# Challenge Campus Biomedico

## Contesto: Telemedicina

La telemedicina è un servizio di supporto medico che consente l'interazione a distanza, tramite un dispositivo elettronico, tra pazienti e professionisti sanitari. La PNT (Piattaforma Nazionale di Telemedicina) ha l'obiettivo di governare e monitorare i processi di telemedicina condotti a livello regionale. L'obiettivo principale è armonizzare questi processi all'interno del sistema sanitario nazionale, coordinandoli con l'ecosistema digitale specifico di ciascuna regione.

Per raggiungere questo obiettivo, vengono forniti risorse e servizi per facilitare l'integrazione e lo sviluppo dei processi, con la possibilità di espandersi il più possibile nel territorio nazionale.

Alcuni degli obiettivi della telemedicina sono:
- Semplificare la gestione delle malattie croniche
- Promuovere la deospedalizzazione
- Migliorare la qualità clinica e l'accesso ai servizi sanitari
- Fornire ai professionisti sanitari nuovi strumenti innovativi

## Contesto: TeleAssistenza

Tra i vari servizi offerti dalla piattaforma, per presentare la sfida, considereremo il servizio di Teleassistenza.
La Teleassistenza è un servizio che prevede visite mediche tra pazienti e professionisti sanitari (infermieri, psicologi, educatori), mirate a fornire controlli e assistenza terapeutica. La piattaforma traccia ogni assistenza fornita e acquisisce tutti i dati necessari per garantire un'adeguata documentazione storica.

## Sfida: Cluster Associati a una Variabile di Esito

L'obiettivo della sfida è profilare i pazienti tenendo conto del loro contributo all'incremento del servizio di Teleassistenza.
È importante identificare schemi e comportamenti comuni in base all'aumento delle teleassistenze dovuto ai pazienti standard.
L'approccio prevede l'identificazione di gruppi di pazienti in relazione a una particolare variabile di esito/obiettivo.
In altre parole, si tratta di identificare gruppi di pazienti con schemi comuni o comportamenti simili in relazione alla variabile target (y=incremento_teleassistenze).

Successivamente, si analizzano le differenze tra i pazienti provenienti da vari gruppi di incremento, per comprendere quali caratteristiche portano all'aumento delle teleassistenze.
Per identificare questi gruppi (cluster), sono richiesti metodi di clustering avanzati che considerino sia le caratteristiche dei pazienti che la variabile di esito (incremento_teleassistenze).
In particolare, per la sfida sono richieste tecniche di clustering supervisionato.

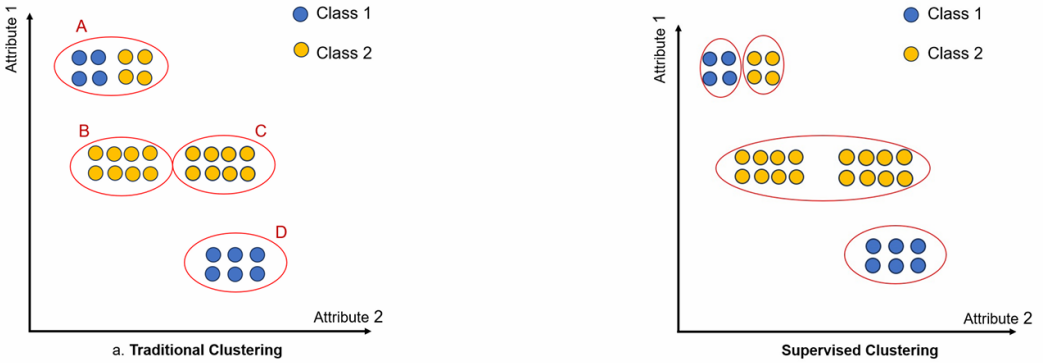
## Scopo dell'analisi

Attraverso l'analisi della profilazione dei pazienti, sarà possibile identificare ed estrarre informazioni rilevanti riguardanti i pazienti. Andando oltre le semplici statistiche aggregate, è possibile ottenere una visione più dettagliata e personalizzata dei loro comportamenti. Questo è un processo fondamentale per comprendere quali fattori influenzano la crescita o il cambiamento nell'utilizzo di questo servizio sanitario a distanza.

## Clustering Supervisionato 1/2

Il clustering tradizionale viene tipicamente applicato in un contesto di apprendimento non supervisionato utilizzando particolari funzioni di errore, ad esempio una funzione di errore che minimizza/massimizza le distanze/la probabilità all'interno di un cluster mantenendo i cluster compatti. Il clustering supervisionato (SC) utilizza dati etichettati come conoscenza preliminare o vincoli per guidare il clustering non supervisionato dei restanti dati non etichettati. Sfruttando i punti di forza degli approcci supervisionati e non supervisionati, il clustering supervisionato (SC) può migliorare la qualità, la coerenza e l'interpretabilità dei cluster quando si lavora con dati che presentano strutture complesse o ambigue.
In questa sfida, si presume di valutare le prestazioni del clustering supervisionato utilizzando le seguenti metriche:
- Purezza delle classi: misurata dalla percentuale di esempi della classe maggioritaria nei diversi cluster di un'iterazione di clustering.
- Qualità del cluster: misurata dal Silhouette Score utilizzando una funzione di distanza euclidea.
- Numero di cluster (k): preferibilmente da mantenere basso.

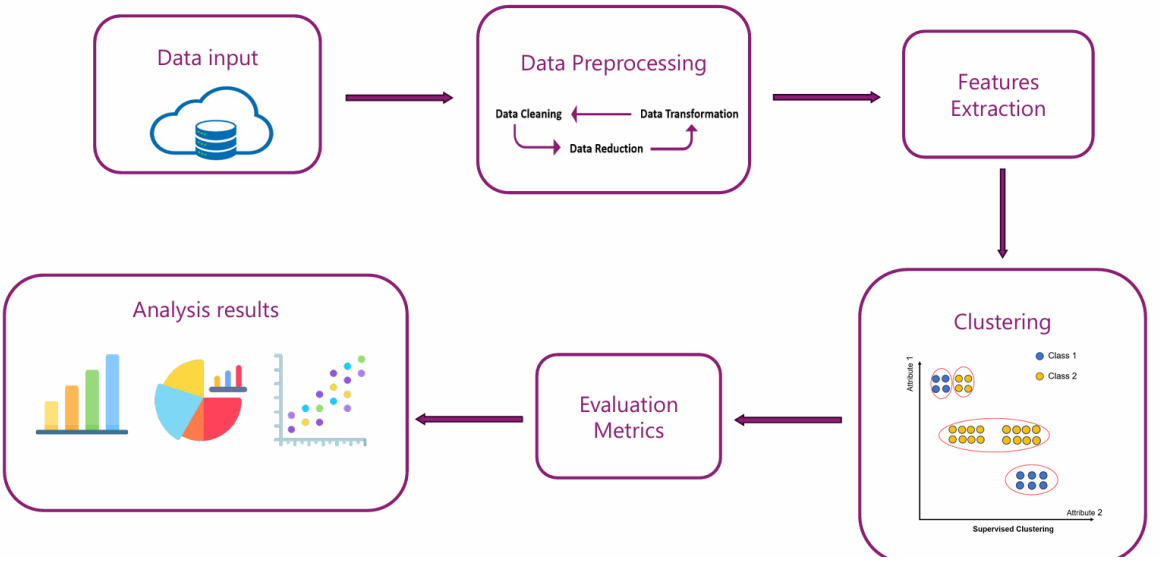
## Supervised Clustering 2/2




**Un algoritmo di clustering tradizionale** identificherebbe i quattro cluster rappresentati nel grafico superiore. Questo perché il clustering tradizionale non tiene conto dell'appartenenza degli esempi alle classi. Se l'obiettivo è generare riassunti per le classi 1 e 2, il clustering tradizionale non sarebbe molto attraente, poiché combinerebbe oggetti di classi diverse nel cluster A e metterebbe esempi della stessa classe in due cluster differenti, B e C.

**Un algoritmo di clustering supervisionato**, che massimizza la purezza delle classi, invece, dividerebbe il cluster tradizionale A in due cluster. Di conseguenza, i cluster tradizionali B e C verrebbero uniti in un unico cluster, senza compromettere la purezza delle classi e riducendo il numero di cluster. Il prototipo del cluster, che è una rappresentazione media dei punti all'interno di un cluster, fornisce informazioni utili per identificare le caratteristiche e la struttura del cluster stesso, e può anche essere utilizzato per identificare l'importanza della variabile target.

## Solution: Algorithmic Flow




## Data input 1/2

The data for this challenge are provided through PARQUET file, which contains the following features:

| Nome Variabile | Description | Type |
|---|---|---|
| **id_prenotazione** | Unique identifier of a single Teleassistance | String |
| **id_paziente** | Patient's unique identifier code | String |
| **data_nascita** | Patient's birth date | String |
| **sesso** | Patient's sex | String |
| **regione_residenza** | Patient's residence region | String |
| **codice_regione_residenza** | Patient's residence region code | String |
| **asl_residenza** | Patient's residence ASL | String |
| **codice_asl_residenza** | Patient's residence ASL code | String |
| **provincia_residenza** | Patient's residence province | String |
| **codice_provincia_residenza** | Patient's residence province code | String |
| **comune_residenza** | Patient's residence city | String |
| **codice_comune_residenza** | Patient's residence city code | String |
| **tipologia_servizio** | Typology of offered service from telemedicine platform | String |
| **descrizione_attivita** | Description of performed activity | String |
| **codice_descrizione_attivita** | Typology of performed activity's | String |
| **data_contatto** | Patient's contact date | String |

## Data input 2/2

| Nome Variabile | Description | Type |
|---|---|---|
| **regione_erogazione** | Service's erogation region | String |
| **codice_regione_erogazione** | Service's erogation region's code | String |
| **asl_erogazione** | Service's erogation ASL | String |
| **codice_asl_erogazione** | Service's erogation ASL code | String |
| **provincia_erogazione** | Service's erogation province | String |
| **codice_provincia_erogazione** | Service's erogation province code | String |
| **struttura_erogazione** | Service's erogation facility name | String |
| **codice_struttura_erogazione** | Service's erogation facility name's code | String |
| **tipologia_struttura_erogazione** | Service's erogation facility typology | String |
| **codice_tipologia_struttura_erogazione** | Service's erogation facility typology code | String |
| **id_professionista_sanitario** | Healtcare professional erogator's unique identifier code | String |
| **tipologia_professionista_sanitario** | Healtcare professional erogator's typology | String |
| **codice_tipologia_professionista_sanitario** | Healtcare professional erogator's typology code | String |
| **data_erogazione** | Service's erogation date | String |
| **ora_inizio_erogazione** | Service's erogation start timestamp (if already permormed) | String |
| **ora_fine_erogazione** | Service's erogation end timestamp (if already permormed) | String |
| **data_disdetta** | Service's erogation cancellation timestamp (if visit cancelled) | String |

# Preprocessing dei Dati

Il preprocessing dei dati è un insieme di tecniche fondamentali nell'apprendimento automatico che coinvolge la trasformazione dei dati grezzi in un formato "comprensibile per il machine learning". Le principali attività nel preprocessing dei dati sono:

### Pulizia dei dati

- Riempire i dati mancanti
- Smussare i dati rumorosi
- Identificare o rimuovere gli outlier
- Rimuovere i duplicati

### Trasformazione dei dati data transformation

- Normalizzazione
- Aggregazione

### Riduzione dei dati data reduction

- Ridurre il volume delle rappresentazioni (mantenendo risultati analitici uguali o simili)
- Rimuovere colonne ridondanti

# Feature Extraction

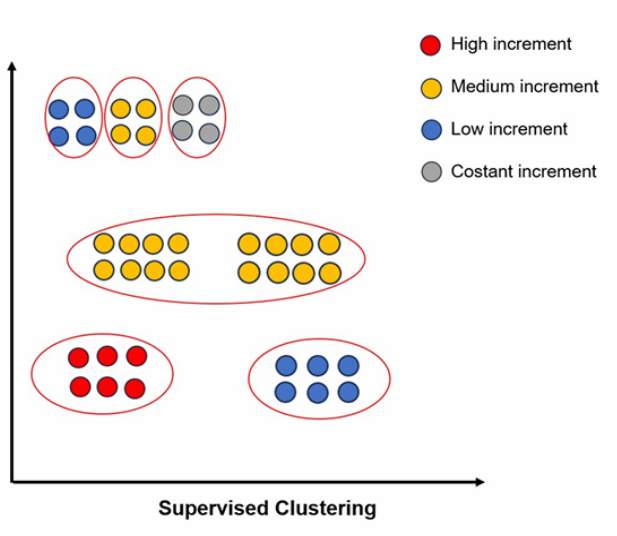
From existing features, extract additional ones to enhance the data analysis process.
Example

| Nome variabile | Descrizione |
|---|---|
| eta | Età del paziente al momento della visita. |
| durata_assistenza | Durata del servizio di Teleassistenza fornito al paziente |
| incremento | Differenza nel numero di servizi di Teleassistenza forniti, ad esempio tra i quadrimestri corrispondenti di un anno e quello successivo. |
| incremento_teleassistenze | La variabile target, calcolata a partire da incremento. |

**incremento_teleassistenze**: la variabile target deve essere discretizzata in 4 classi:
- Incremento costante
- Basso incremento
- Medio incremento
- Alto incremento

# Clustering

Raggruppare i pazienti in cluster omogenei basati sulle loro caratteristiche (schemi comuni o
comportamenti simili) e sul target di interesse (**incremento_teleassistenze**), con l'obiettivo di massimizzare sia la coerenza complessiva che la purezza del clustering.




# Metriche di Valutazione Proposte 1/3

**Purezza**: misura la purezza di ogni cluster, valutando quanto ciascun cluster contenga elementi della stessa classe. La metrica, per ogni cluster, identifica la classe più rappresentata e conta quanti elementi le appartengono. Successivamente, normalizza la somma dei valori ottenuti.
La purezza è calcolata come segue:

$$\text{Purezza} = \frac{1}{N} \sum_{k=1}^{K} \max\left(C_k \cap L_j\right)$$

- N: numero totale di elementi da raggruppare.
- k: numero di cluster.
- C_k: insieme di elementi nel cluster k.
- L_j: insieme di elementi appartenenti alla classe j.
- C_k \cap L_j: numero di elementi nel cluster k appartenenti alla classe j.
- $\max(C_k \cap L_j)$: numero massimo di elementi in C_k appartenenti alla stessa classe.

La purezza varia tra 0 e 1; un valore di 1 indica un clustering perfetto, in cui ogni cluster contiene solo elementi di una singola classe.

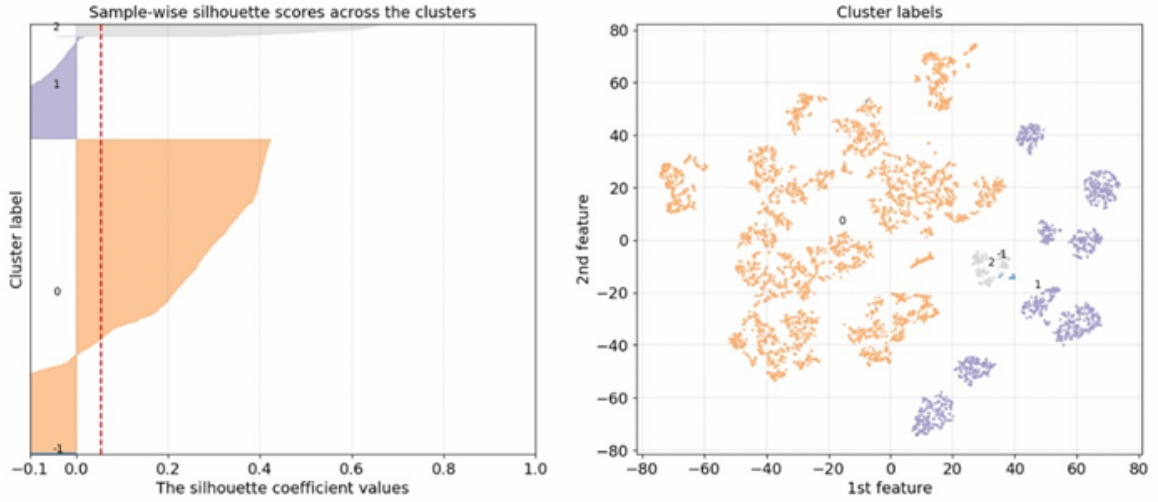
## Metriche di Valutazione Proposte 2/3

La metrica **Silhouette** può essere utilizzata per valutare l'accuratezza del clustering.

**Silhouette Score**: misura quanto un oggetto è simile al proprio cluster (coesione) rispetto agli altri cluster (separazione). La silhouette varia da -1 a +1. Per renderla più interpretabile, viene spesso normalizzata a un intervallo tra 0 e 1.

Un punteggio di silhouette normalizzato di:

- 1 indica che i campioni sono assegnati al cluster corretto.
- 0.5 indica che i campioni si trovano al confine tra due cluster.
- 0 indica che i campioni potrebbero essere stati assegnati al cluster sbagliato.




## Metriche di Valutazione Proposte 3/3

Per questa sfida:

1. **Calcolo delle metriche per valutare il clustering supervisionato**: Utilizzare il punteggio di Silhouette per la coerenza del clustering e la purezza per la corretta classificazione dell'incremento in ogni cluster.
2. **Normalizzazione delle metriche tra 0 e 1**: Assicurarsi che le metriche per entrambi i compiti siano normalizzate, in modo che rientrino tra 0 e 1. Questo passaggio è importante per garantire che le metriche siano comparabili.
3. **Calcolo della metrica finale**: Calcolare la media delle due metriche normalizzate e sottrarre un termine di penalizzazione pari a 0,05 volte il numero di cluster per ottenere una valutazione complessiva.

# Risultati dell'Analisi

Una parte fondamentale del lavoro di un data scientist è la capacità di narrare e descrivere i risultati (**storytelling**), poiché fornisce intuizioni cruciali per prendere decisioni informate, personalizzare le strategie aziendali e migliorare la comunicazione con i clienti.

## Visualizzazione della Distribuzione delle Caratteristiche per i Cluster e Interpretazione dei Risultati

Fornire grafici rappresentativi per i cluster, in modo da visualizzare la distribuzione delle caratteristiche all'interno dei diversi cluster. Questa analisi sarà utile per comprendere come le varie caratteristiche sono distribuite tra i gruppi identificati dal clustering. Inoltre, interpretare e validare i risultati ottenuti.