LA 2.1: Analisi ed elaborazione di dataset eterogenei in un formato interoperabile indirizzato allo sviluppo di tecniche di features selection e patterns recognition

Relatore: Giovanni Brunaccini

Analisi dataset di letteratura e Variabili raccolte in prove di carica/scarica

Dataset N.1 - Faraji-Niri (Università di Warwick)

Cicli di carica e scarica su celle cilindriche, interponendo EIS e capacity check.

Dataset N.2 – Dataset CALCE (Università del Maryland)

Misure su celle litio di varia natura ➔ Per la costruzione del database dimostrativo, sono stati scelti i dati relativi a test su celle INR 18650-20R cilindriche, test a bassa corrente (circa C/20, C/25) in carica e scarica (3 temperature e 4 rate).

Dataset N.3 - Severson (MIT), Attia (Università di Stanford)

124 celle commerciali (A123 Systems) ciclate a temperatura costante di 30°C (impulsi e corrente costante)

Dataset N.4 – Kollmeyer (tramite Mendeley Data)

Misure su celle Panasonic 18650PF a 5 differenti temperature ➔ selezionata solo 25°C per limitare il tempo di esecuzione delle query e il relativo consumo di risorse per debug e tuning.

Dataset N.5 (prove in laboratorio CNR-ITAE)

Test di invecchiamento su celle litio commerciali.

- Data_Point (id della misura registrata, univoco e seriale)
- Test_Time (tempo misurato dall'avvio del test, in s)
- DateTime (un timestamp posix della registrazione del datapoint)
- Step_Time (tempo misurato dall'avvio del singolo step del test, si azzera all'inizio di ciascuno step del test complessivo)
- Step_Index (valore progressivo interno al singolo ciclo di test per identificare lo step corrente del test)
- Cycle_Index (valore progressivo incrementato all'inizio di un nuovo ciclo del test)
- Current (corrente in carica o scarica della cella)
- Voltage (tensione della cella)
- Charge_Capacity (quantità di carica accumulata nel test di carica)
- Discharge_Capacity (quantità di carica rilasciata durante il test di scarica)
- Charge_Energy (energia immessa nella cella durante il test di carica)
- Discharge_Energy (energia rilasciata dalla cella durante il test di scarica)
- dV/dt (variazione della tensione a cui è stato sensibile lo strumento di misura per individuare un nuovo datapoint)
- Internal_Resistance (resistenza ohmica della cella calcolata sul singolo datapoint)
- Temperature (temperatura della camera di test)

Nel dataset sono presenti le informazioni di 10 celle invecchiate in camera climatica e per valutare l'impatto di differenti profili sulla vita utile della batteria (ciclo del costruttore e regolazione primaria della frequenza).

Metodologia: lo sviluppo del primo benchmark tramite il dataset CNR-ITAE

- **Dati impiegati**: Check-up test svolti nei laboratori del CNR-ITAE su dieci celle
- **Modalità di invecchiamento**: Ogni profilo è un caso di **regolazione di frequenza** in condizioni di lavoro differenti
- **Tipologia di benchmark**: **regressione lineare** su dati di **impedenza elettrochimica** (EIS) a differenti stati di invecchiamento e SoC

Condizioni per lo sviluppo della regressione:

i) Un numero identico di cicli di carico per ciascuna cella (12 valori tra 0 e 425)

ii) Un check-up test alla fine di ciascun set di cicli di invecchiamento, costituito da:

- tre misure di ***capacità*** (capacity check) tramite scarica e carica completa delle celle,
- un set di cinque misure di EIS (V➔I) a cinque **SoC** differenti, registrando la risposta in frequenza ***Z'+jZ''*** intorno all'**OCV** in scansioni di 50 frequenze,

Dataset totale:
600 record
102 features (input)
1 variabile target

Le misure saranno anche utilizzate in analisi e nel training dei modelli di machine learning.

Pre-processing per la normalizzazione del dataset :

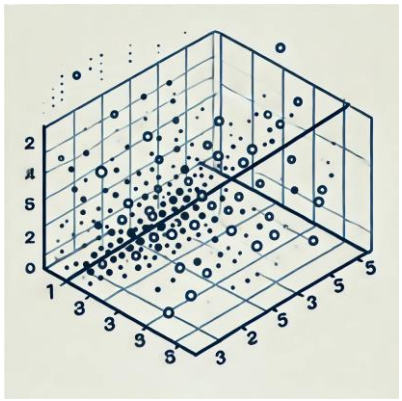
Nelle variabili usate per la stima dell'SoH è riscontrabile un'ampia variabilità (anche più ordini di grandezza) di:

- Valori assoluti (resistenze in Ω o mΩ, SoC in %, tensione in V)
- Range di variazione relativo nella stessa prova
- Range di variazione al procedere dell'invecchiamento del dispositivo

Analisi ed elaborazione di dataset eterogenei in un formato interoperabile indirizzato allo sviluppo di tecniche di features selection e patterns recognition

Metodologia: strumenti per l'elaborazione delle misure disponibili

Dati grezzi da strumenti di laboratorio



```python
from influxdb import InfluxDBClient
def scrivi_tsdb_locale(local_tags, local_fields, timestamp_locale):
    client = InfluxDBClient(tsdb_host, tsdb_port, tsdb_username,
                            tsdb_password, tsdb_database)
    tags = local_tags
    fields = local_fields
    # Create a JSON data point
    data_point = {
#            "time": timestamp.strftime('%Y-%m-%dT%H:%M:%SZ'),
        "measurement": measurement,
        "tags": tags,
        "fields": fields,
        "time": timestamp_locale,
    }
    client.write_points([data_point])
#    print("Ho scritto la ts")
    client.close()
    return
```

1) Script per la raccolta di ''time series'' (capacity test)

influxdb

2) Script per la raccolta dei dati '' scalari '' (EIS)

```python
import time, datetime
table_name = "dati_impianto"
now = datetime.datetime.now()
backup_file = "backup_"+table_name+"_"+now.strftime("%Y%m%d_%H%M%S")

if restore:
    restore_file_name = "backup_" + table_name + "_" + restore_datetime
    df = pd.read_feather('{}.feather'.format(restore_file_name))
    ''' ricopio il backup selezionato come nuovo WORKING FILE '''
    df.to_feather('{}.feather'.format(table_name))
else:
    if aggiorno_dati == True:
        df = pd.read_csv(csv_file_path)
        ''' salvataggio con feather FILE DI BACKUP'''
        df.to_feather('{}.feather'.format(backup_file_name))
        ''' salvataggio con feather WORKING FILE '''
        df.to_feather('{}.feather'.format(table_name))
    else:
        ''' uso feather invece di sqlite '''
        df = pd.read_feather('{}.feather'.format(table_name))
```

3) Estrazione da dataset preliminari

Features ordinate e pesate per fitness (''importanza'') del contenuto

```
Feature Ranking:
[88 83 60 61  5 66 24 91  4 37 85 71 90 55  6 73 41 82 81 78 79 84 50 56
 89 92 68 86  9 49 10 35 74 45  7 69  8 48 20 64 21 38 12 39 11 77 62 43
 42 76 75 59 13 63 14 16 58 46 57 36  2 44  1 93 15 65  3  1 72 27  1 25
 28 17 52  1 47 22 70 80 34 23  1 29  1  1 51  1 26 30 31 32 40  1 33 87
 54 18 67 19 53]
Selected Features:
['I31', 'R34', 'I35', 'R38', 'I41', 'I42', 'R43', 'R44', 'R47']
RFE Score: 0.6211571395841097
```

4) Normalizzazione

	cap	SoC	OCV	R1	I1	R2	I2	R3	I3	R4	...	R46	I46	R47	I47	R48	I48	R49	I49	R50	I50
0	100.000000	100	4.180333	0.022903	0.000566	0.023362	0.001180	0.023870	0.001669	0.024401	...	0.038200	0.008332	0.039213	0.009569	0.040289	0.011027	0.041389	0.012721	0.042608	0.014638
1	100.000000	80	4.024433	0.022699	0.000285	0.023126	0.000873	0.023587	0.001338	0.024069	...	0.035679	0.007267	0.036353	0.008208	0.037122	0.009219	0.038022	0.010364	0.039172	0.011695
2	100.000000	50	3.740439	0.022770	0.000261	0.023196	0.000841	0.023654	0.001293	0.024151	...	0.034116	0.006021	0.034679	0.006838	0.035335	0.007750	0.036143	0.008776	0.037179	0.009967
3	100.000000	20	3.476580	0.022918	0.000321	0.023330	0.000922	0.023799	0.001401	0.024288	...	0.034557	0.006037	0.035055	0.006879	0.035664	0.007829	0.036442	0.008985	0.037459	0.010369
4	100.000000	0	3.126237	0.023756	0.000834	0.024247	0.001504	0.024766	0.002064	0.025330	...	0.047726	0.016345	0.049361	0.019375	0.051039	0.023192	0.052798	0.027849	0.054723	0.033309
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
595	93.444648	100	4.168203	0.024816	-0.001656	0.025205	-0.000719	0.025619	0.000022	0.026036	...	0.059828	0.012980	0.061446	0.014158	0.062924	0.015575	0.064512	0.017301	0.066158	0.019423
596	93.444648	80	4.025332	0.024436	-0.001978	0.024772	-0.001042	0.025112	-0.000297	0.025471	...	0.043251	0.008312	0.044145	0.009202	0.045205	0.010194	0.046470	0.011359	0.048067	0.012779
597	93.444648	50	3.751479	0.024427	-0.002030	0.024733	-0.001095	0.025083	-0.000352	0.025439	...	0.039505	0.006370	0.040333	0.007318	0.041208	0.008433	0.042166	0.009681	0.043239	0.011102
598	93.444648	20	3.479544	0.024515	-0.001986	0.024858	-0.001041	0.025196	-0.000280	0.025566	...	0.040450	0.006598	0.041165	0.007441	0.042050	0.008529	0.043065	0.009899	0.044225	0.011547
599	93.444648	0	3.142405	0.025015	-0.001646	0.025378	-0.000655	0.025778	0.000159	0.026223	...	0.056276	0.016349	0.057918	0.018845	0.059681	0.022150	0.061480	0.026279	0.063300	0.031070

600 rows × 103 columns

Analisi ed elaborazione di dataset eterogenei in un formato interoperabile indirizzato allo sviluppo di tecniche di features selection e patterns recognition

Recursive Feature Elimination (RFE)

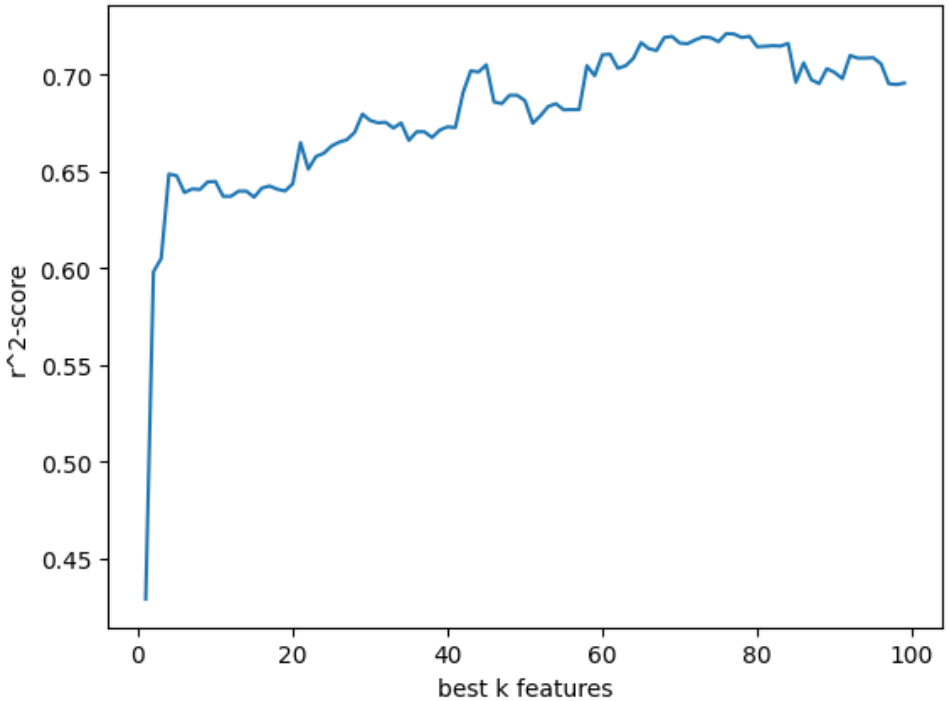
E’ la tecnica per identificare le grandezze maggiormente determinanti nella stima della capacità residua della cella

E’ un processo iterativo dei seguenti passaggi:

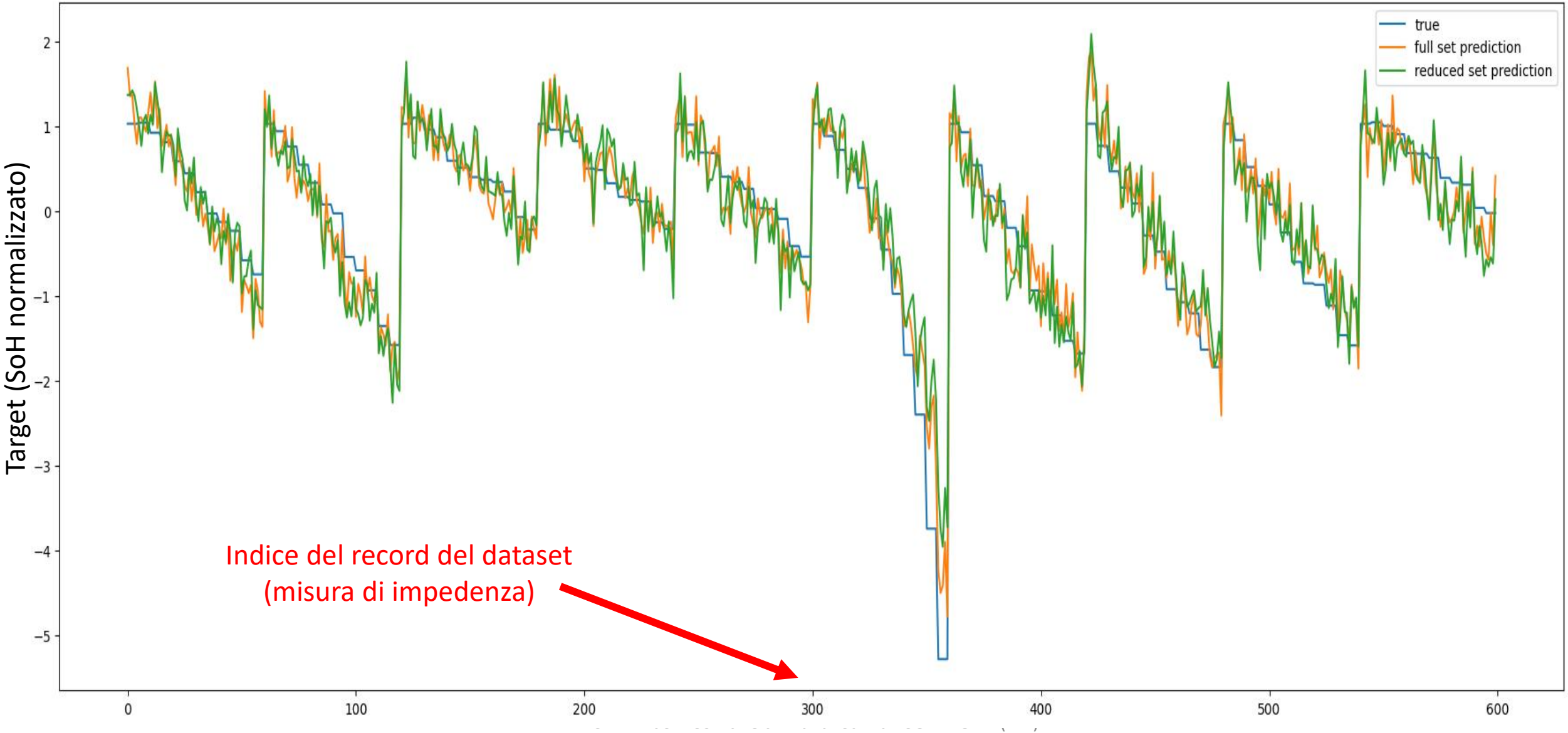
1) Costruzione del Modello Iniziale,
2) Valutazione dell'Importanza delle Caratteristiche,
3) Eliminazione delle Caratteristiche Meno Rilevanti

OSSERVAZIONI:

- E’ computazionalmente **onerosa** rispetto ad altri procedimenti
- Richiede un passaggio di **normalizzazione** propedeutica
- possiede il vantaggio **dell’interpretabilità** dei risultati (non usa variabili sintetiche)
- la **dimensione ridotta** del dataset iniziale consente di accettare un calcolo più oneroso
- l'obiettivo di ottenere una **procedura di stima come benchmark** rispetto ad altre tecniche rende accettabile un tempo di calcolo più lungo in funzione di una maggiore accuratezza attesa dei risultati.



Primo benchmark: *Confronto risultati da regressioni lineari a set completo e ridotto*



Principal Component Analysis (PCA) - Explained Variance (EV)

La **PCA** permette di sintetizzare un set di variabili per la rappresentazione della variabilità del dataset originario con un minor numero di features

L'indicatore usato per classificare le componenti è noto come "Explained Variance" (EV).

LA EV è una misura di quanta parte della varianza totale nel set di dati originale è "spiegata" (ossia contenuta rispetto al totale) da ciascuna componente principale.

Dal punto di vista del calcolo matriciale, la EV di una componente principale è uguale all'autovalore associato a quella componente.

In Sklearn PCA, è possibile accedere alla EV di ciascun componente principale tramite l'attributo explained_variance_[i], che fornisce la EV dell'i-esima componente principale

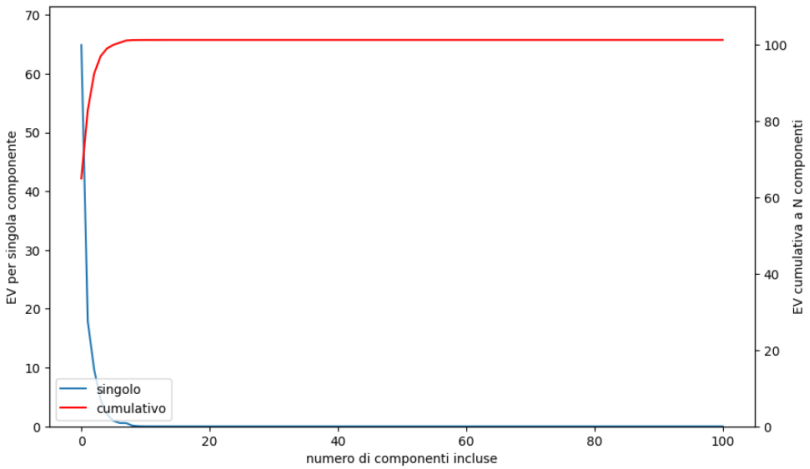
	Explained Variance (valore relativo) per singola componente e cumulata							
Componenti	1	2	3	4	5	6	7	8
Singola	0.641	0.177	0.095	0.045	0.02	0.009	0.006	0.005
Cumulata	0.641	0.818	0.913	0.958	0.978	0.987	0.993	0.998

Analisi ed elaborazione di dataset eterogenei in un formato interoperabile indirizzato allo sviluppo di tecniche di features selection e patterns recognition

RdS RICERCA DI SISTEMA

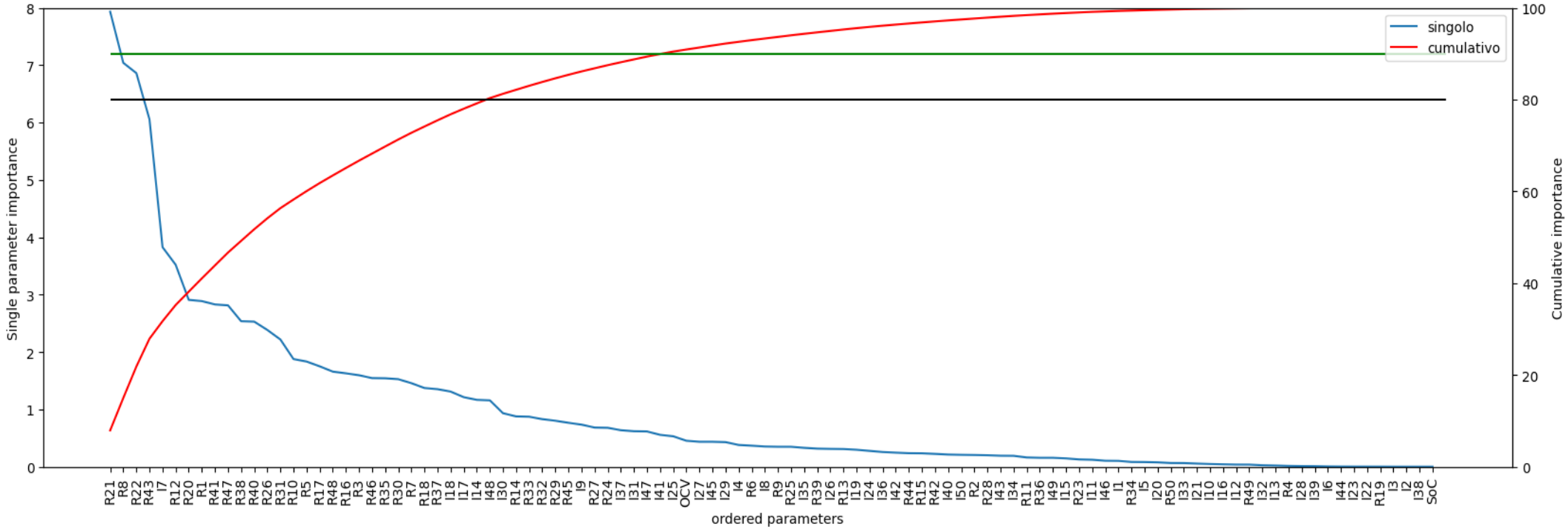
Consiglio Nazionale delle Ricerche

Identificazione dei parametri di maggiore contenuto informativo



Explained Variance dopo PCA

Explained Variance del set completo di parametri estratti



Confronto visuale con stime basate su tecniche di ML

Gaussian Process Regression

$$y_i = f(\mathbf{x}_i) + \epsilon_i$$

Ipotesi di rumore additivo

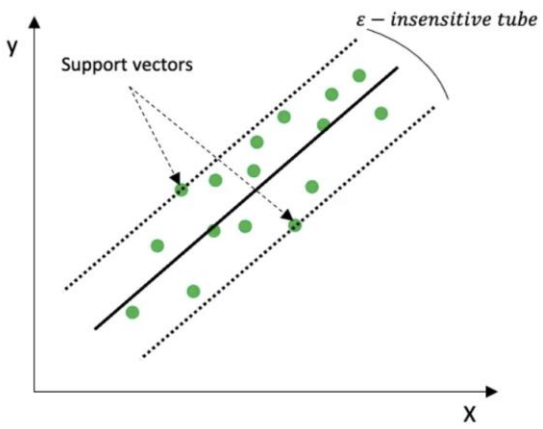
$$k(x, x') = \sigma_f^2 \exp\left[\frac{-(x - x')^2}{2l^2}\right] + \sigma_n^2 \delta(x, x')$$

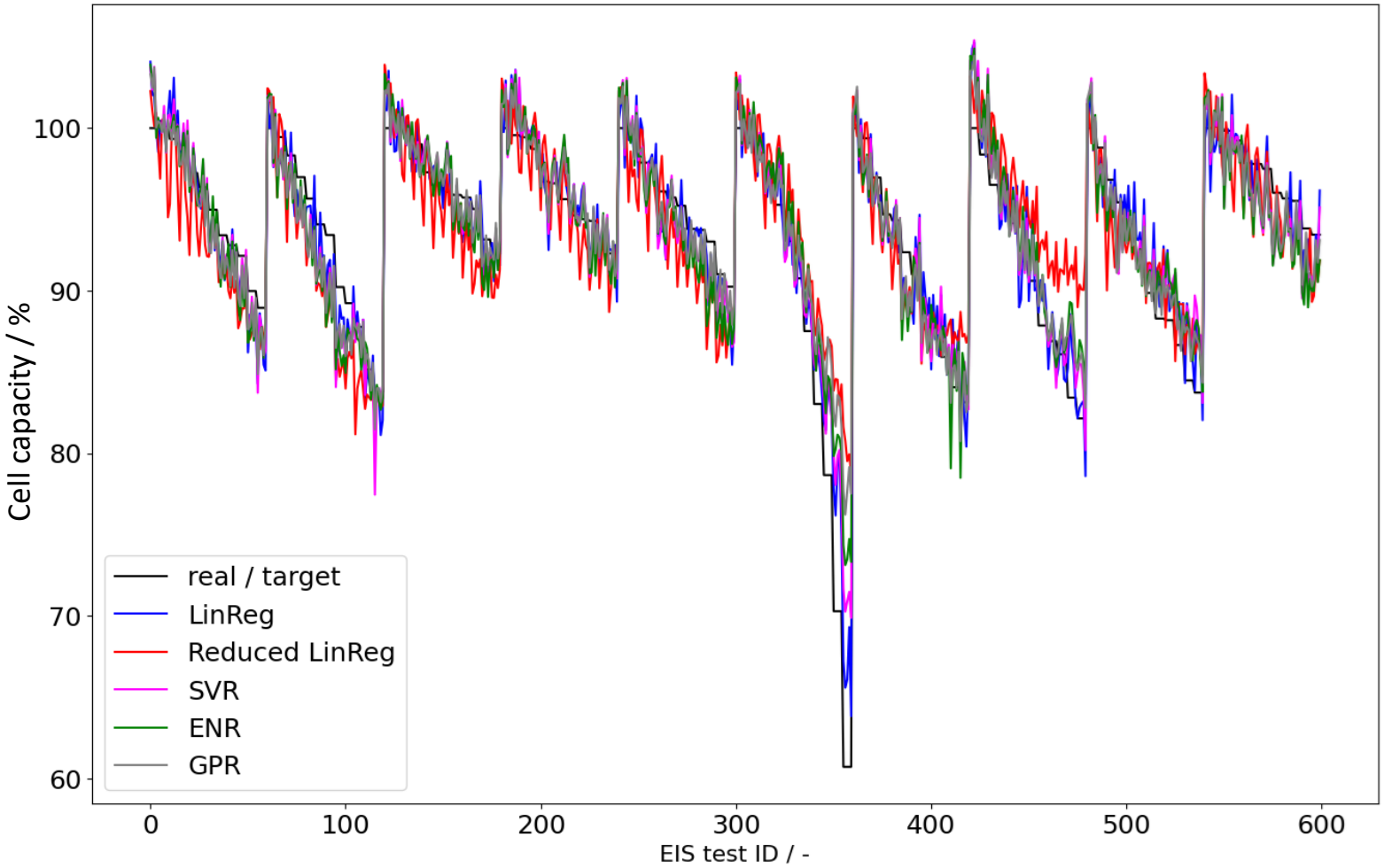
Elastic-Net Regression

Combinazione lineare delle metriche L1 e L2 (penalità "Lasso" e "Ridge" al variare di α)

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\left[(1-\alpha)||\beta||_2^2/2 + \alpha||\beta||_1\right]$$

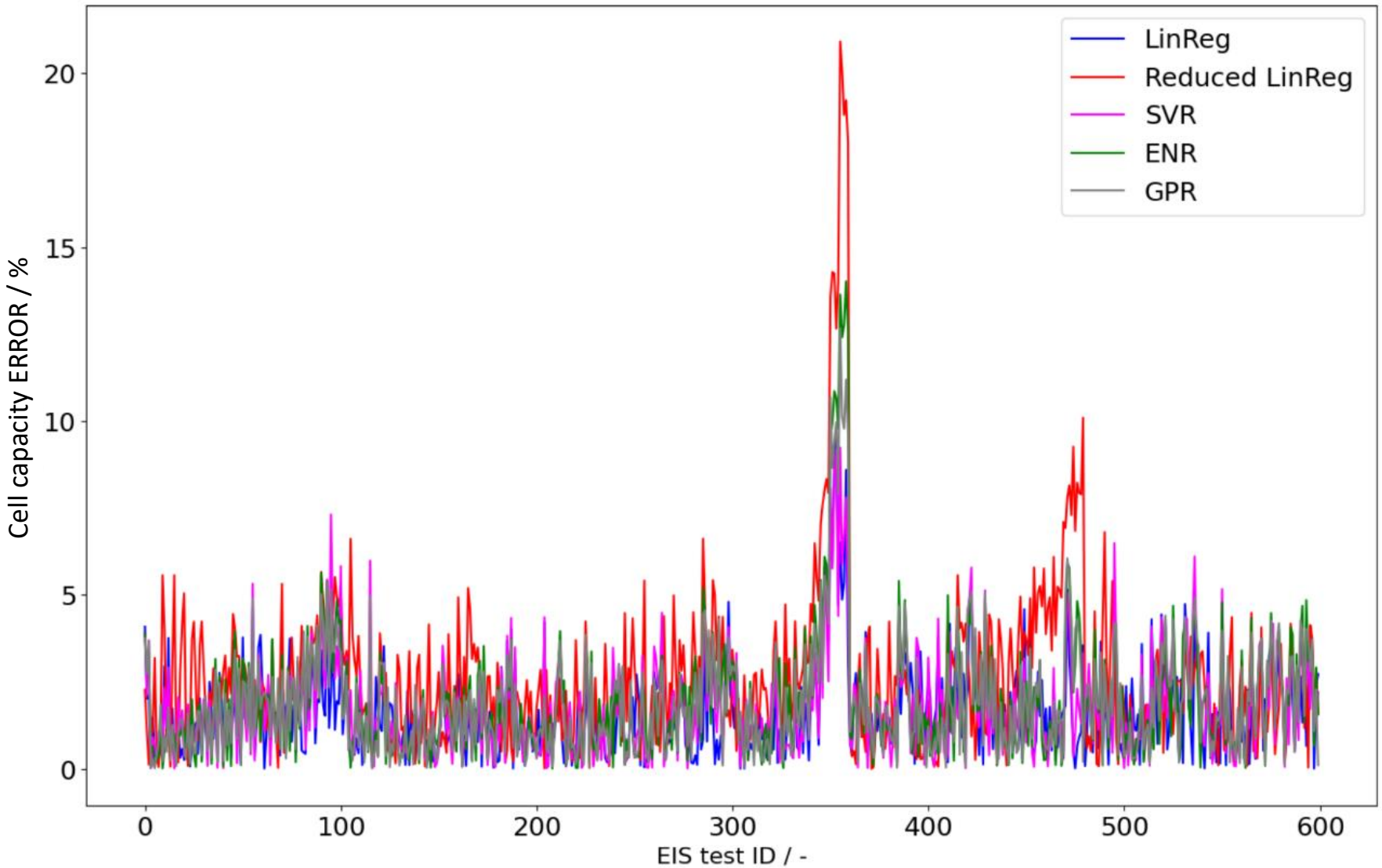
Support Vector Regression





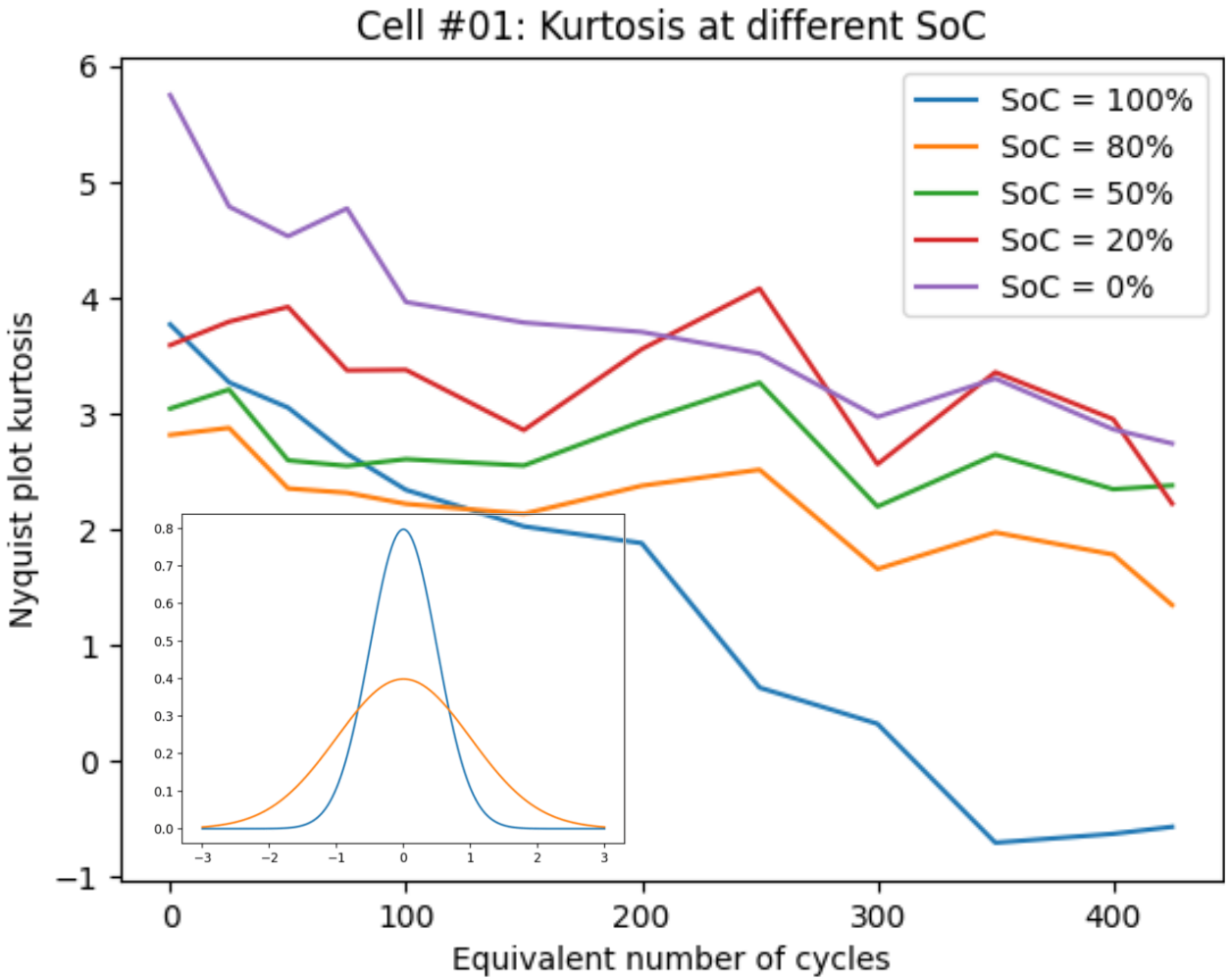
Analisi ed elaborazione di dataset eterogenei in un formato interoperabile indirizzato allo sviluppo di tecniche di features selection e patterns recognition

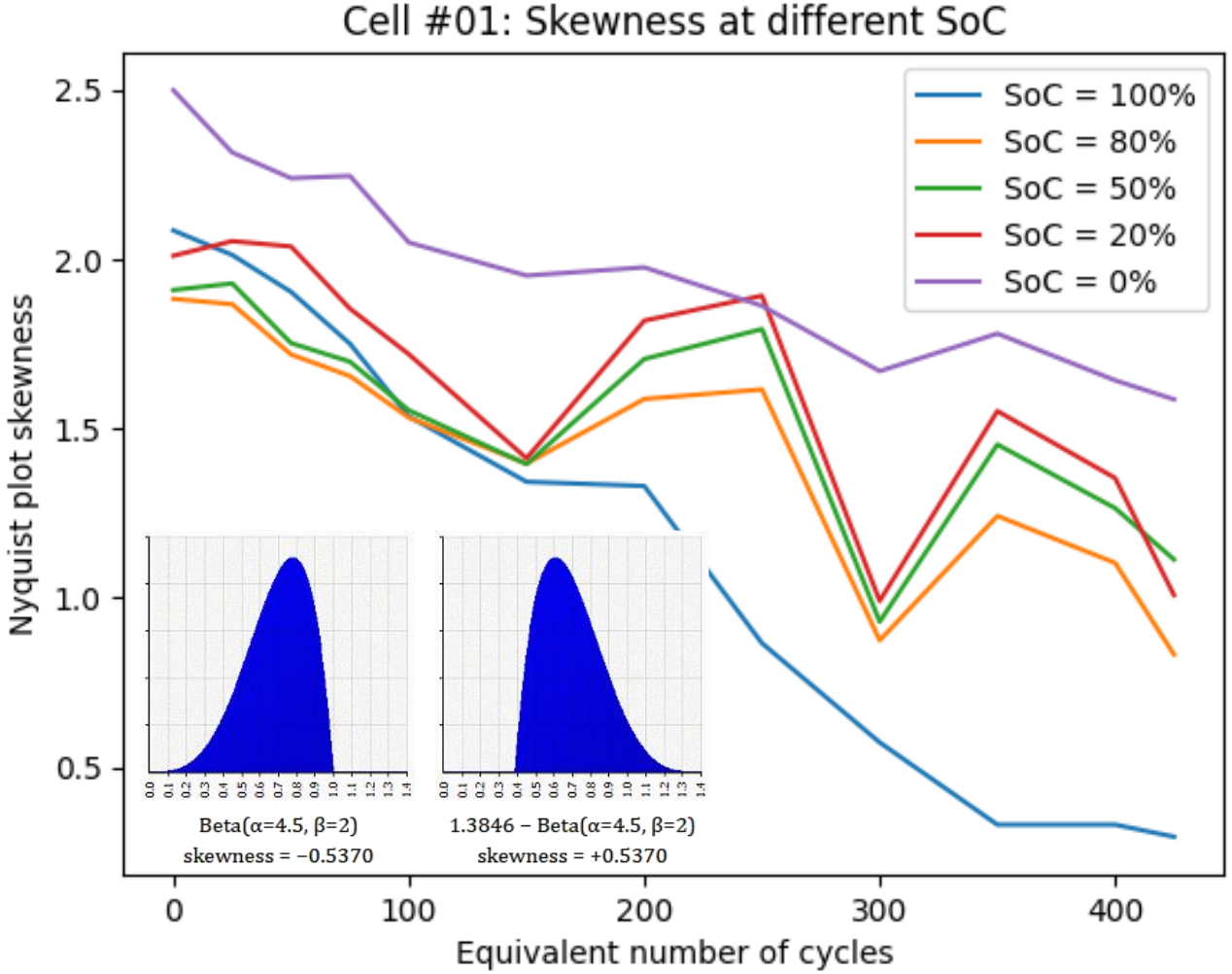
Valutazione quantitativa dell'errore commesso



Analisi ed elaborazione di dataset eterogenei in un formato interoperabile indirizzato allo sviluppo di tecniche di features selection e patterns recognition

Valutazione di Curtosi e Skewness per la ‘’cella n.1’’ del set da attività sperimentale





Grazie per l'attenzione

LA 2.1: Analisi ed elaborazione di dataset eterogenei in un formato interoperabile indirizzato allo sviluppo di tecniche di features selection e patterns recognition

Relatore: Giovanni Brunaccini