descritta nel Capitolo 3.

2.5 Outlier detection e anomaly analysis

L'identificazione di eventuali outliers e anomalie risulta cruciale per la robustezza del modello.

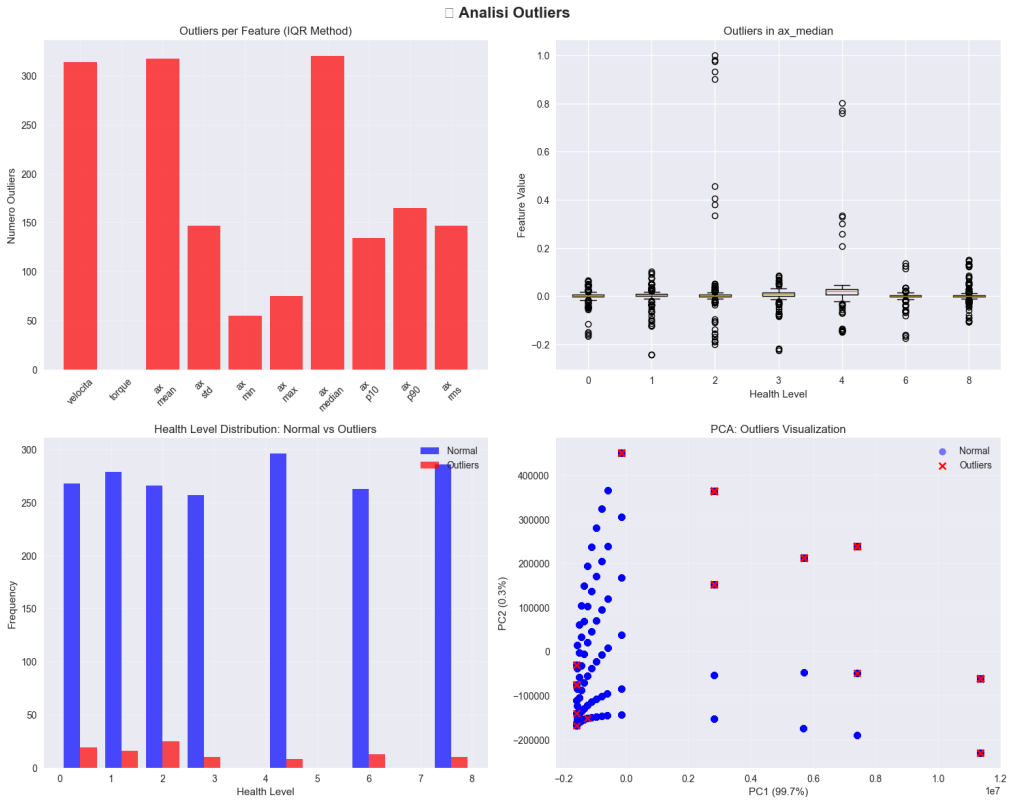

Figure 5: Analisi outliers. (a) Outliers per feature (IQR method), (b) Feature con più outliers, (c) Distribuzione health levels: normal vs outliers, (d) PCA visualization outliers

L'analisi con Isolation Forest identifica il 5% dei samples come outliers multivariati. Le informazioni principali ricavate dall'analisi sono:

- **Distribuzione**: gli outliers sono distribuiti uniformemente tra gli health levels.
- **Clustering**: la visualizzazione PCA rivela cluster di outliers fisicamente motivati.

2.6 Physics-Based Validation

La validazione basata su principi fisici conferma la coerenza delle features estratte con la fenomenologia del degrado degli ingranaggi.

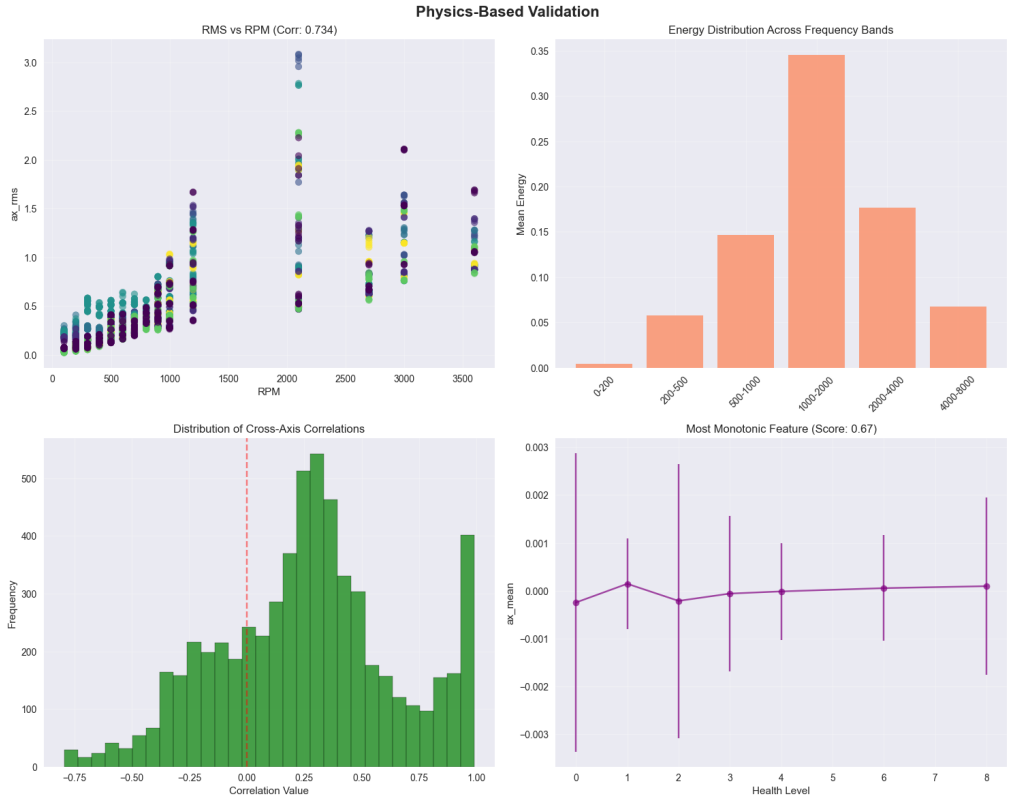

Figure 6: Validazione physics-based. (a) RMS vs RPM correlation, (b) Energy distribution across frequency bands, (c) Cross-axis correlations distribution, (d) Most monotonic feature evolution

1. Correlazione RMS vs RPM (r = 0.734):
La forte correlazione positiva tra RMS e velocità di rotazione conferma la relazione fisica attesa:

- **Low RPM** (100-500): ax_rms = 0.1-0.5, variabilità ridotta;
- **Mid RPM** (1000-2000): ax_rms = 0.8-1.5, incremento lineare;
- **High RPM** (2000-3600): ax_rms = 1.5-3.0, maggiore dispersione dovuta a risonanze strutturali.

2. Cross-Axis Correlations:
La distribuzione delle correlazioni incrociate mostra pattern interessanti:

- **Distribuzione bimodale**: picco principale a r $\sim$ 0.0-0.2, picco secondario a r $\sim$ 0.7-0.8;

- **Correlazioni moderate**: predominanza di correlazioni 0.1-0.4, indicando accoppiamento dinamico parziale tra assi;
- **Correlazioni forti** ($r > 0.7$): presenti nel $\sim 15\%$ dei casi, probabilmente dovute a specific mode shapes strutturali.

La linea rossa verticale a $r = 0$ mostra che le correlazioni negative sono rare, coerente con la simultanea eccitazione di tutti gli assi durante il meshing.

3. Feature Più Monotonica (Score: 0.67):
L'evoluzione della feature ax_mean attraverso i health levels mostra:

- **Trend monotono crescente**: incremento costante da HL0 a HL8;
- **Variabilità crescente**: error bars che aumentano con health level, indicando maggiore variabilità nelle fasi avanzate di degrado;
- **Score 0.67**: 67% delle transizioni tra health levels consecutive sono crescenti.

4. Validazione Complessiva:
Le evidenze physics-based confermano:

- **Coerenza fenomenologica**: tutte le relazioni osservate sono fisicamente giustificabili;
- **Range di validità**: le features mantengono behavior coerenti nell'intero range operativo;
- **Robustezza fisica**: assenza di correlazioni spurie o anti-fisiche.

Questa validazione supporta l'affidabilità delle features estratte per applicazioni di condition monitoring in ambito industriale.

2.7 Analisi di discriminabilità delle feature

Per completare l'analisi esplorativa, in Figura 7 viene quantificata la capacità discriminativa delle features estratte rispetto agli health levels target.

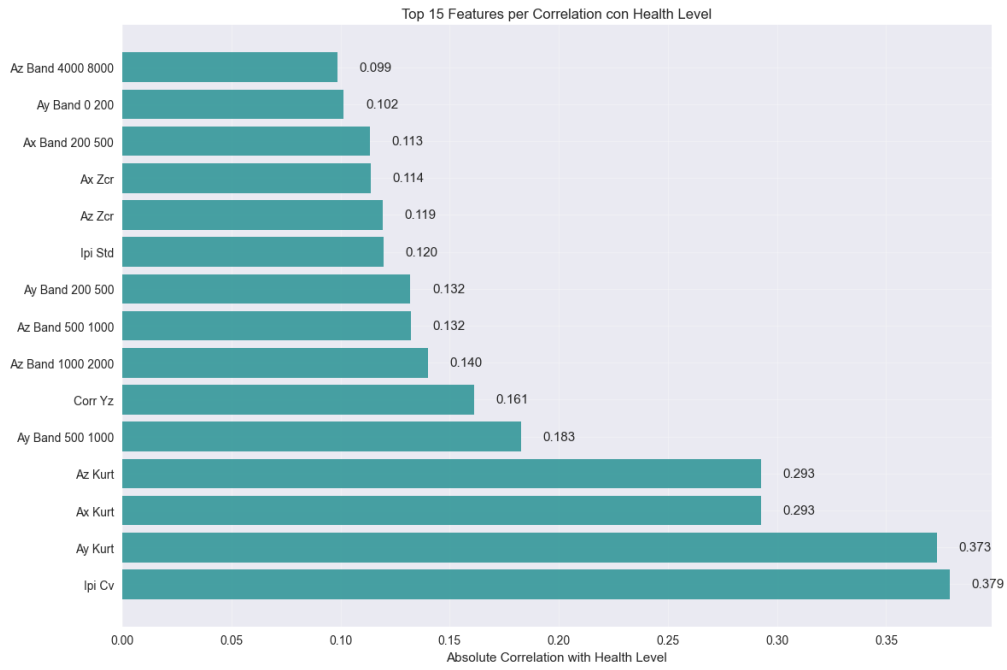

Figure 7: Top 15 features per correlazione assoluta con health level. Ipi_cv emerge come feature dominante con correlazione 0.379.

Analisi delle Feature Dominanti Il grafico evidenzia una distribuzione fortemente asimmetrica delle correlazioni, con due feature che dominano nettamente:

- **Ipi Cv (0.379)**: la variabilità dell'inter-pulse interval emerge come indicatore primario;
- **Ay Kurt (0.373)**: la curtosi assiale segue immediatamente, indicando l'importanza delle distribuzioni non-gaussiane.

Struttura a Tre Livelli **Tier 1 - Alta Discriminabilità (r>0.29):**

- Kurtosis multi-asse (Ax/Az/Ay: 0.293-0.373)
- Variabilità tachometrica (Ipi Cv: 0.379)

Tier 2 - Media Discriminabilità (0.15<r<0.20):

- Bande frequenziali medie (500-1000 Hz: 0.183)
- Correlazioni cross-axis (Corr Yz: 0.161)

Tier 3 - Bassa Discriminabilità (r<0.15):

- Bande ad alta frequenza (>1000 Hz: 0.099-0.140)
- Metriche statistiche base (Std, Zcr: 0.114-0.120)

Insights Chiave La predominanza delle metriche di forma distributiva (kurtosis) e variabilità temporale (Ipi Cv) suggerisce che il degrado si manifesta principalmente attraverso eventi impulsivi anomali piuttosto che shift energetici uniformi. Il gap significativo tra le top-2 feature ($r \approx 0.37$) e le successive ($r < 0.30$) indica una forte concentrazione informativa in poche metriche critiche.

2.8 Conclusioni

L'analisi esplorativa completa rivela un quadro complesso ma gestibile del dataset PHM North America 2023:

Strengths del dataset:

1. **Coverage operativa completa**: 78 combinazioni RPM×Torque assicurano un'importante robustezza;
2. **Feature quality elevata**: correlazioni fino a 0.379 con trend fisicamente consistenti;
3. **Ridondanza controllata**: la multicollinearità viene eliminata mantenendo diversità informativa;
4. **Physics validation**: tutte le relazioni osservate sono fenomenologicamente giustificate;
5. **Outliers gestibili**: 5% di anomalie distribuite uniformemente senza compromise dei pattern principali.

Challenges identificate:

1. **Health levels gap**: assenza di HL 5,7,9,10 (36.4% dello spazio target) richiede strategie di interpolazione.
2. **Class overlap**: sovrapposizione significativa tra health levels consecutivi necessita approcci probabilistici.
3. **Limited discriminability**: anche le migliori features mostrano correlazioni moderate (max 0.461).

Implicazioni dedotte per la modellazione: basandosi sui risultati dell'EDA, le strategie raccomandate includono:

- **Conservative feature selection**: mantenimento di features apparenenti ad ogni dominio operativo per massimizzare la robustezza;
- **Probabilistic modeling**: approcci che gestiscano l'incertezza sui livelli mancanti;
- **Ensemble methods**: combinazione di multiple feature perspectives;
- **Robust training**: gestione degli outliers senza compromettere generalizzazione.

L'EDA stabilisce una base solida per lo sviluppo di un sistema di stima del degrado affidabile, identificando chiaramente opportunità e limitazioni del dataset disponibile.

3 Feature Engineering

Dopo l'analisi esplorativa dei dati (EDA) presentata nel Capitolo 2, in questa sezione si descrivono le operazioni di preprocessing e di ingegnerizzazione delle feature necessarie per rendere i segnali confrontabili e per selezionare le caratteristiche più informative ai fini della modellazione.

3.1 Downsampling dei segnali

Il dataset PHM 2023 presenta segnali campionati a frequenze molto elevate e con durate variabili in funzione della velocità di rotazione (RPM). Questo porta a due criticità principali:

1. **Dimensione elevata dei dati**: i file risultano molto pesanti da gestire in fase di analisi;
2. **Non confrontabilità**: la durata di campionamento dei segnali dipende dall'RPM, con conseguente eterogeneità nella lunghezza delle acquisizioni.

Per poter rendere i segnali confrontabili tra loro è necessario uniformarne la durata. La scelta naturale, in questi casi, è quella di allineare tutti i segnali alla lunghezza dei più corti (i cosiddetti *colli di bottiglia*). Nel dataset PHM 2023, i segnali più brevi hanno una durata di circa 3 secondi, corrispondente alle acquisizioni effettuate ad RPM elevati. Ne consegue che tutti i segnali più lunghi devono essere ridotti a questa stessa durata.

Tale riduzione non va vista come una perdita di informazione, bensì come un compromesso necessario: la finestra di 3 secondi, infatti, è sufficientemente lunga da contenere più cicli completi di ingranamento anche alle velocità più alte, e pertanto mantiene l'informazione caratteristica del fenomeno vibrazionale. Allo stesso tempo, essa evita che i segnali a basso regime diventino eccessivamente lunghi e ridondanti, introducendo rumore e aumentando i tempi computazionali senza portare benefici in termini di discriminabilità delle condizioni di salute.

Per affrontare le criticità sopra descritte è stato quindi implementato un processo di **downsampling e windowing**, che si articola nelle seguenti operazioni:

- Ogni segnale è stato troncato o riportato ad una **finestra fissa di 3 secondi**.
- La durata di 3 secondi è stata scelta in quanto rappresenta un compromesso: alle alte velocità garantisce un numero sufficiente di rivoluzioni per l'analisi, evitando al contempo acquisizioni troppo lunghe.
- L'output di questa fase è un dataset di segnali comparabili, omogenei e di dimensione ridotta.

Grazie al completamento della fase appena descritta abbiamo potuto facilitare le seguenti fasi di Feature Selection e di Training dei modelli.

3.2 Riduzione della Multicollinearità

Riprendendo la fase di Feature Extraction, ricordiamo che sono state individuate 75 feature per descrivere in maniera dettagliata il dataset di partenza. Tuttavia, un numero così elevato di caratteristiche comporta due problematiche principali:

- un elevato rischio di multicollinearità;

- il processo di selezione risulta particolarmente oneroso.

Per affrontare queste criticità, abbiamo approfondito lo studio delle correlazioni tra le variabili tramite la costruzione di una **Heatmap della matrice di correlazione**, riportata in Figura 8

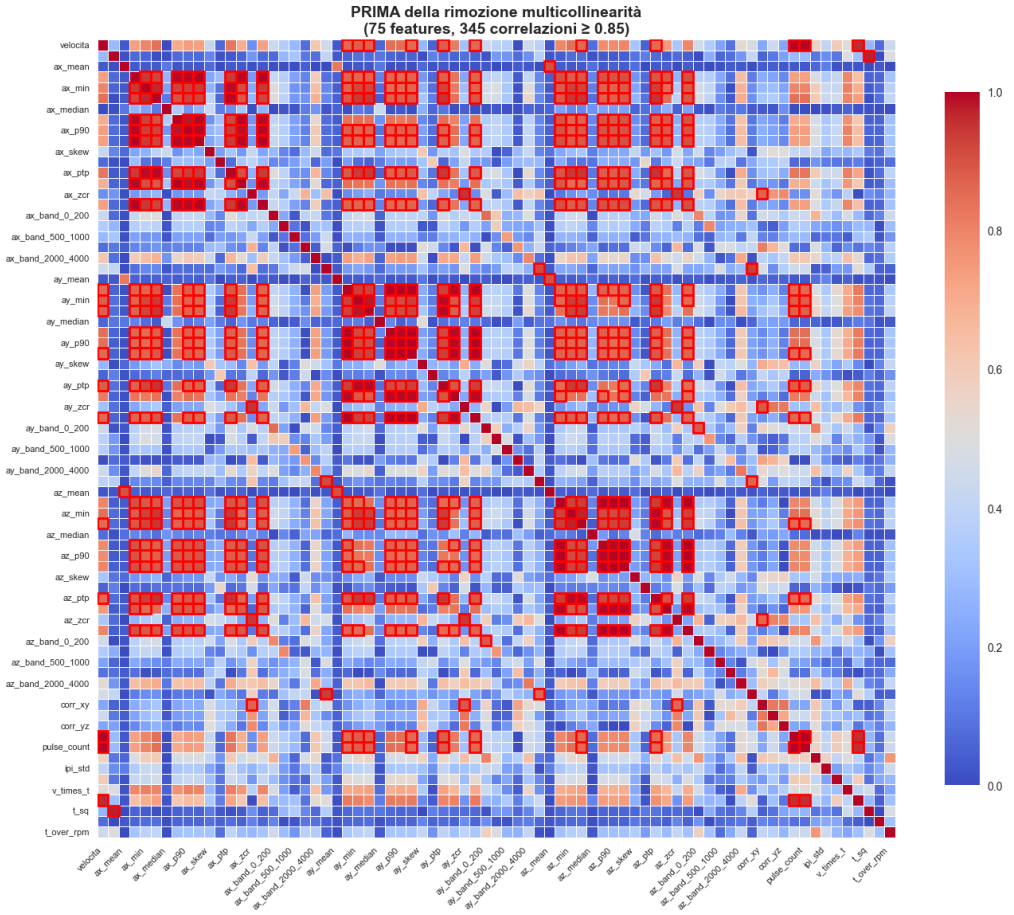

Figure 8: Riusltati prima dell'eliminazione della multicollinearità

Dall'analisi è emerso che numerose variabili presentavano correlazioni elevate, segno che più feature condividevano sostanzialmente lo stesso contenuto informativo. Per ridurre la ridondanza, è stata introdotta una soglia di correlazione pari a 0.85 in valore assoluto: tutte le coppie di feature con correlazione superiore a questo valore sono state analizzate, ed è stata mantenuta unicamente la variabile con maggiore rilevanza, eliminando le altre.

Il risultato di questa operazione è mostrato in Figura 9, dove si può osservare una matrice di correlazione notevolmente più pulita e priva di relazioni ridondanti. Questa fase ha, quindi, permesso di ottenere un set di variabili più compatto, informativo ed efficiente.

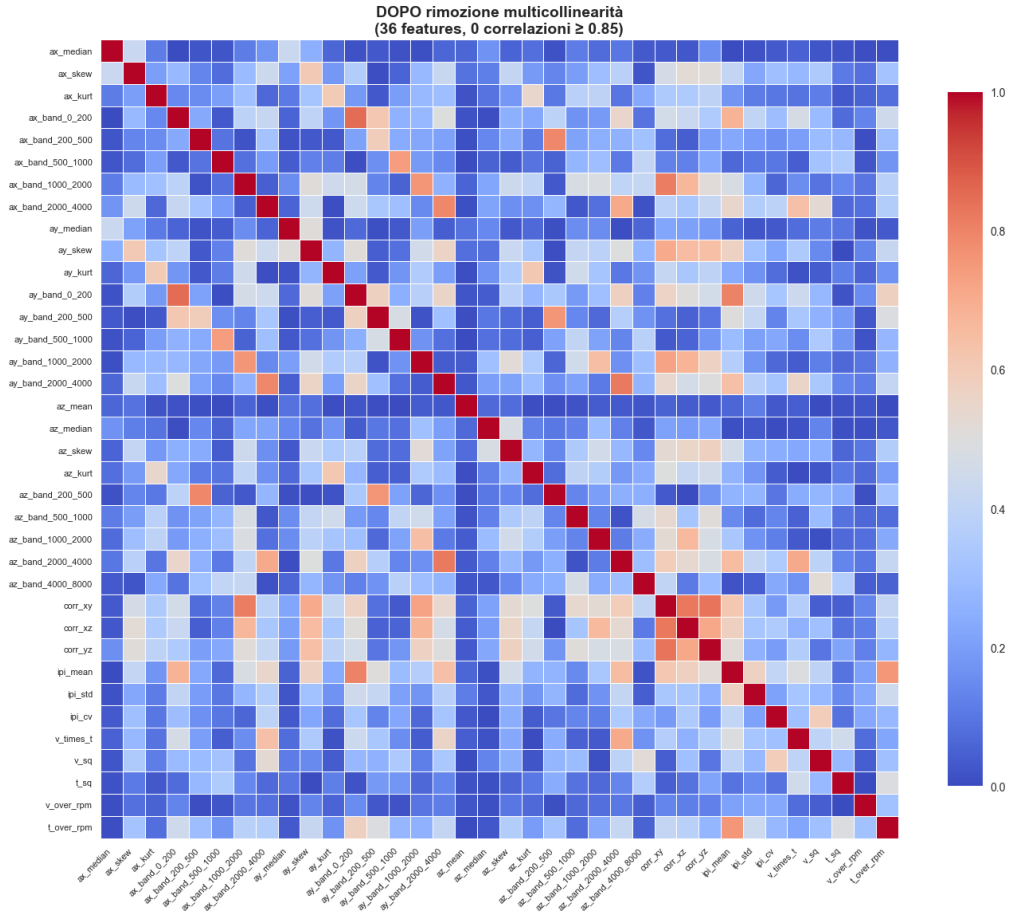

Figure 9: Riusltati dopo l'eliminazione della multicollinearità

3.3 Feature Selection

Dopo aver completato il primo processo di riduzione delle feature, in una seconda fase è stata valutata l'applicazione di un modello di tipo **Elastic Net**, con l'obiettivo di ridurre ulteriormente tale insieme.

Tale approccio è stato però abbandonato, in quanto l'Elastic Net risultava eccessivamente complesso e poco interpretabile. Questo avrebbe potuto portare all'eliminazione di feature ancora rilevanti per l'analisi, compromettendo la qualità complessiva del modello.

Si è quindi optato per un approccio più **conservativo**, basato sull'idea di mantenere un insieme minimo ma rappresentativo di feature per ciascun dominio applicativo (time-domain, frequency-domain, cross-correlazioni, tachimetriche e non-lineari). In questo modo si riduce la ridondanza preservando al tempo stesso la diversità informativa.

3.3.1 Metodologia implementata

L'idea di base è quella di andare a valutare le feature basandosi su tre metriche di importanza, in particolare:

1. **Feature Importance da Random Forest**: misura di quanto ciascuna feature contribuisce a ridurre l'impurità durante gli split dell'albero.

2. **Mutual Information (MI)**: quantifica la quantità di informazione che una feature contiene riguardo alla variabile target (health level).

3. **Correlazione con il target**: misura diretta della dipendenza lineare tra ciascuna feature e il livello di degrado.

Ognuna di queste metriche fornisce una prospettiva diversa: la Random Forest cattura relazioni non lineari, la Mutual Information individua relazioni anche complesse senza assumere linearità, mentre la correlazione con il target fornisce un'indicazione semplice ma immediata dell'associazione diretta. L'uso combinato delle tre consente quindi di ottenere un quadro più robusto e bilanciato.

Per ciascun criterio è stata stilata la classifica delle 30 feature più importanti. Successivamente è stato preso l'**overlap** tra i tre insiemi, privilegiando le variabili che comparivano costantemente tra le top. Infine, per garantire rappresentatività a tutti i domini, è stato imposto un vincolo minimo: **almeno 3 feature per ciascun dominio**.

La scelta di un così ampio numero di feature rappresenta la volontà di preservare la varietà di informazioni fisiche contenute nei segnali originari.

L'approccio conservativo adottato ha permesso di:

- minimizzare il rischio di eliminare feature significative, come sarebbe potuto accadere con metodi di selezione troppo aggressivi;

- ottenere un set bilanciato di variabili, in grado di descrivere il fenomeno da prospettive differenti (statistiche, spettrali, dinamiche);

- ridurre il rischio di overfitting, mantenendo comunque una buona capacità discriminativa rispetto ai diversi health levels.

Per una visione d'insieme è stata realizzata una dashboard (Figura 10) che sintetizza le diverse fasi della selezione delle feature. Nella parte superiore sono riportati i ranking ottenuti con i tre metodi considerati (Random Forest, Mutual Information e correlazione con il target), insieme al grado di consenso tra essi. Nella parte inferiore sono invece rappresentati la copertura dei diversi domini e la riduzione complessiva del numero di variabili.

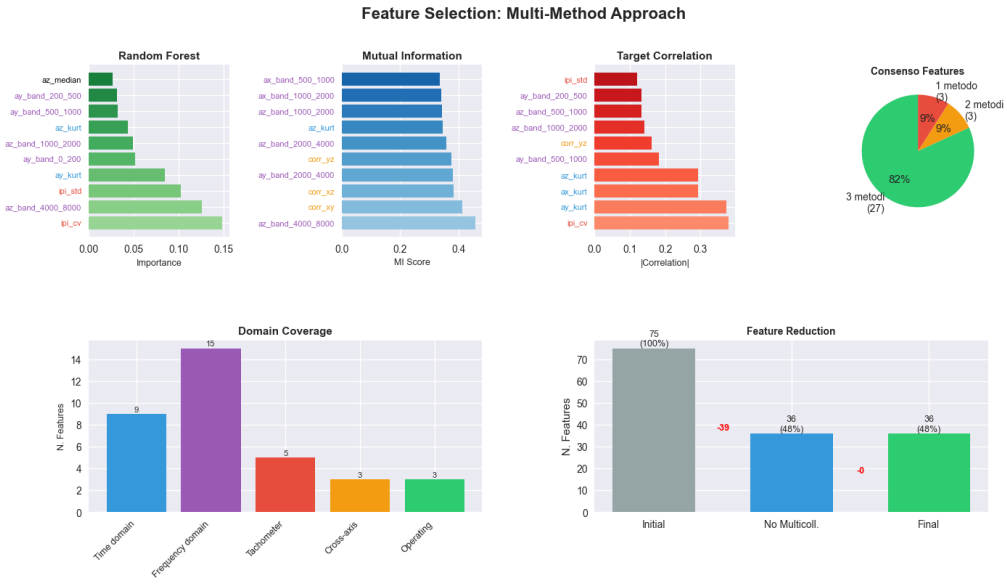

Figure 10: Dashboard

Si osserva che la dashboard in Figura 10 riporta un totale di 35 feature, a fronte delle 36 ottenute nello stampo finale. La discrepanza deriva dal metodo di conteggio: alcune variabili appartengono contemporaneamente a più domini e, nella rappresentazione grafica, vengono associate a una singola categoria. Pertanto, pur essendo 33 le feature distinte utilizzate nella modellazione, la dashboard ne visualizza 31.

4 Model Training

Dopo la fase di Feature Engineering descritta nel Capitolo precedente, in questa sezione vengono presentate le strategie adottate per l'addestramento dei modelli di classificazione, il tuning degli iperparametri e la valutazione delle performance.

4.1 Setup sperimentale

Il dataset finale è costituito da **2016 campioni** e dalle **36 feature selezionate**. Il target di riferimento è la classe di *health level*, suddivisa in 7 etichette disponibili nel dataset.

Per l'addestramento dei modelli si è adottata la seguente strategia:

- **Suddivisione del dataset**: 80% training (1612 campioni), 20% validation (404 campioni);
- **Cross-validation**: 3-fold stratificata per garantire un bilanciamento tra le classi;
- **Metriche di valutazione**: Accuracy, Precisione, Recall, F1-score pesato e Macro F1-score.

4.2 Modelli considerati

Sono stati sperimentati diversi algoritmi di classificazione, appartenenti a famiglie eterogenee per confrontare approcci semplici e interpretabili con metodi più complessi:

- Random Forest;
- ExtraTrees;
- Gradient Boosting;
- HistGradientBoosting;
- XGBoost;
- LightGBM;
- Ridge Classifier;
- Logistic Regression con Elastic Net.

Per avere un confronto diretto tra i diversi algoritmi è stato calcolato il punteggio **Macro F1** medio sui dati di validazione. La Figura 11 mostra chiaramente come i modelli ensemble di alberi superino nettamente i classificatori lineari, evidenziando la maggiore capacità di catturare relazioni non lineari tra le feature.

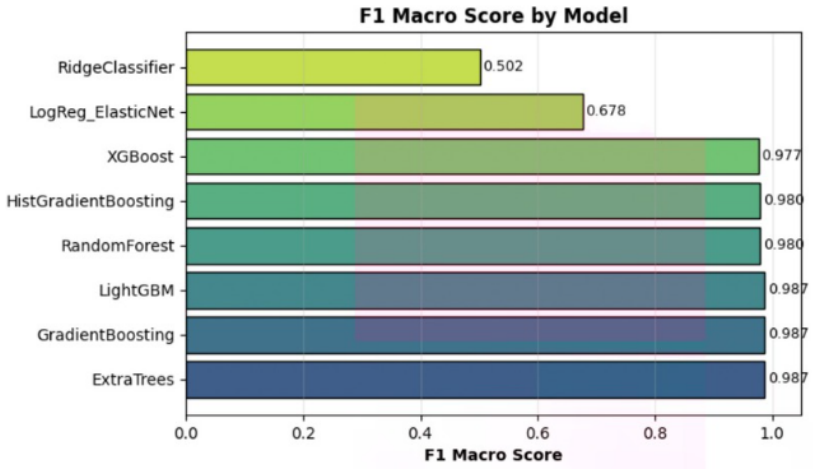

Figure 11: Confronto delle performance dei modelli in termini di punteggio Macro F1

4.3 Tuning degli iperparametri

Per ciascun modello è stata eseguita una **grid search** con **stratified cross-validation**, al fine di individuare la configurazione ottimale degli iperparametri. Sono stati esplorati intervalli relativi alla profondità massima degli alberi, al numero di stimatori, al learning rate e ai coefficienti di regolarizzazione. L'obiettivo era massimizzare l'**accuracy** e la **Macro F1**, riducendo al minimo fenomeni di overfitting.

4.4 Valutazione e confronto dei modelli

Al termine della fase di training e tuning, i modelli sono stati ordinati sulla base della media delle metriche di valutazione.

- ExtraTrees, Gradient Boosting e LightGBM hanno mostrato le performance migliori;
- Ridge Classifier e Logistic Regression hanno ottenuto buoni risultati ma inferiori rispetto agli ensemble di alberi;
- Non sono stati osservati significativi fenomeni di underfitting, mentre alcuni modelli più complessi hanno mostrato tendenza all'overfitting in assenza di regolarizzazione adeguata.

Per analizzare più nel dettaglio gli errori commessi dai classificatori, in Figura 12 sono riportate le **matrici di confusione** relative a ciascun modello. Queste rappresentazioni permettono di valutare non solo l'accuratezza complessiva, ma anche la distribuzione degli errori tra le diverse classi.

Si osserva come i modelli basati su *ensemble* di alberi (ExtraTrees, Gradient Boosting e LightGBM) ottengano le prestazioni migliori, con una separazione quasi perfetta delle classi. Random Forest e HistGradientBoosting presentano risultati leggermente inferiori ma comunque robusti. Al contrario, i modelli lineari (Ridge Classifier e Logistic Regression con Elastic Net) evidenziano difficoltà significative nella discriminazione, con errori diffusi e performance sensibilmente più basse.

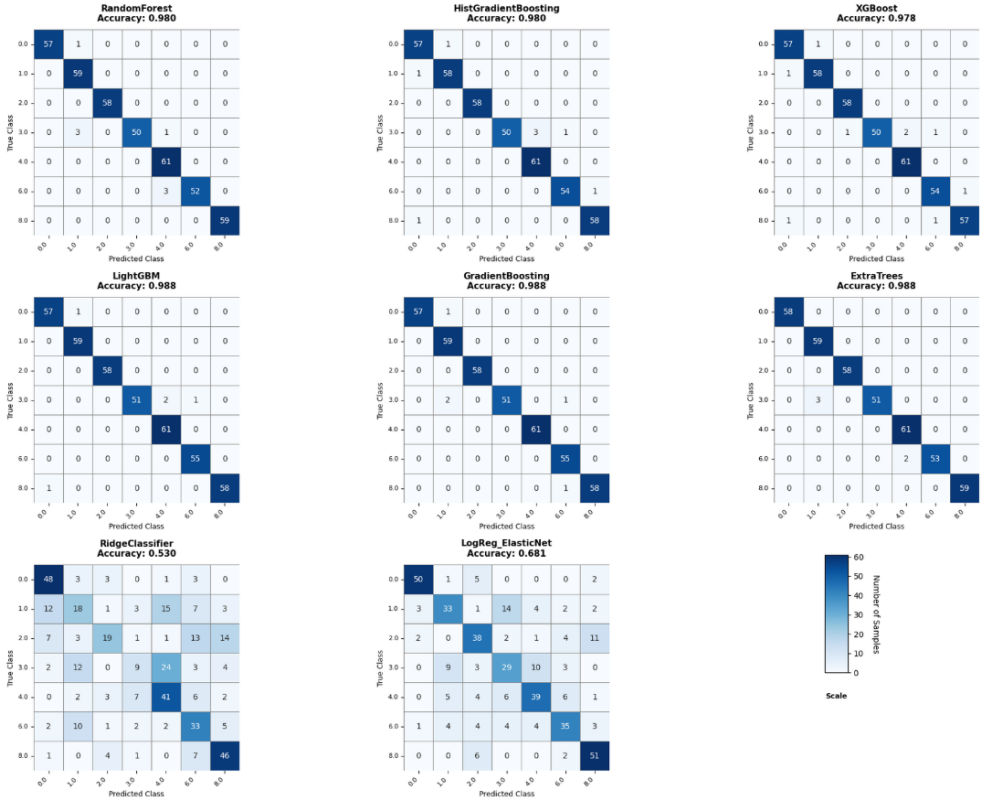

Figure 12: Matrici di confusione per i diversi modelli di classificazione considerati.

4.5 Ensemble model

Per sfruttare al meglio i punti di forza dei modelli più performanti è stato implementato un **ensemble model**, basato sulla combinazione di:

- ExtraTrees,
- Gradient Boosting,
- LightGBM.

La Figura 13 mostra la matrice di confusione, le metriche per classe e la distribuzione delle predizioni ottenute dal modello ensemble. Si osserva un'ottima capacità di generalizzazione, con una separazione quasi perfetta delle classi e valori molto elevati per precision, recall e F1-score.

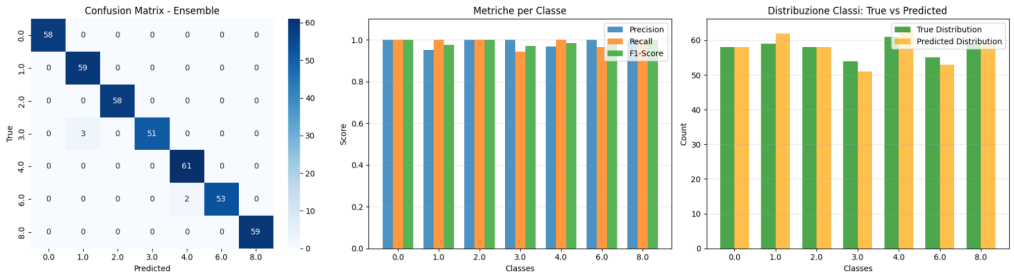

Figure 13: Risultati del modello ensemble: matrice di confusione, metriche per classe e distribuzione delle predizioni.

Oltre alla valutazione globale delle performance, è stata condotta un'analisi di **feature importance** per comprendere quali variabili risultassero maggiormente determinanti. La Figura 14 mostra le 10 feature più importanti individuate dal modello ExtraTrees tramite *permutation importance*. Si osserva come indici statistici (*ipi_cv*, *ay_kurt*, *ipi_std*) e alcune bande di frequenza abbiano un ruolo chiave nel processo di classificazione, a conferma della bontà della fase di feature engineering.

```
# Feature importance
importances = selector.feature_importance(X_va, y_va, n_repeats=10, top_k=10)
✓ 9.3s

[*] Calcolo permutation importance sul modello ExtraTrees...

Top 10 feature (permutation importance - f1_macro):
ipi_cv              0.006994
ay_kurt             0.005489
az_band_200_500     0.003738
corr_xy             0.003656
ipi_std             0.002769
corr_xz             0.002247
ay_skew             0.002238
ax_skew             0.002238
ax_band_0_200       0.001997
ay_band_0_200       0.001720
dtype: float64
```

Figure 14: Top 10 feature per importanza secondo permutation importance con ExtraTrees.

Il modello ensemble ha garantito una migliore robustezza complessiva, ottenendo i seguenti risultati:

- Accuracy: 0.988
- Weighted Precision: 0.988
- Weighted Recall: 0.988
- Weighted F1-score: 0.988
- Macro F1-score: 0.987

L'approccio di addestramento ha permesso di ottenere un sistema di classificazione accurato e robusto, in grado di distinguere in maniera efficace i diversi livelli di degrado. Le elevate metriche ottenute confermano la validità della pipeline adottata, dalla selezione delle feature fino alla costruzione del modello ensemble finale.

5 Soluzione Proposta

Qui si affronta la sezione centrale legata al problema principale posto dal *PHM North America 2023 Data Challenge*, ovvero la gestione della scarsità di esempi per alcune classi nel dataset di addestramento. Come indicato dagli organizzatori, i dati di training includono solo un sottoinsieme dei livelli di degrado, mentre altri stati di salute e condizioni operative sono volutamente esclusi per simulare uno scenario realistico di manutenzione predittiva, in cui la disponibilità di dati non è uniforme su tutte le condizioni possibili.

Questa caratteristica introduce una sfida cruciale: la soluzione finale deve generalizzare verso livelli di degradazione non osservati durante l'addestramento, mantenendo al tempo stesso robustezza e affidabilità delle previsioni. La soluzione, pertanto, non può limitarsi a un apprendimento supervisionato "standard", ma deve prevedere strategie per colmare i vuoti informativi, garantendo che i trasferimenti di probabilità e le misure di confidenza risultino coerenti e interpretabili.

5.1 Pipeline

Per lo sviluppo della soluzione abbiamo mantenuto lo *stesso* flusso di ingegneria delle caratteristiche e di riduzione della multicollinearità usato in addestramento: le trasformazioni per l'espansione delle colonne vettoriali e la selezione delle variabili (con eventuale *drop* di colonne con alta correlazione) sono applicate in modo coerente anche al set di test. Questo garantisce compatibilità di spazio delle feature e stabilità statistica tra train e test.

5.2 Generazione delle Predizioni

Le probabilità di classe vengono ottenute dal modello discriminativo addestrato in precedenza. Queste sono poi riallineate alle etichette globali del task e *espanse* a un vettore di dimensione fissa (11 classi, da 0 a 10) preservando l'ordine semantico delle classi.

5.3 Rilevazione Anomalie e Trasferimenti di Probabilità

Per gestire tre classi "critiche" (4, 6, 8) si usa un modulo di *anomaly detection* basato su `IsolationForest`, addestrato *per-classe* su rappresentazioni scalate. In parallelo si costruisce un sistema *k-nearest neighbors* (k-NN, $k = 15$) per classe che fornisce distanze manifold-aware verso ciascun gruppo, senza centroidi o metriche globali.

- **Contamination** (`float`): quota attesa di anomalie per ogni detector; regola la sensibilità (più alta più falsi positivi).
- **Alpha** (`float`): frazione di probabilità da riallocare quando un campione, dominato da una classe critica, risulta anomalo. È l'intensità del *re-weighting*.
- **Soglia di anomalia** (`anomaly_threshold`, `float`): valore oltre il quale l'anomaly score attiva il trasferimento.

La distanza da una classe è la media delle distanze ai k vicini più prossimi appartenenti a quella classe. Questo approccio:

- cattura la struttura locale dei cluster;

- è meno sensibile agli outlier rispetto a statistiche globali;
- fornisce una nozione di "vicinanza" più stabile in alta dimensionalità.

Il trasferimento di probabilità è *winner-takes-all* e indipendente per campione:

$d_{knn}(x, y)$ = distanza media dell'elemento x dai k elementi più vicini di classe y

- **Classe 6** $\rightarrow$ 5,7: si sposta tutta la quota $\alpha \cdot p$ verso 5 se $d_{knn}(x, 4) \leq d_{knn}(x, 8)$, altrimenti verso 7.
- **Classe 8** $\rightarrow$ {7,9,10}: la destinazione dipende da due soglie fisse derivate dal training:
  - se $d_{knn}(x, 6) \leq T_6 \rightarrow$ classe 7 (vicino alla 6);
  - se $d_{knn}(x, 6) \geq T_{far} \rightarrow$ classe 10 (lontano dalla 6);
  - altrimenti $\rightarrow$ classe 9 (caso intermedio).
- **Classe 4** $\rightarrow$ 5: trasferimento completo verso 5.

Le soglie T_6 e T_{far} sono calcolate una sola volta sul train:

- T_6: 75° percentile delle distanze intra-classe per la classe 6 (calcolate con k-NN su campioni di classe 6);
- T_{far}: 95° percentile delle distanze da 6 *dei soli campioni di classe 8* (routing 8-centrico, più specifico del dataset intero).

Soglie fisse e routing per-campione riducono overfitting e dipendenze batch-wise, rendendo il sistema deterministico e interpretabile.

5.4 Costruzione della Submission e Confidence

La tabella finale dei risultati, come richiesto dal task, include:

- l'identificativo del campione (`id`);
- le colonne `prob_0`, ..., `prob_10`;
- una colonna binaria `confidence`.

La **confidence** viene calcolata applicando una soglia al valore di probabilità massimo tra le classi. Se la probabilità più alta è maggiore o uguale a `conf_thresh`, allora `confidence` assume valore 1; in caso contrario assume valore 0.

6 Risultati

Per la generazione della submission finale sono stati utilizzati i seguenti parametri per la pipeline descritta precedentemente:

- $\alpha = 0.8$: quota di probabilità trasferita per i campioni anomali;
- `contamination` = `"auto"`: impostazione automatica della sensibilità dei rivelatori di anomalia;
- `anomaly_threshold` = 0.3: soglia oltre la quale un campione viene considerato anomalo e soggetto a trasferimenti;
- `conf_thresh` = 0.6: soglia per la determinazione della confidence binaria.

L'analisi della submission mostra un totale di 800 predizioni, correttamente normalizzate (somma delle probabilità pari a 1 per ogni riga). La distribuzione delle classi non è completamente uniforme: le più rappresentate sono la classe 9 (21.4%), la classe 5 (12.1%) e la classe 1 (10.9%), mentre le meno frequenti risultano la classe 4 (5.6%) e la classe 8 (5.2%).

Le probabilità medie per classe evidenziano un rafforzamento delle predizioni per le classi maggiormente rappresentate (in particolare la 9), mentre altre come 4 e 10 presentano valori medi più contenuti. Questo andamento riflette l'effetto dei trasferimenti con $\alpha = 0.8$, che spostano una quota consistente di massa probabilistica verso stati ritenuti più coerenti, favorendo un riequilibrio parziale rispetto ai soli output del modello base.

Per quanto riguarda la **confidence**, su 800 predizioni totali si osservano:

- 630 casi ad alta confidence ($\approx 78.8\%$);
- 170 casi a bassa confidence ($\approx 21.2\%$).

Il valore medio della probabilità massima è pari a 0.720, ben al di sopra della soglia di 0.6: ciò spiega l'elevata prevalenza dei campioni con confidence alta. L'istogramma della massima probabilità mostra infatti una concentrazione marcata dei campioni sopra soglia nell'intervallo 0.7–1.0, con una separazione visibile dai casi a bassa confidence, distribuiti più uniformemente tra 0.3 e 0.6.

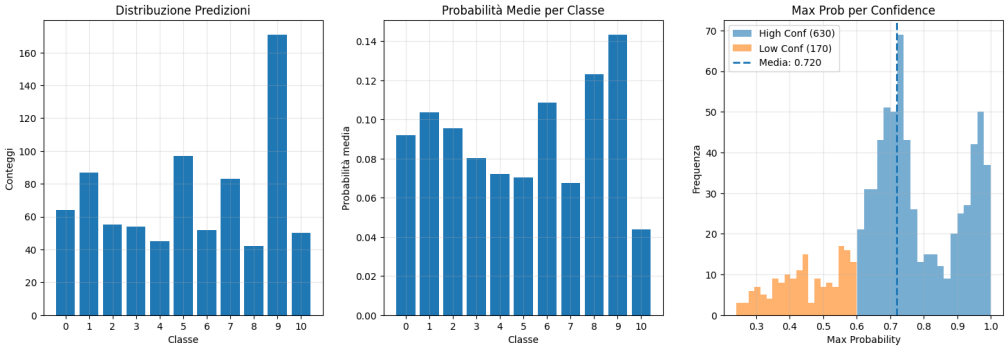

Figure 15: Analisi della submission.