Lab Data Challenge

Luca Iaria

Marco Emanuele Saini

Alessandro Colello

Introduzione

Che cosa rappresentano i dati?

I dati rappresentano l'ampiezza e la fase nel tempo dei segnali wifi delle stanze. Le prime 112 variabili indicano le serie temporali relative al segnale wifi in circa 430k istanti temporali, escluse le ultime 3 variabili target oggetto di studio mutualmente esclusive indicanti:

- **nessuna persona nella stanza:** vale 1 se la stanza è vuota, 0 altrimenti;
- **una sola persona (ferma) nella stanza:** vale 1 se nella stanza è presente una persona ferma, 0 altrimenti;
- **una sola persona (in movimento) nella stanza:** vale 1 se nella stanza è presente una persona in movimento, 0 altrimenti.

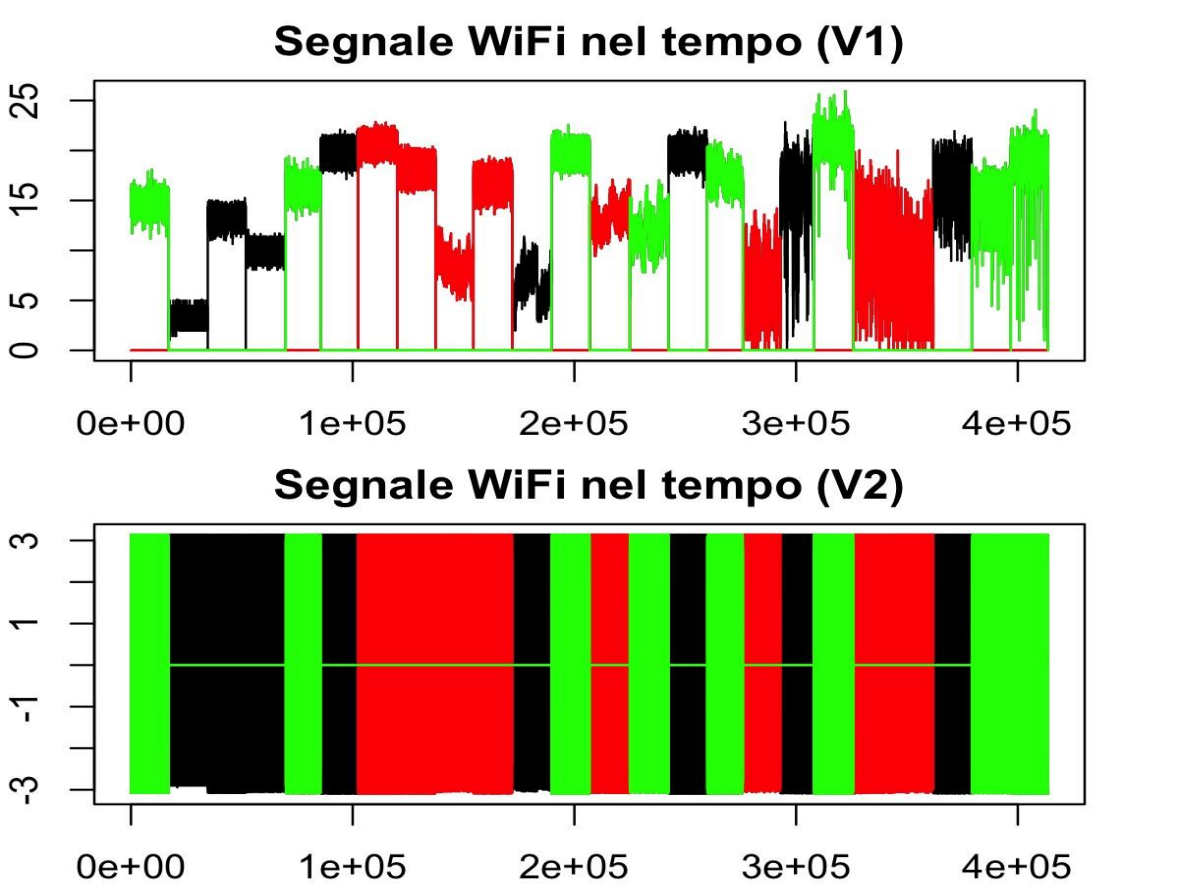
Analisi esplorativa del training set originale

Summary sulle prime 4 features e sulle ultime 2 features del training set originale:

	Media	Primo quartile	Terzo quartile	Minimo	Massimo
V1	14.18	10.05	18.60	0	25.94
V2	-0.02	-1.64	1.57	-3.09	3.14
V3	13.68	10.00	17.49	0	26.40
V4	-0.04	-1.67	1.57	-3.09	3.14
...	…	…	…	…	…
V111	13.90	11.00	18.03	0	25.50
V112	-0.02	-1.65	1.57	-3.10	3.14

Le variabili pari e dispari sembrano avere un andamento diverso.

Analisi grafica del training set originale

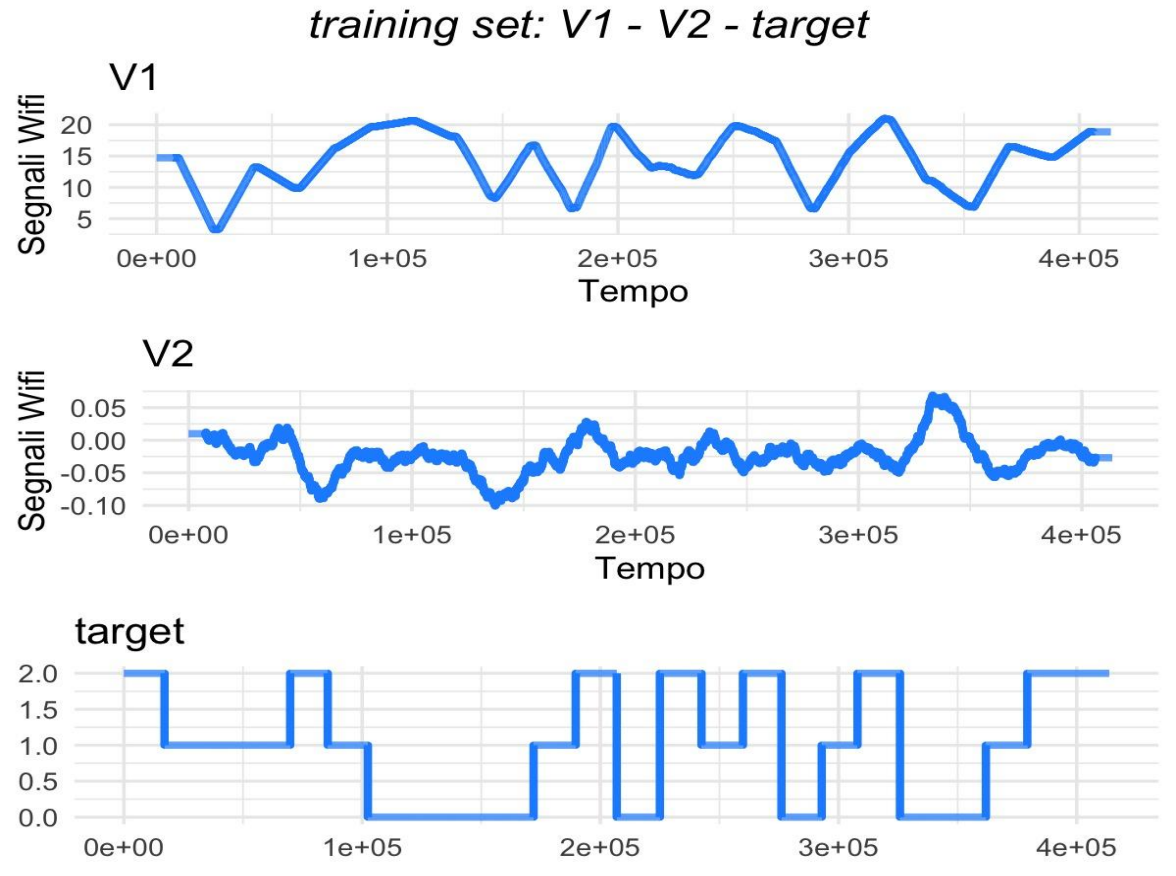


Nessuna persona

persona ferma

Persona in movimento

Analisi grafica training set con media mobile



Andamento nel tempo di due variabili del training set, trasformato tramite media mobile.

La variabile target assume i seguenti valori:

- **0**, nessuna persona;
- **1**, persona ferma;
- **2**, persona in movimento.

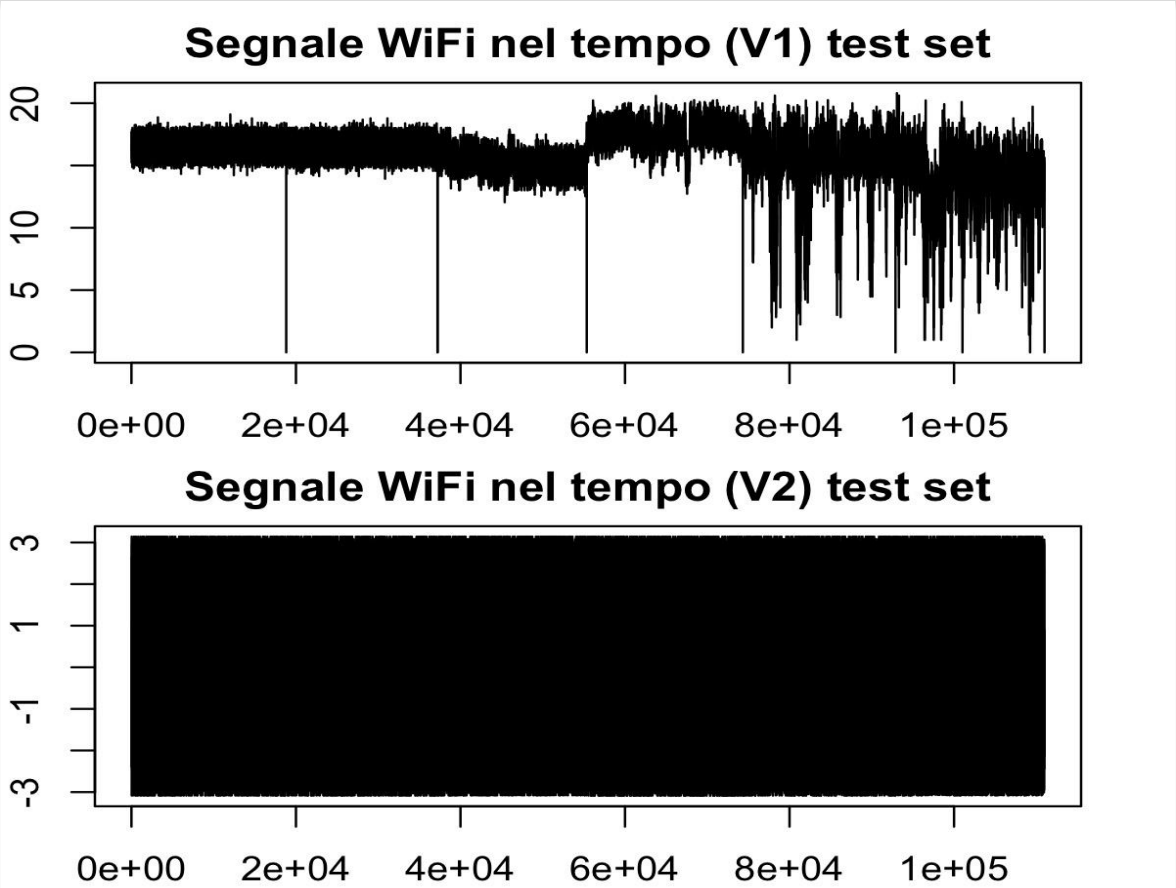
Analisi descrittiva del test set originale

Summary sulle prime 4 features e sulle ultime 2 features del test set originale:

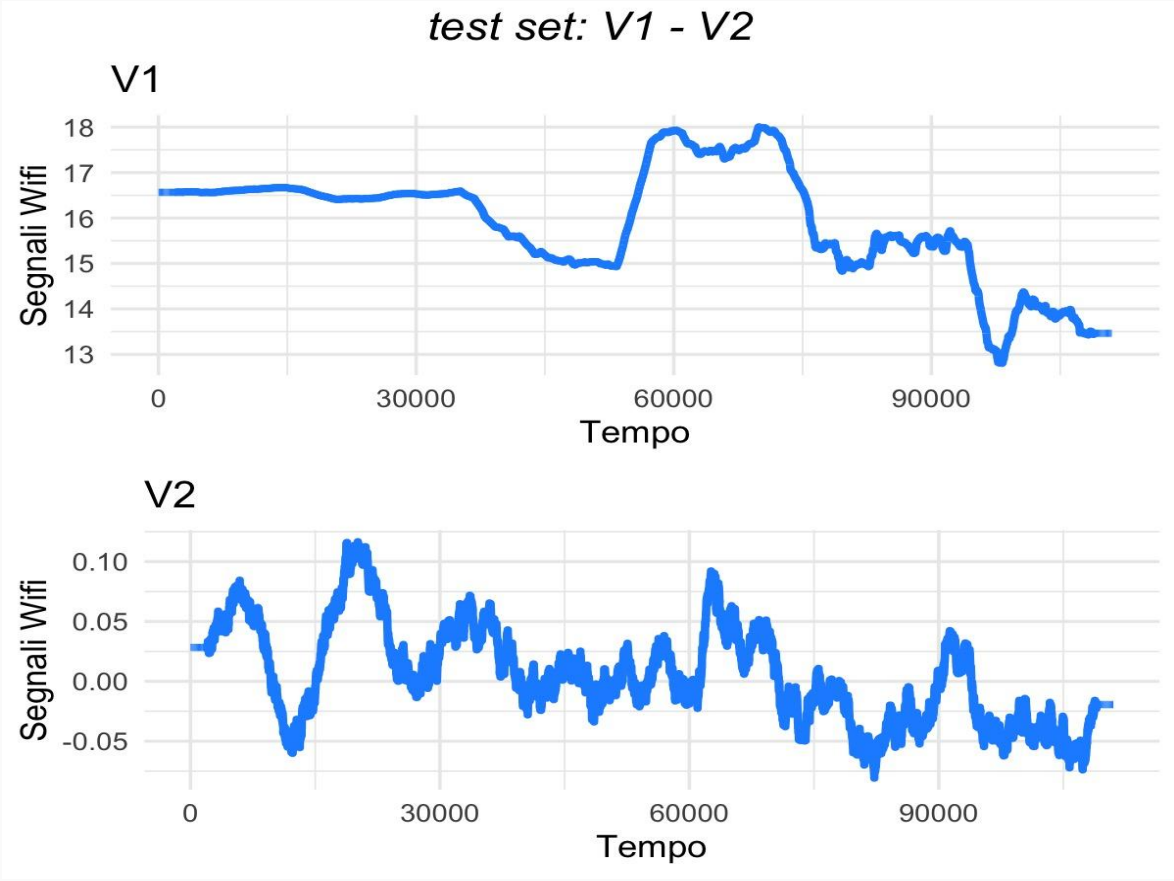
	Media	Primo quartile	Terzo quartile	Minimo	Massimo
V1	14.62	10.63	18.79	0	25.94
V2	-0.01	-1.62	1.57	-3.09	3.14
V3	14.15	10.30	17.80	0	26.40
V4	-0.03	-1.64	1.57	-3.09	3.14
...	…	…	…	…	…
V111	14.30	11.05	18.44	0	25.50
V112	-0.01	-1.63	1.57	-3.10	3.14

Le variabili pari e dispari sembrano avere un andamento diverso.

Analisi grafica del test set



Analisi grafica del test set con media mobile



Andamento delle variabili del test set trasformate con la media mobile.

Creazione del modello

Considerazioni preliminari:

- Anche il test set segue la stessa struttura del training set;
- Le ultime osservazioni del training set risultano temporalmente collegate al test set;
- Per ridurre il rumore dei dati, è stata applicata una media mobile con finestra di 15.000 osservazioni (pari al 3.62% del totale);
- Il dataset è stato suddiviso in training (85%) e validation (15%);
- Per ridurre la dimensionalità, è stata utilizzata l'analisi delle componenti principali (PCA): le prime 10 componenti spiegano circa il 95% di variabilità.

Modelli sperimentati:

- RNN, ha mostrato problemi di overfitting, con un'accuracy intorno al 33%;
- Random Forest, buona generalizzazione, con un'accuracy di circa il 43%.

Tuning degli iperparametri della random forest

- Per trovare un modello ottimale, è stato necessario ottimizzare gli iperparametri, in particolare il **numero di alberi** e la **profondità di un albero**;
- Per il numero di alberi sono stati considerati i seguenti valori: 50, 100, 500 e 1000. Invece per la profondità massima di ciascun albero sono stati considerati: 10, 50, 100 e 150;
- Combinando questi valori, sono stati confrontati 16 modelli diversi, ottenendo i seguenti risultati:

	deep = 10	**deep = 50**	**deep = 100**	**deep = 150**
trees = 50	0.9974855	0.9998227	0.9998227	0.9998227
trees = 100	0.9975661	0.9998388	0.9998388	0.9998227
trees = 500	0.9975500	0.9998465	0.9998388	0.9998388
trees = 1000	0.9975016	0.9998227	0.9998388	0.9998388

Previsione sul validation set

Il **miglior modello** individuato utilizza 500 alberi ed una profondità massima di 50 nodi. Applicando il modello ottimale sul validation set, si ottiene la seguente matrice di confusione:

	Prediction			
Reference		**0**	**1**	**2**
	0	21190	1	2
	1	1	20636	2
	2	1	3	20206

Accuracy: 99.9%.

Dalla matrice di confusione, si evince che non sono presenti segni di overfitting. Il modello sembra generalizzare bene anche su nuovi dati.

Previsione sul test set

Applicando il modello ottimale al test set, si ottiene la seguente matrice di confusione ottenuta è:

	Prediction			
Reference		**0**	**1**	**2**
	0	6858	15847	15469
	1	4077	23268	9609
	2	798	18141	17735

Accuracy: 42.4%.

Il modello risulta essere discreto.

Considerazioni finali

- Il rumore dei dati influenza significativamente le stime; trovare una trasformazione ottimale delle variabili non è semplice;
- Il modello prevede con maggiore accuratezza la presenza di una persona ferma nella stanza;
- Il modello mostra invece maggiori difficoltà nel riconoscere quando la stanza è vuota.

Grazie per l’attenzione