

Smart House - Riconoscimento di attività quotidiane

794324 - Francesco Morano

793414 - Stefano Pelliccioli

Introduzione

L'invecchiamento della popolazione sta avendo un impatto sempre più significativo sul sistema sanitario. Secondo stime recenti, entro il 2025 più di un miliardo di persone avranno un'età superiore ai 60 anni.

A fronte di questo problema, si rende necessaria la creazione di sistemi automatici per monitorare lo status di salute della popolazione. La soluzione prevalente è data dalle Wireless Sensor Network (WSN), una rete di sensori di diverso tipo, posti negli infissi ed elettrodomestici delle abitazioni e coordinati da un'unità centrale che elabora i dati ricevuti. Essa fornisce informazioni sulle attività quotidiane, permettendo di monitorare il comportamento degli abitanti, senza risultare eccessivamente intrusiva.

Sono già state fatte numerose sperimentazioni, utilizzando diversi modelli computazionali probabilistici per ottenere un riconoscimento efficiente di queste attività, tra cui Reti Bayesiane, Conditional Random Fields e Hidden Markov Models.

In particolar modo, questi ultimi si prestano efficacemente all'apprendimento dei dati e alla ricerca di pattern quando i dati da analizzare formano una serie temporale.

Dataset

Il dataset è formato da due sperimentazioni, denominate OrdonezA e OrdonezB. Per ognuna di esse, sono riportati 3 file di testo:

- **ADLs:** questo file contiene le attività svolte sotto forma di triple **"Start Time - End Time - Activity"**;

Start time	End time	Activity
2011-11-28 02:27:59	2011-11-28 10:18:11	sleeping
2011-11-28 10:21:24	2011-11-28 10:23:36	Toileting
2011-11-28 10:25:44	2011-11-28 10:33:00	showering
2011-11-28 10:34:23	2011-11-28 10:43:00	Breakfast
2011-11-28 10:49:48	2011-11-28 10:51:13	Grooming
2011-11-28 10:51:41	2011-11-28 13:05:07	Spare_Time/TV
2011-11-28 13:06:04	2011-11-28 13:06:31	Toileting
2011-11-28 13:09:31	2011-11-28 13:29:09	Leaving
2011-11-28 13:38:40	2011-11-28 14:21:40	Spare_Time/TV
2011-11-28 14:22:38	2011-11-28 14:27:07	Toileting

Figura 1: Dataset delle attività

- **Description:** una descrizione sommaria della durata della sperimentazione, attività possibili, tipologia e posizione dei sensori;
- **Sensors:** le singole rilevazioni effettuate dai sensori, indicate come "Start Time - End Time - Location - Type - Place". Location fa riferimento al sensore, Type alla tipologia del sensore e Place indica la stanza nel quale si trova.

Start time	End time	Location	Type	Place
2011-11-28 02:27:59	2011-11-28 10:18:11	Bed	Pressure	Bedroom
2011-11-28 10:21:24	2011-11-28 10:21:31	Cabinet	Magnetic	Bathroom
2011-11-28 10:21:44	2011-11-28 10:23:31	Basin	PIR	Bathroom
2011-11-28 10:23:02	2011-11-28 10:23:36	Toilet	Flush	Bathroom
2011-11-28 10:25:44	2011-11-28 10:32:06	Shower	PIR	Bathroom
2011-11-28 10:34:23	2011-11-28 10:34:41	Fridge	Magnetic	Kitchen
2011-11-28 10:34:44	2011-11-28 10:37:17	Cupboard	Magnetic	Kitchen
2011-11-28 10:38:00	2011-11-28 10:42:41	Toaster	Electric	Kitchen
2011-11-28 10:38:33	2011-11-28 10:38:40	Fridge	Magnetic	Kitchen
2011-11-28 10:41:29	2011-11-28 10:41:36	cupboard	Magnetic	Kitchen

Figura 2: Dataset dei sensori

Hidden Markov Model (HMM)

Un HMM è un modello generativo probabilistico definito da variabili osservabili x_t , dette osservazioni o evidenze, e variabili nascoste y_t , dette stati, dove la variabile t indica il tempo. Nel nostro caso le osservazioni sono le letture dei sensori, mentre gli stati rappresentano le attività.

In un HMM classico lo stato y_t dipende solamente dallo stato precedente y_{t-1} , mentre le variabili osservabili x_t dipendono solo dallo stato y_t . Si può quindi descrivere un HMM come un processo Markoviano del primo ordine, nel quale gli stati del processo non sono osservabili.

I parametri principali che descrivono un modello di Markov nascosto sono:

- π il vettore delle probabilità a priori $p(y_i)$;
- T la matrice delle probabilità di transizione tra gli stati $p(y_i|y_{i-1})$;
- O la matrice delle probabilità di emissione delle osservazioni $p(x_i|y_i)$.

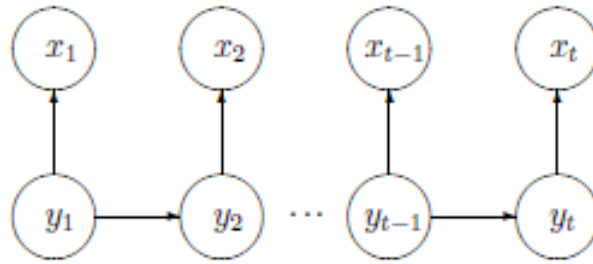


Figura 3: Struttura e dipendenze di un HMM

Software

Per lo svolgimento di questa attività abbiamo utilizzato il linguaggio open-source Python, utilizzando la libreria Pomegranate che fornisce l'implementazione di diversi tipi di modelli probabilistici, partendo dalle singole distribuzioni di probabilità fino a strutture più complesse come Reti Bayesiane, HMM, GMM e classificatori Bayesiani. Questo package fornisce, inoltre, diversi metodi per la definizione della struttura e per il fitting dei parametri di ogni modello.

Integrazione

Prima di definire la struttura del modello, abbiamo ritenuto necessario integrare i dati in un unico file. Per fare ciò, abbiamo implementato una funzione che, leggendo i file delle attività e dei sensori, genera un solo file dove le osservazioni sono riportate come quadruple **Start Time - End Time - Location - Activity**, dove i primi due termini fanno riferimento al tempo di inizio e fine dell'attivazione del sensore.

Durante questo procedimento, dal dataset dei sensori è stato considerato l'attributo Location, in quanto più informativo rispetto alla tipologia dei sensori o all'ambiente in cui si trovano.

Il dataset definitivo contiene quindi, per ogni istanza, i tempi di attivazione di un sensore, la sua posizione e l'attività rilevata corrispondente.

Start Time	End Time	Location	Activity
2011-11-28 02:27:59	2011-11-28 10:18:11	Bed	Sleeping
2011-11-28 10:21:24	2011-11-28 10:21:31	Cabinet	Toileting
2011-11-28 10:21:44	2011-11-28 10:23:31	Basin	Toileting
2011-11-28 10:23:02	2011-11-28 10:23:36	Toilet	Toileting
2011-11-28 10:25:44	2011-11-28 10:32:06	Shower	Showering
2011-11-28 10:34:23	2011-11-28 10:34:41	Fridge	Breakfast

Figura 4: Risultato della procedura di integrazione

Modelli utilizzati

L'obiettivo principale che ci siamo posti, per questo progetto, è stato quello di creare un modello HMM in grado di rispondere a questa domanda: "È possibile predire le attività di una persona a partire dai dati forniti dai sensori?". Inoltre, questo ci ha permesso di rispondere ad un altro quesito: "Il modello dimostra la stessa precisione per tutte le attività?". Considerando questi obiettivi, abbiamo deciso di effettuare tre diverse sperimentazioni:

- Modello HMM equiprobabile (Baseline);
- Modello HMM equiprobabile con addestramento;
- Modello HMM creato basato sulle osservazioni.

La validazione dei modelli è stata effettuata tramite la procedura di K - Fold Validation, dove il dataset è partizionato in K insiemi e uno di essi è utilizzato come Test, mentre gli altri vengono usati come Training. Nel nostro caso, abbiamo effettuato una variante detta One-Day Validation, dove il dataset è separato in base ai giorni rilevati dalla rete di sensori; il numero di folds K è quindi pari al numero di giorni rilevati e iterativamente uno di essi viene utilizzato per il test e la misura delle performance, e tutti gli altri come training.

Per ognuno di essi, sono state calcolate due misure di performance:

- Differenza punto a punto(%): è un valore che misura l'errore di previsione, calcolato come differenza tra le attività previste dal modello e quelle reali contenute nella groundtruth ad ogni istante (espresso in percentuale).

- F-Measure(%): misura di performance largamente utilizzata nell'ambito di Machine Learning, viene calcolata tramite la seguente formula:

$$F\text{-measure} = \frac{(2PrecisionRecall)}{(Precision+Recall)} \text{ (espressa in percentuale).}$$

Modello HMM Equiprobabile

Questo modello è stato scelto per essere utilizzato come Baseline. Una Baseline è un punto di riferimento per la valutazione delle performance di un sistema di Machine Learning ottenuto tramite tecniche euristiche, statistiche semplici o casualità.

Il modello è stato inizializzato considerando di non conoscere nulla sulle probabilità di inizio, transizione ed emissione del modello, per cui ci poniamo nel caso peggiore, ovvero quello in cui tutte le probabilità sono identiche.

Intuitivamente ci aspettiamo che i risultati ottenuti da questo modello non siano in grado di fornire una previsione accurata delle attività rilevate, perciò l'abbiamo scelto come riferimento per misurare la differenza di prestazioni con gli altri modelli.

Modello HMM Equiprobabile con addestramento

Questo modello inizializza la struttura dell'HMM basandosi sull'ipotesi che le probabilità che lo caratterizzano siano sconosciute, quindi ogni parametro è descritto da una distribuzione uniforme. In seguito, i parametri vengono modificati tramite un algoritmo di training supervisionato che addestra il modello utilizzando solamente la parte dei dati etichettata come Training. Abbiamo deciso di creare questo modello per evidenziare l'importanza dell'inizializzazione del modello (Structure Learning), anche rispetto alla fase di addestramento (Fitting). In molti casi, conoscere dei buoni valori con cui inizializzare i parametri può portare ad avere risultati migliori. Prevediamo di ottenere dei risultati migliori rispetto al modello precedente, grazie alla procedura di fitting, ma non i più accurati possibili a causa dell'inizializzazione.

Modello HMM basato sulle osservazioni

Per ultimo, abbiamo realizzato un Hidden Markov Model i cui parametri vengono inizializzati tramite delle funzioni ad-hoc, da noi implementate, che calcolano le probabilità tramite enumerazione, in modo da descrivere le tre distribuzioni multinomiali corrispondenti ai parametri. Le funzioni sono:

- `init-prob()`: Stima del vettore delle probabilità iniziali normalizzate.

- `trans-prob()`:Stima della matrice di transizione con probabilità normalizzate.
- `ems-prob()`:Stima della matrice delle osservazioni con probabilità normalizzate.

Tutte le funzioni elencate utilizzano degli pseudocontatori per evitare errori in fase di Test, relativi ad attività non presenti nell'insieme di Training. L'utilizzo di questo modello è dovuto alla necessità di ottenere una maggiore precisione nel riconoscimento delle attività, sfruttando la presenza di dati già verificati per imparare la struttura del modello. Ovviamente, ci aspettiamo di ottenere delle performance migliori rispetto ai modelli precedenti.

Risultati e discussione

i-esimo "giorno"	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Diff.	76.66	87.09	87.05	81.81	100	77.77	82.61	85.29	85.19	89.19	73.68	78.26	71.43	77.77
F-Measure	3.78	2.28	2.22	3.07	0	3.63	2.96	2.56	2.58	1.95	4.16	3.57	4.44	3.63

Figura 5: Risultati modello Baseline - dataset A

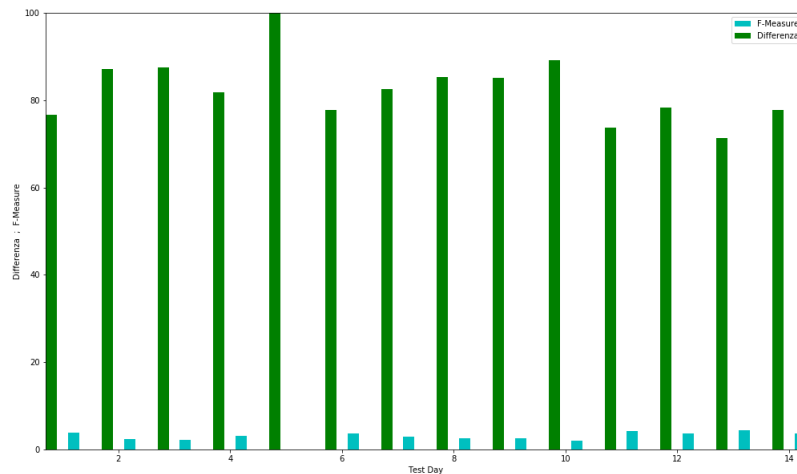


Figura 6: Grafico risultati modello Baseline - dataset A

i-esimo "giorno "	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Diff.	50	48.38	56.25	68.18	62.5	61.11	56.52	55.88	55.55	59.46	63.15	65.21	47.61	38.88
F-Measure	27.91	36.43	30.30	15.60	15.63	21.24	29.09	22.90	30.92	30.64	18.79	17.57	19.86	30.87

Figura 7: Risultati modello equiprobabile - dataset A

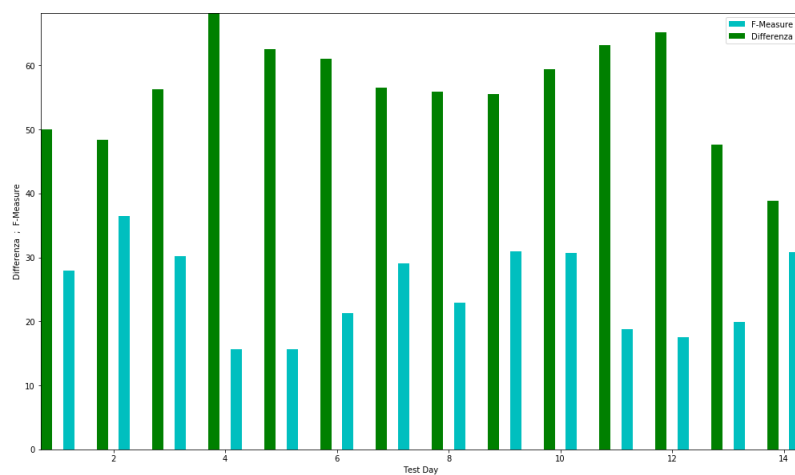


Figura 8: Grafico risultati modello equiprobabile - dataset A

i-esimo "giorno"	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Diff.	13.33	6.45	18.75	27.27	6.25	5.55	8.69	5.88	3.70	16.21	0	0	0	5.55
F-Measure	83.16	65.66	76.14	58.72	48.71	57.48	75.49	78.54	77.60	75.51	60	70	60	75

Figura 9: Risultati modello basato su osservazioni - dataset A

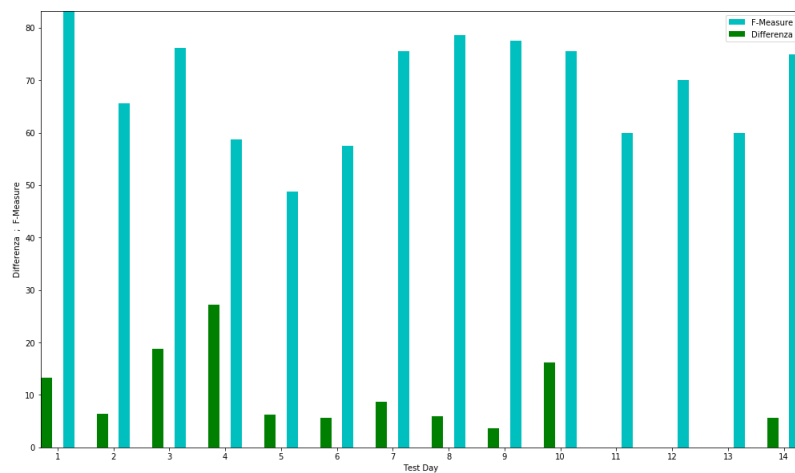


Figura 10: Grafico risultati modello basato su osservazioni - dataset A

i-esimo giorno	1	2	3	4	5	6	7	8	9	10	11
Diff.	100	96.11	92.3	87.23	90.62	94.29	91.04	93.51	89.55	91.66	89.83
F-Measure	0	0.74	1.42	2.27	1.71	1.08	1.64	1.22	1.89	1.54	1.85

12	13	14	15	16	17	18	19	20	21	22
92.5	94.60	93.33	93.33	88.37	91.66	89.41	88.88	79.76	92.85	92.99
1.40	1.03	1.25	1.25	2.09	1.54	1.91	2	3.37	1.33	1.31

Figura 11: Risultati modello Baseline - dataset B

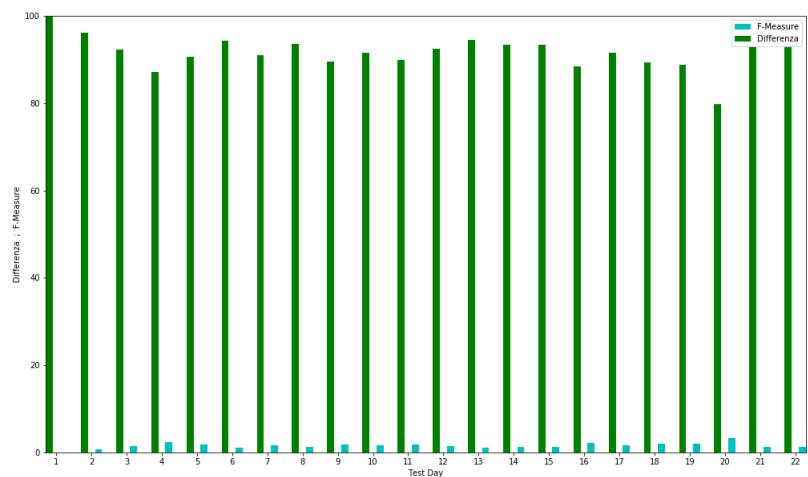


Figura 12: Grafico risultati modello Baseline - dataset B

i-esimo giorno	1	2	3	4	5	6	7	8	9	10	11
Diff.	0	32.04	28.21	25.54	46.87	21.42	22.39	31.17	37.31	20	25.42
F-Measure	10	57.62	59.41	50.37	45.53	59.15	52.87	56.20	47.73	67.22	58.02

12	13	14	15	16	17	18	19	20	21	22
57.50	47.30	49.33	28.33	23.25	23.33	36.47	42.85	40.47	47.62	42.10
44.68	43.89	33.09	52.30	46.64	49.30	49.87	39.43	42.91	48.83	45.27

Figura 13: Risultati modello equiprobabile - dataset B

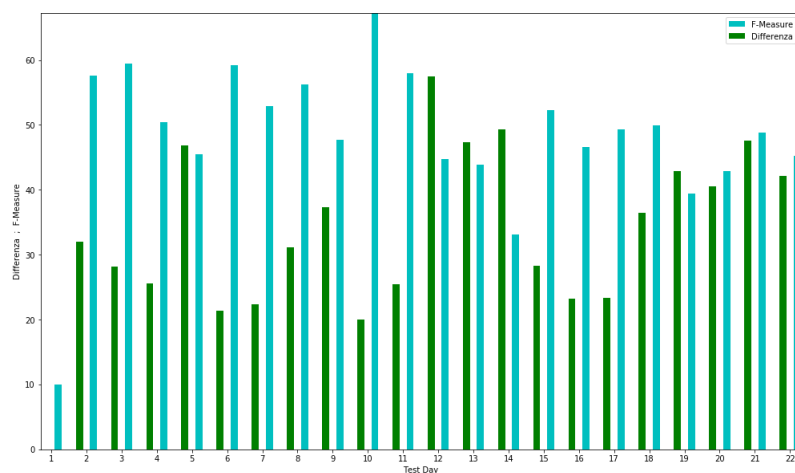


Figura 14: Grafico risultati modello equiprobabile - dataset B

i-esimo giorno	1	2	3	4	5	6	7	8	9	10	11
Diff.	0	35.92	25.64	31.91	29.56	10	5.97	24.67	34.33	16.66	18.64
F- Measure	10	69.77	62.95	59	56.50	80.98	64.90	70.13	59.31	70.75	64.06

12	13	14	15	16	17	18	19	20	21	22
45	29.73	41.33	35	9.30	16.66	23.53	28.58	30.95	23.90	31.58
50.68	52.55	56.41	57.60	56.48	55.39	57.08	50.39	42.91	48.83	45.27

Figura 15: Risultati modello basato su osservazioni - dataset B

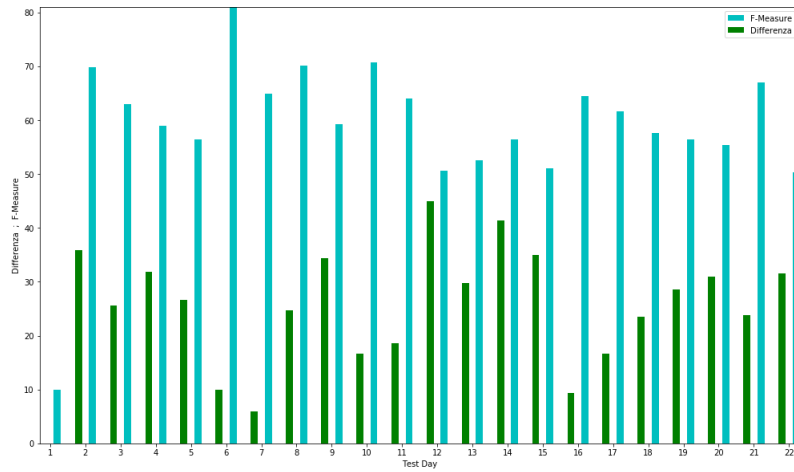


Figura 16: Grafico risultati modello basato su osservazioni - dataset B

Modello	Differenza OrdenezA	F-measure OrdenezA	Differenza OrdenezB	F-measure OrdenezB
Modello totalmente equiprobabile	82.41	2.91	91.53	1.54
Modello equiprobabile con addestramento	56.34	24.84	33.14	48.20
Modello estratto dai dati	8.40	68.72	24.80	58.73

Figura 17: Valori medi delle misure di performance

Valutando le tabelle ed i grafici, qui riportati, possiamo osservare come i risultati rispettino le nostre aspettative; mostrano infatti un miglioramento delle misure di performance incrementale nelle diverse sperimentazioni proposte.

Possiamo notare come il secondo modello fornisca già un notevole miglioramento rispetto alla baseline, mentre si evidenzia che le prestazioni migliori sono ottenute dal terzo modello, che sfrutta le istanze del dataset per determinarne i parametri. Inoltre, è stato effettuato un ulteriore tentativo di miglioramento, sfruttando la fase di Fitting sul modello costruito a partire dai dati, nella speranza di riuscire a meglio identificare le attività quotidiane; purtroppo questo esperimento non ha portato i risultati desiderati, al contrario, si è verificato un peggioramento delle performance.

Demo

In aggiunta alle sperimentazioni effettuate, è stata implementata un'interfaccia grafica per visualizzare in modo dinamico i risultati ottenuti dai modelli. Questa GUI dà all'utente l'opportunità di:

- Scegliere un file del dataset;
- Selezionare il numero di giorni di Training;
- Visualizzare i risultati ottenuti dai tre modelli.

In questo caso, la partizione di Test è composta dai giorni rimanenti dopo la selezione della durata dell'addestramento.

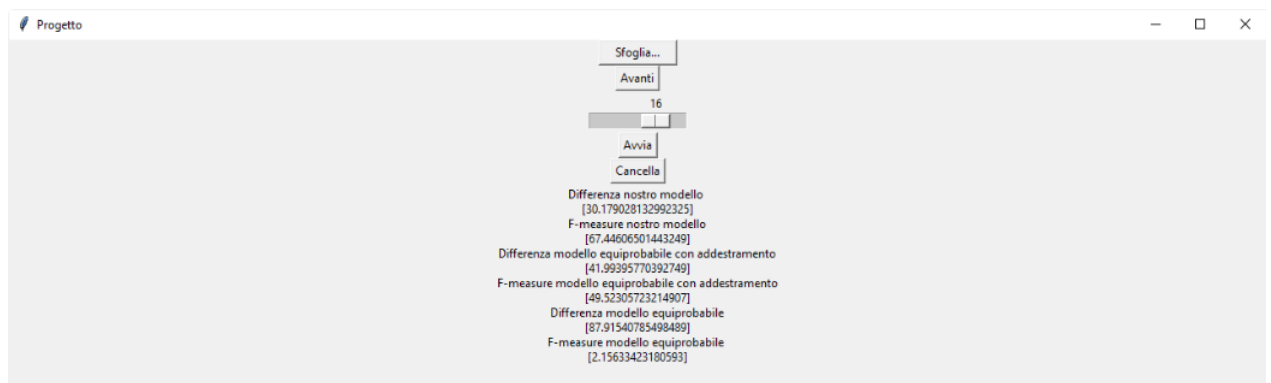


Figura 18: Interfaccia grafica

Conclusioni e sviluppi futuri

Lo scopo principale del progetto era rispondere a queste due domande: "È possibile predire le attività di una persona a partire dai dati forniti dai sensori?" e "Il modello dimostra la stessa precisione per tutte le attività?".

Dopo aver analizzato i modelli costruiti e i corrispettivi risultati, possiamo dire di aver raggiunto un buon risultato per quanto concerne la prima domanda. Per quanto riguarda la seconda, invece, si dimostra che non tutte le attività sono riconosciute con la stessa precisione; in particolare le attività **{Breakfast,Lunch,Snack,Dinner}** e **{Grooming, Toileting}** sono spesso confuse tra loro, poiché i sensori che le caratterizzano sono gli stessi. Come sviluppi futuri e proposte di miglioramento, possiamo avanzare due ipotesi:

- Modificare il dataset, sostituendo gli insiemi sopracitati con un'unica attività più generale (es. Eating), questo però comporta una perdita nella distinzione delle attività;
- Sfruttare le informazioni temporali dei sensori per meglio distinguere certe attività (Morning -> Breakfast, Evening->Lunch) costruendo un HHMM (Hierarchical HMM) o un HMM equivalente.