**Operazioni Effettuate – Dataset ‘sdo\_2023.csv’**

**1. Scopo del Documento**

Il presente documento ha l’obiettivo di illustrare in maniera dettagliata le operazioni di *pre-processing* effettuate sul dataset **‘sdo\_2023\_semianonim.csv’**, che rappresenta una raccolta strutturata di informazioni sanitarie anonime.  
L’attività è stata condotta in vista della pubblicazione dei dati in formato **Open Data**, con l’intento di renderli maggiormente accessibili, comprensibili e riutilizzabili da parte di ricercatori, analisti, istituzioni sanitarie, cittadini e stakeholder pubblici e privati.

Il trattamento dei dati si è concentrato principalmente sulla **miglior qualità informativa**, sulla **standardizzazione semantica**, sulla **rimozione di ambiguità**, nonché sulla **coerenza strutturale** dell’intero dataset. Questi interventi risultano fondamentali per garantire un utilizzo efficace dei dati sia in ambito clinico che statistico.

**2. Fasi Principali del Pre-processing**

L’intervento di pre-processing ha previsto una serie di fasi consecutive, ciascuna mirata a migliorare un aspetto specifico della struttura e del contenuto del dataset:

* **Analisi iniziale**
* **Verifica e tipizzazione dei dati**
* **Rinomina e riformulazione dei nomi delle colonne**
* **Decodifica e mapping dei valori**
* **Pulizia semantica e formattazione**
* **Gestione dei dati duplicati**
* **Controllo delle dimensioni e della coerenza**

Ognuna di queste fasi è descritta nei paragrafi seguenti.

**3. Analisi Iniziale e Statistiche Descrittive**

Dopo il caricamento del file sdo\_2023\_semianonim.csv, è stata condotta un’analisi esplorativa preliminare con lo scopo di ottenere una visione generale della struttura interna del dataset. In particolare:

* Sono state analizzate **71 colonne** iniziali, corrispondenti a variabili di tipo anagrafico, clinico e amministrativo.
* È stato rilevato che molte colonne contenevano valori **codificati numericamente**, il che ne comprometteva la leggibilità.
* È stato eseguito un primo censimento dei **tipi di dati**, con classificazione in int64, float64 e object (stringhe o categorie).

Questo passaggio ha consentito di impostare le linee guida per la successiva riformulazione semantica e per l’uniformazione dei tipi.

**4. Tipizzazione dei Dati**

L’analisi dei tipi ha evidenziato alcune incongruenze semantiche:

* Alcuni campi numerici risultavano correttamente tipizzati, come ad esempio:
  + Anno della scheda di ricovero
  + Numero progressivo della scheda
  + Scheda del parto
* Tuttavia, molte colonne presentavano **codici numerici rappresentati come stringhe**, ad esempio:
  + Genere del paziente
  + Livello di istruzione del paziente

Questi elementi sono stati mantenuti temporaneamente come object in attesa della **decodifica e trasformazione in etichette testuali**, così da migliorarne la leggibilità.

**5. Rinomina e Standardizzazione dei Nomi delle Colonne**

Uno dei principali interventi di miglioramento ha riguardato la riformulazione dei nomi delle colonne. Sono state applicate le seguenti trasformazioni:

* Conversione in **Title Case** (prima lettera maiuscola per ogni parola)
* Rimozione di simboli, parentesi e caratteri speciali
* Eliminazione degli spazi superflui
* Uso di terminologia coerente, chiara e descrittiva

**Esempi di modifiche:**

* 'REGIONE' → 'Regione'
* 'Genere del paziente (es. 1 per maschio, 2 per femmina)' → 'Genere Del Paziente Es 1 Per Maschio 2 Per Femmina'

Nel complesso, **tutte le colonne finali** sono state modificate, migliorando significativamente la leggibilità e l’interoperabilità del dataset.

**6. Decodifica dei Valori Codificati e Mapping Semantico**

Una parte fondamentale dell’intero processo ha riguardato la sostituzione dei codici numerici presenti all’interno del dataset con **etichette testuali esplicite**. Questo intervento ha incluso:

* **REGIONE**: 140 → *Molise*
* **SESSO**: 1 → *Maschio*, 2 → *Femmina*
* **ICD, Diagnosi, Procedure e DRG**: codici clinici decodificati con la descrizione corrispondente
* **CODCAUSAES**: categorie esplicite (es. *incidenti stradali*, *avvelenamenti*, *suicidi*, ecc.)
* **ISTRUZIONE**: codici numerici sostituiti con livelli scolastici (es. *Licenza media*, *Diploma*, *Laurea*)
* **Priorità del ricovero**: tradotta in termini clinici umani (es. *Urgente*, *Programmato*)
* **TIPO\_RIC** e **Provenienza**: etichette mediche rese in linguaggio comprensibile

**Obiettivi di questa trasformazione:**

* Rendere il dataset leggibile anche a non esperti del dominio sanitario
* Prevenire errori interpretativi in fase di analisi
* Allinearsi agli standard nazionali (ICD9, DRG, codifiche ISTAT)
* Preparare i dati per analisi descrittive, esplorative e predittive

**7. Verifica delle Dimensioni del Dataset**

Dopo il pre-processing, il dataset risultava composto da:

* **38.400 righe**
* **42 colonne pulite, tipizzate e documentate**

Questa dimensione conferma la ricchezza del dataset, adatta ad analisi complesse su larga scala, incluse elaborazioni cliniche, gestionali e di pianificazione sanitaria regionale o nazionale.

**8. Controllo dei Duplicati**

È stato condotto un controllo sull’intera tabella al fine di identificare eventuali righe duplicate.  
Il risultato dell’analisi è stato positivo: **nessuna duplicazione è stata riscontrata**; pertanto, non si è resa necessaria alcuna rimozione.

**9. Estrazione di nuovi dataset**

Dal dataset relativo alle schede di dimissioni generali sono stati estratti due dataset specifici per quanto riguarda la parte anagrafica e il processo di trasferimento tra i vari reparti.

**10. Considerazioni Finali e Conclusione**

L’intero processo di pre-processing ha trasformato un dataset originariamente tecnico, frammentario e codificato in un **dataset documentato, coerente e leggibile**.

Tali operazioni, svolte con un approccio metodico e orientato alla qualità, hanno reso il dataset adatto alla pubblicazione come **Open Data**, promuovendone la trasparenza e l’accessibilità.  
Grazie alla rinomina, al mapping semantico e alla standardizzazione, oggi il file può essere impiegato efficacemente in:

* Studi statistici a livello nazionale o regionale
* Analisi epidemiologiche
* Supporto a decisioni politiche e sanitarie
* Progetti di ricerca accademica o data journalism