



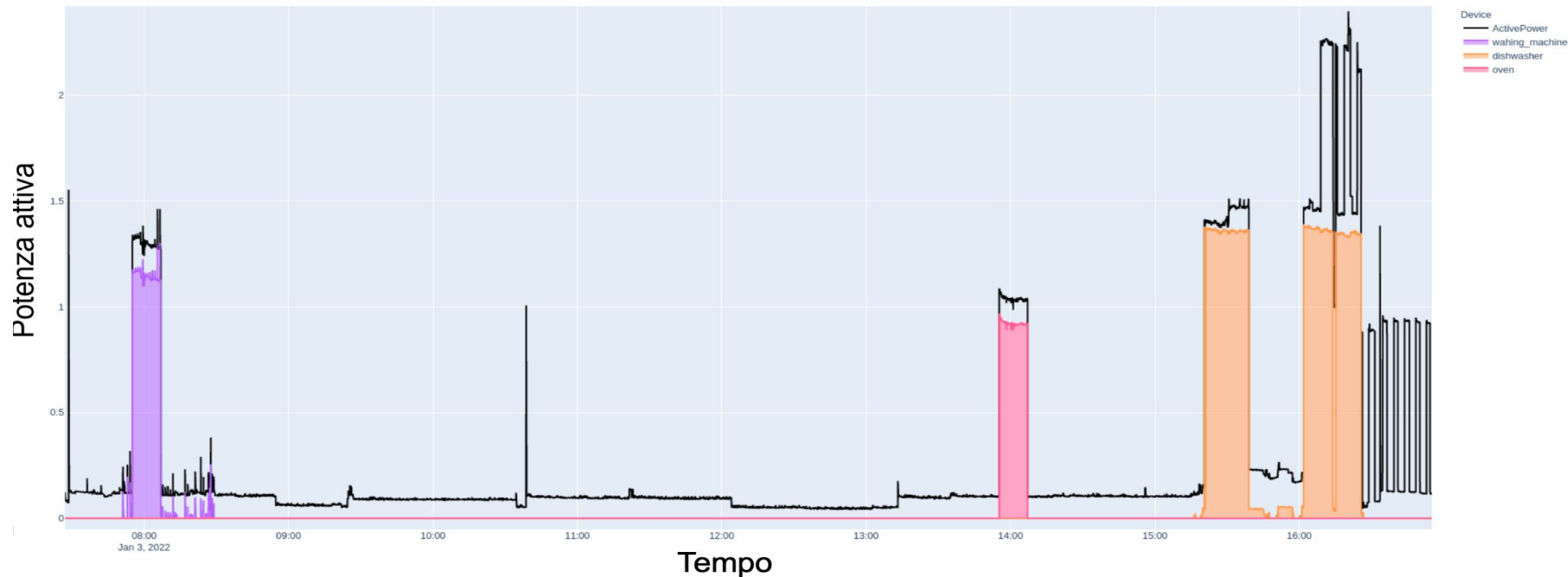
Non Intrusive Load Monitoring (NILM)

Team Bravo



Presentazione problema

Il Non-Intrusive Load Monitoring (NILM) è una sfida che si pone l'obiettivo di disaggregare i singoli dispositivi elettrici da un segnale di consumo aggregato.





Presentazione problema

Il NILM si sviluppa principalmente con l'obiettivo di evitare l'installazione di numerosi dispositivi di monitoraggio per ciascun apparato, contribuendo così a ridurre i costi associati al monitoraggio per proprietari di abitazioni o gestori di uffici.

Inizialmente complesso da utilizzare, il NILM è ora oggetto di interesse da parte di diverse società e gruppi di ricerca, dando luogo a dispositivi commerciali e numerose pubblicazioni scientifiche. Il progetto NILM mira a analizzare i consumi energetici aggregati di un'utenza, identificando i dispositivi attivi e calcolando i consumi associati. Utilizzando un sensore sul contatore, il NILM registra i consumi in un database e, attraverso sofisticati algoritmi, riconosce le firme energetiche uniche di ciascun dispositivo installato, separandole dal consumo energetico totale.

Per le aziende, tali dati sono cruciali per comprendere il comportamento dei consumatori, migliorare la pianificazione energetica e offrire servizi personalizzati.

Per gli utenti finali, l'utilità risiede nella capacità di monitorare e ottimizzare i consumi energetici dei propri dispositivi, identificando e gestendo quelli più inefficienti. Inoltre, l'accesso a informazioni dettagliate consente agli utenti di prendere decisioni più informate sulla sostituzione di dispositivi e di sfruttare tariffe energetiche variabili per ottimizzare i costi. Il tutto è reso visibile attraverso un cruscotto di sintesi, che fornisce sia una visione d'insieme che dettagli specifici a livello di singolo dispositivo.



Dataset

Il NILM per una serie di fattori non prevedono un dataset con labels, e quindi va risolto in maniera non supervisionata, nel nostro caso però i dati vengono semplificati e quindi prevedono anche delle colonne dedicate all'accensione dei dispositivi rendendo così il problema risolvibile in maniera supervisionata.

In particolare i dataset forniti rappresentano 105 giorni di consumo effettivo registrati da un sensore installato in un appartamento. Per ognuno dei giorni in esame vengono indicati, tramite opportune features:

- Le informazioni sui consumi nella loro totalità;
- Il consumo specifico registrato relativamente a tre device: lavatrice, lavastoviglie e forno.



Record

I dati sono campionati al secondo, quindi, per ogni giorno si hanno a disposizione 9072000 record, relativi a tutti i secondi della giornata. Ogni riga del dataset, quindi, rappresenta le diverse ampiezze relative ai consumi energetici registrati dal sensore sull'intero appartamento e sui singoli dispositivi presi in esame.

	DateTime	ActivePower	ReactivePower	Voltage	Current	harmonic1_Real	harmonic1_Imaginary	harmonic3_Real	harmonic3_Imaginary	harmonic5_Real	harmonic5_Imaginary
0	2023-01-01 00:00:00	66116.00000	67088.31250	237.00000	279.00000	13.00000	-300.00000	17.00000	-55.00000	-16.00000	-21.00000
1	2023-01-01 00:00:01	66561.00000	68342.31250	237.00000	281.00000	13.00000	-302.00000	17.00000	-55.00000	-16.00000	-22.00000
2	2023-01-01 00:00:02	66107.00000	67583.31250	237.00000	279.00000	13.00000	-301.00000	16.00000	-55.00000	-16.00000	-22.00000
3	2023-01-01 00:00:03	66309.00000	67101.31250	236.00000	281.00000	16.00000	-297.00000	17.00000	-54.00000	-15.00000	-22.00000
4	2023-01-01 00:00:04	66309.00000	65259.31250	236.00000	281.00000	15.00000	-300.00000	17.00000	-54.00000	-16.00000	-22.00000



Features

Le diverse features presenti nel dataset, che descrivono i consumi a livello sia totale che del singolo dispositivo considerato, sono:

- **DateTime**: data e orario dell'evento, quindi l'istante cui fanno riferimento i dati raccolti;
- **ActivePower**: potenza attiva consumata;
- **ReactivePower**: potenza reattiva;
- **Voltage**: tensione elettrica;
- **Current**: corrente erogata;
- **harmonic1_Real**: parte reale dell'armonica 1 registrata;
- **harmonic1_Imaginary**: parte immaginaria dell'armonica 1 registrata;
 - tali informazioni sono presenti anche per le armoniche 3, 5 e 7
- **washing_machine**: potenza attiva consumata dalla lavatrice in quell'istante di tempo;
- **dishwasher**: potenza attiva consumata dalla lavastoviglie in quell'istante di tempo;
- **oven**: potenza attiva consumata dal forno nell'istante di tempo di riferimento.



Preprocessamento dei dati

- Eliminare campioni duplicati controllando la presenza di righe con lo stesso timestamp e mantenendo solo una di esse.
- Verificare la presenza di lacune nel campionamento. Per farlo, si confronta la data di inizio e fine campionamento con la lunghezza teorica del dataset senza lacune, considerando un tasso di campionamento di 1 secondo. Se vengono individuate lacune, queste vengono colmate attraverso l'interpolazione lineare.



Preprocessamento dei dati

- Verificare l'assenza di valori nulli (NaN) nell'intero dataset; nel nostro caso, abbiamo constatato l'assenza totale di tali valori.
- Riorganizzazione degli indici del dataset: per migliorare la leggibilità e la gestione del dataset, abbiamo scelto di sostituire gli indici delle righe con i relativi timestamp.

	ActivePower	ReactivePower	Voltage	Current	harmonic1_Real	harmonic1_Imaginary	harmonic3_Real	harmonic3_Imaginary	harmonic5_Real	harmonic5_Imaginary
2023-01-01 00:00:00	66.11600	67.08831	237.00000	279.00000	13.00000	-300.00000	17.00000	-55.00000	-16.00000	-21.00000
2023-01-01 00:00:01	66.56100	68.34231	237.00000	281.00000	13.00000	-302.00000	17.00000	-55.00000	-16.00000	-22.00000
2023-01-01 00:00:02	66.10700	67.58331	237.00000	279.00000	13.00000	-301.00000	16.00000	-55.00000	-16.00000	-22.00000
2023-01-01 00:00:03	66.30900	67.10131	236.00000	281.00000	16.00000	-297.00000	17.00000	-54.00000	-15.00000	-22.00000
2023-01-01 00:00:04	66.30900	65.25931	236.00000	281.00000	15.00000	-300.00000	17.00000	-54.00000	-16.00000	-22.00000



Preprocessamento dei dati

Estrazione nuove features

Incrementare il numero di feature attraverso l'estrazione da quelle già esistenti può consentire di identificare nuovi comportamenti distintivi per ciascun dispositivo. In questo contesto, vengono estratte nuove feature temporali basate sul timestamp, tra cui:

- giorno della settimana (dayofweek)
- giorno dell'anno (dayofyear)
- anno (year)
- mese (month)
- trimestre (quarter)
- ora (hour)
- settimana dell'anno (weekofyear)
- giorno del mese (dayofmonth)
- data (date)
- stagione (season)



Preprocessamento dei dati

Preparazione dei dataset X e Y

Per la classificazione dei dispositivi, inizialmente si è scelto di convertire i valori di potenza attiva consumata dai singoli dispositivi in una variabile booleana. Di conseguenza, è stata creato un dataset (Y) con una colonna dedicata per ciascun dispositivo, in cui ogni dispositivo assume il valore 1 se la potenza attiva in quell'istante è superiore a 0, e 0 se la potenza attiva è pari a 0.

	wahing_machine	dishwasher	oven
2023-01-02 18:52:36	1	0	1
2023-01-02 18:52:37	1	0	1
2023-01-02 18:52:38	1	0	1
2023-01-02 18:52:39	1	0	1
2023-01-02 18:52:40	1	0	1



Preprocessamento dei dati

Preparazione dei dataset X e Y

A questo punto, le combinazioni dei valori booleani nelle 3 colonne saranno tradotte in un valore unico nella colonna finale denominata "label". In particolare, tale valore sarà ottenuto mediante la conversione delle combinazioni di valori booleani, che possono corrispondere a un numero binario, in un valore decimale. Nel dataset X invece andranno tutte le colonne tranne le colonne dedicate al labelling.

	wahing_machine	dishwasher	oven
2023-01-02 18:52:36	1	0	1
2023-01-02 18:52:37	1	0	1
2023-01-02 18:52:38	1	0	1
2023-01-02 18:52:39	1	0	1
2023-01-02 18:52:40	1	0	1











	label
2023-01-02 18:52:36	5
2023-01-02 18:52:37	5
2023-01-02 18:52:38	5
2023-01-02 18:52:39	5
2023-01-02 18:52:40	5



Classificazione

La colonna label rappresenta ogni stato di classificazione possibile:

- 0  Tutto spento (AO)
- 1  Forno acceso (OW)
- 2  Lavastoviglie acceso (DW)
- 3  Lavastoviglie e Forno accesi (DW_OW)
- 4  Lavatrice accesa (WM)
- 5  Lavatrice e Forno accesi (WM_OW)
- 6  Lavatrice e Lavastoviglie accese (WM_DW)
- 7  Tutto acceso (WM_DW_OW)



Classificazione

Algoritmo

Per selezionare l'algoritmo di classificazione ottimale, abbiamo principalmente esaminato tre algoritmi:

- **KNN**
- **Random Forest Classifier**
- **SVN**

Sui dati ricampionati ogni 10 secondi, il Random Forest Classifier ha dimostrato prestazioni superiori, pertanto abbiamo scelto di adottare questo approccio. Inoltre, il Random Forest Classifier presenta caratteristiche e funzionalità rilevanti per la risoluzione delle sfide legate al Non-Intrusive Load Monitoring (NILM).

Le prestazioni dell'algoritmo sono state testate verificando l'accuracy, la recall, l'f1 score e la tabella di confusione ottenuti da ogni algoritmo.

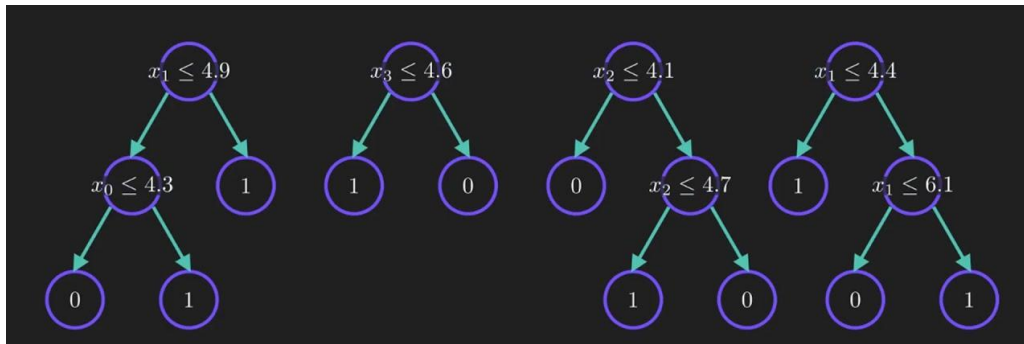


Classificazione

Random Forest Classifier

Il Random Forest Classifier è un potente algoritmo di apprendimento automatico che si basa sulla creazione di un insieme di alberi decisionali, noti come "forest". Ogni albero viene allenato su un sottoinsieme casuale dei dati di addestramento, e le predizioni finali sono ottenute mediante la votazione delle predizioni di ciascun albero.

Questo approccio rende il Random Forest Classifier particolarmente efficace nel gestire complessità e rumore nei dati. Inoltre, l'algoritmo offre diverse caratteristiche che si rivelano vantaggiose nella risoluzione delle sfide associate al Non-Intrusive Load Monitoring (NILM). Una delle ragioni principali per la scelta del Random Forest Classifier nel contesto del NILM è la sua capacità di gestire un elevato numero di feature. Nel monitoraggio energetico, i dati sono spesso caratterizzati da una vasta gamma di variabili, e il Random Forest riesce a sfruttare questa eterogeneità per ottenere predizioni accurate.





Classificazione

Random Forest Classifier

la feature importance fornita dal Random Forest è cruciale. Questa metrica consente di valutare l'impatto di ciascuna feature sull'efficacia del modello, aiutando a identificare quali aspetti dei dati sono più rilevanti nella distinzione tra diversi dispositivi e i loro comportamenti energetici.

Nel caso specifico del NILM, la gestione dei class weight rappresenta un ulteriore elemento chiave. Modificare i pesi delle classi consente di influenzare la fase di addestramento in modo da enfatizzare la rilevazione di situazioni meno frequenti ma altrettanto importanti. Ad esempio, assegnando un peso maggiore alle situazioni in cui tutti i dispositivi sono accesi rispetto a quelle in cui sono spenti, il modello sarà incentivato a prestare più attenzione a queste condizioni meno comuni ma significative.

I pesi assegnati ad ogni classe sono i seguenti:

- $AO = 0.3$
- $OW, DW, WM, DW_OW, WM_OW, DW_WM = 1$
- $DW_WM_OW = 3$



Classificazione

Prova Iniziale

Inizialmente abbiamo trainato il modello attraverso un set di training e un set di test, il set di testing è stato ottenuto ricavando il 30% dei dati dal set originale, dunque il set di training è stato ottenuto rimuovendo i dati presenti dal set di testing.

Inizialmente, abbiamo addestrato il nostro modello utilizzando un set di training e un set di test. Per creare il set di test, abbiamo estratto il 30% dei dati dal set originale, mentre il set di training è stato generato eliminando i dati presenti nel set di test. Nonostante il modello abbia raggiunto un'accuratezza molto elevata su tutte le classi, eccezion fatta per la classe in cui tutti i dispositivi sono accesi, dove l'accuratezza è stata massima (1), la recall era esattamente la metà (0,50).

Questa discrepanza è stata attribuita all'eterogeneità nell'estrazione dei dati, in cui specifici comportamenti durante l'accensione simultanea dei tre dispositivi erano stati completamente esclusi, influenzando le prestazioni del modello su quella particolare classe.



Classificazione

Prova Iniziale

Inizialmente abbiamo trainato il modello con 100 alberi e attraverso un set di training e un set di test, il set di testing è stato ottenuto ricavando il 30% dei dati dal set originale, dunque il set di training è stato ottenuto rimuovendo i dati presenti dal set di testing.

Inizialmente, abbiamo addestrato il nostro modello utilizzando un set di training e un set di test. Per creare il set di test, abbiamo estratto il 30% dei dati dal set originale, mentre il set di training è stato generato eliminando i dati presenti nel set di test. Nonostante il modello abbia raggiunto un'accuratezza molto elevata su tutte le classi, eccezion fatta per la classe in cui tutti i dispositivi sono accesi, dove l'accuratezza è stata massima (1), la recall era esattamente la metà (0,50).

Questa discrepanza è stata attribuita all'eterogeneità nell'estrazione dei dati, in cui specifici comportamenti durante l'accensione simultanea dei tre dispositivi erano stati completamente esclusi, influenzando le prestazioni del modello su quella particolare classe.



Classificazione

Seconda Prova

Per migliorare le performance del modello, abbiamo deciso di aumentare il numero di alberi nel nostro Random Forest da 100 a 500. Questa scelta è motivata dalla capacità del Random Forest di migliorare la precisione e la stabilità del modello con l'aumentare del numero di alberi, consentendo una migliore generalizzazione ai dati di test.

Per ottimizzare l'estrazione dei dati e affrontare l'eterogeneità precedentemente riscontrata, abbiamo adottato un approccio stratificato. Abbiamo estratto il 20% dei valori da ogni classe in modo proporzionale alla distribuzione originale delle classi. Questo metodo assicura che ogni classe sia rappresentata in modo equo nei set di training e test, evitando squilibri che potrebbero compromettere l'addestramento del modello.



Classificazione

Risultati Seconda Prova

I risultati ottenuti dopo i suddetti cambiamenti hanno portato ad un grande miglioramento soprattutto riguardo la classe Tutti accesi con una baseline accuracy di 0.9988

	precision	recall	f1-score	support
AO	1.00	1.00	1.00	8501970
OW	1.00	1.00	1.00	69501
DW	1.00	1.00	1.00	182445
DW_OW	1.00	1.00	1.00	20340
WM	1.00	1.00	1.00	238119
WM_OW	1.00	1.00	1.00	21308
WM_DW	1.00	1.00	1.00	37804
WM_DW_OW	1.00	1.00	1.00	513
accuracy			1.00	9072000
macro avg	1.00	1.00	1.00	9072000
weighted avg	1.00	1.00	1.00	9072000

Classification report su un predict su tutto il dataset

	precision	recall	f1-score	support
AO	1.00	1.00	1.00	1700394
OW	1.00	1.00	1.00	13900
DW	1.00	1.00	1.00	36489
DW_OW	1.00	1.00	1.00	4068
WM	0.98	0.98	0.98	47624
WM_OW	1.00	1.00	1.00	4262
WM_DW	1.00	1.00	1.00	7561
WM_DW_OW	1.00	1.00	1.00	102
accuracy			1.00	1814400
macro avg	1.00	1.00	1.00	1814400
weighted avg	1.00	1.00	1.00	1814400

Classification report su un predict sul dataset di testing



Classificazione

Risultati Seconda Prova

	AO	OW	DW	DW_OW	WM	WM_OW	WM_DW	WM_DW_OW
AO	8500974	1	15	0	980	0	0	0
OW	6	69487	6	0	2	0	0	0
DW	87	2	182338	3	11	0	4	0
DW_OW	0	0	2	20337	0	1	0	0
WM	984	1	8	0	237125	1	0	0
WM_OW	0	1	0	0	5	21300	2	0
WM_DW	0	0	2	1	2	1	37798	0
WM_DW_OW	0	0	0	0	0	0	0	513

matrice di confusione sul predict su tutto il dataset



Classificazione

Terza Prova

Per ulteriormente migliorare le performance del nostro modello, abbiamo inizialmente eliminato le feature poco rilevanti grazie alla feature importance (anno year) e successivamente abbiamo intrapreso due ulteriori strategie. Innanzitutto, abbiamo introdotto un filtro per mitigare l'impatto dei rumori, concentrandoci specificamente sul **filtro di media**. Questa scelta è stata fatta per ridurre le variazioni casuali nei dati e stabilizzare le informazioni, permettendo al modello di apprendere da pattern più consistenti e rilevanti.

Inoltre, per ampliare la capacità del modello di individuare comportamenti specifici dei dispositivi, abbiamo introdotto **nuove features**. Queste nuove variabili sono state selezionate per fornire al modello ulteriori informazioni e pattern distintivi, consentendogli di affinare ulteriormente la sua capacità di classificazione. L'obiettivo è quello di catturare aspetti specifici dei dati che potrebbero essere stati precedentemente trascurati, migliorando così la comprensione complessiva del comportamento dei dispositivi.

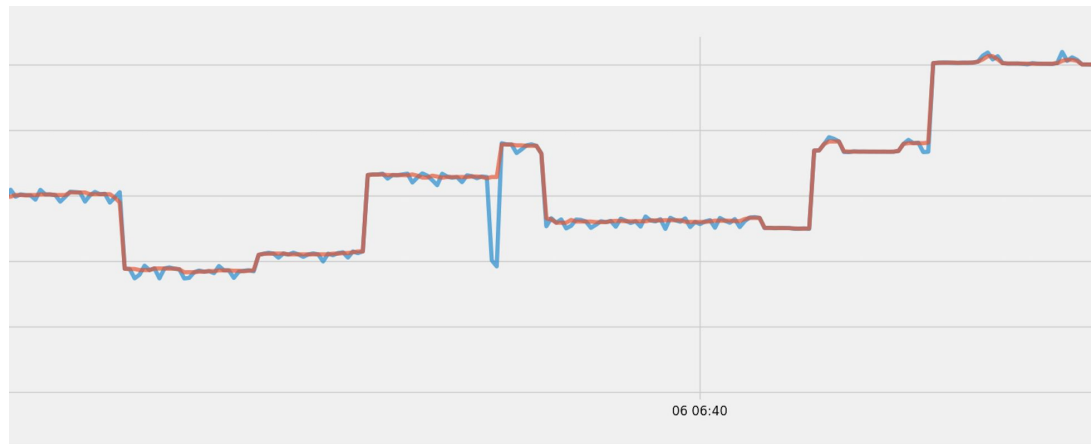


Classificazione

Filtro di media

L'aggiunta di un filtro di mediana contribuisce a "lisciare" la serie temporale dei dati, riducendo la presenza di fluttuazioni indesiderate e rendendo il segnale complessivo più coerente. Questo può essere particolarmente benefico nel contesto del Non-Intrusive Load Monitoring (NILM), dove la precisione nella rilevazione dei comportamenti energetici dei dispositivi è cruciale e dove c'è bisogno di distinguere lo steady state.

Il filtro di mediana è una tecnica di filtraggio che mira a ridurre il rumore in un segnale o una serie temporale mantenendo al contempo le caratteristiche salienti del segnale. Questo tipo di filtro opera sostituendo ogni valore nel segnale con la mediana dei valori presenti in una finestra di dimensioni specifiche centrata su di esso. Ho utilizzato una finestra di dimensioni 5 per il filtro di mediana. Ciò significa che, per ogni punto dati nella serie temporale, il valore sarà sostituito con la mediana dei cinque valori circostanti: due valori prima, il valore stesso, e due valori successivi.



● Segnale di ingresso

● Segnale di uscita (filtrato)

Esempio
di Filtraggio
sull'Active Power



Classificazione

Feature extraction

- **Instant Power** (Potenza istantanea) : La potenza istantanea può essere calcolata come il prodotto della tensione istantanea e la corrente istantanea in un dato istante di tempo.
- **Apparent Power** (Potenza Apparente): La potenza apparente è la magnitudine complessa del vettore della potenza istantanea. Può essere calcolata come la radice quadrata della somma dei quadrati della potenza attiva e la potenza reattiva
- **Power Factor** (Fattore di Potenza): Il fattore di potenza è il coseno dell'angolo tra la potenza attiva e la potenza apparente. Può essere calcolato come potenza attiva fratto potenza apparente
- **Power Fluctuation** (Fluttuazione di Potenza): le fluttuazioni di potenza in un sistema elettrico si calcolano attraverso la differenza tra le letture di potenza attiva in istanti di tempo consecutivi.
- **Resistance** (Resistenza): La resistenza può essere stimata utilizzando la legge di Ohm e quindi attraverso tensione fratto corrente



Classificazione

Risultati Terza Prova

A questo punto grazie a queste migliorie si è notato che il modello è più preciso ad individuare le classi ma con un leggero peggioramento sulla lavatrice con una baseline accuracy di 0.9989

	precision	recall	f1-score	support
AO	1.00	1.00	1.00	8501970
OW	1.00	1.00	1.00	69501
DW	1.00	1.00	1.00	182445
DW_OW	1.00	1.00	1.00	20340
WM	1.00	1.00	1.00	238119
WM_OW	1.00	1.00	1.00	21308
WM_DW	1.00	1.00	1.00	37804
WM_DW_OW	1.00	1.00	1.00	513
accuracy			1.00	9072000
macro avg	1.00	1.00	1.00	9072000
weighted avg	1.00	1.00	1.00	9072000

Classification report su un predict su tutto il dataset

	precision	recall	f1-score	support
AO	1.00	1.00	1.00	1700394
OW	1.00	1.00	1.00	13900
DW	1.00	1.00	1.00	36489
DW_OW	1.00	1.00	1.00	4068
WM	0.98	0.98	0.98	47624
WM_OW	1.00	1.00	1.00	4262
WM_DW	1.00	1.00	1.00	7561
WM_DW_OW	1.00	1.00	1.00	102
accuracy			1.00	1814400
macro avg	1.00	1.00	1.00	1814400
weighted avg	1.00	1.00	1.00	1814400

Classification report su un predict sul dataset di testing



Classificazione

Risultati Terza Prova

	AO	OW	DW	DW_OW	WM	WM_OW	WM_DW	WM_DW_OW
AO	8500822	0	15	0	1133	0	0	0
OW	8	69493	0	0	0	0	0	0
DW	79	1	182365	0	0	0	0	0
DW_OW	0	0	1	20339	0	0	0	0
WM	997	3	0	0	237118	1	0	0
WM_OW	0	1	0	0	2	21305	0	0
WM_DW	0	0	0	0	0	0	37804	0
WM_DW_OW	0	0	0	0	0	0	0	513

matrice di confusione sul predict su tutto il dataset



Conclusioni

Nel complesso, i risultati ottenuti nel progetto sono promettenti, indicando un'elevata performance del modello proposto. Tuttavia, per confermare in modo più robusto l'effettivo successo del modello e la sua capacità di generalizzazione su dati non visti, sarebbe ottimale applicare una procedura di cross validation. Questo consentirebbe di valutare la stabilità delle prestazioni del modello su diverse partizioni del dataset e di rilevare eventuali problemi di overfitting o underfitting. È importante sottolineare che il processo di training dei modelli ha richiesto mediamente circa 200 minuti. Questo aspetto va considerato nell'ottica di eventuali implementazioni pratiche del modello, poiché il tempo di addestramento può influenzare la fattibilità e l'efficienza dell'applicazione del modello in ambienti operativi.



Conclusioni

Possibili evoluzioni

Le future evoluzioni del progetto potrebbero concentrarsi su diverse direzioni per affinare ulteriormente la precisione e l'applicabilità pratica del modello di Non-Intrusive Load Monitoring (NILM).

- **Applicazione di Filtri Avanzati:** Integrare filtri avanzati come il filtro di Kalman potrebbe essere una strategia interessante. Questo consentirebbe di distinguere non solo tra gli stati transitori e stabili, ma anche di catturare in modo più accurato le transizioni di stato, contribuendo a una migliore comprensione delle dinamiche di consumo energetico dei dispositivi.
- **Estrazione di Feature Specifiche per Ogni Stato:** Considerare l'estrazione di feature specifiche per ciascun stato (transiente o stabile) potrebbe migliorare la capacità del modello di distinguere comportamenti sottili e fornire informazioni dettagliate su come i dispositivi si comportano in diverse condizioni. Ciò potrebbe essere ottenuto identificando estraendo feature rilevanti per ciascun stato.
- **Esplorare Modelli di Classificazione Alternativi:** Esplorare modelli di classificazione alternativi potrebbe rappresentare un'evoluzione significativa. Adottare un modello più leggero e veloce potrebbe rendere il processo di training più efficiente e l'applicazione in tempo reale più praticabile. Algoritmi come le reti neurali leggere o gli algoritmi basati su alberi decisionali potrebbero essere opzioni da esplorare in questa prospettiva.



Conclusioni

Progetto Svolto dal Team Bravo

Manuele Longo

Davide Cologgi

Lorenzo Nicotera