**Operazioni Effettuate – ‘Dataset SDO\_2024\_per\_OpenDATA.csv’**

Lo scopo di questo documento è descrivere in modo dettagliato il processo di pre-processing applicato al dataset ‘*SDO\_2024\_per\_OpenDATA.csv’*.  
Il trattamento dei dati è stato realizzato per garantire la qualità, l’affidabilità e la chiarezza delle informazioni prima della pubblicazione come Open Data, facilitandone la comprensione e il riutilizzo da parte di utenti e stakeholder esterni.

Le operazioni di pre-processing includono una serie di fasi mirate alla pulizia, standardizzazione, arricchimento e verifica del dataset. Questi passaggi sono fondamentali per assicurare che i dati siano coerenti, privi di errori e ben documentati, elementi indispensabili per la trasparenza e la fruibilità tipiche di un dataset aperto**.**

**Statistiche Pre e Post Processing**

Inizio Pre-processing

* Numero di righe: 516
* Numero di colonne: 2
* Dimensioni del dataset: *(515, 2)*
* Max Numero Ricoveri: 2973, Min: 1

Fine Pre-processing

* Numero di righe: 516
* Numero di colonne: 3
* Dimensioni del dataset: *(516, 3)*
* Max Numero Ricoveri: 2973, Min: 1

**1. Controllo e Standardizzazione delle Intestazioni**

In questa fase preliminare, le intestazioni del dataset sono state analizzate per garantirne **coerenza**, **uniformità** e **chiarezza**. Sono stati rimossi eventuali caratteri speciali e spazi superflui per evitare ambiguità e facilitare la gestione dei dati nelle fasi successive.

È stata applicata la convenzione **Title Case** (prima lettera maiuscola di ogni parola) per migliorare la leggibilità.  
Nel dettaglio, i nomi delle colonne sono stati esaminati e, dove necessario, rinominati per migliorarne la descrizione e renderli più intuitivi, specialmente per utenti non specialisti.

**Modifiche effettuate:**

* Drg→ **Gruppo Diagnostico Correlato**
* ricoveri 2024→ **Numero Di Ricoveri**

Questa operazione ha assicurato una nomenclatura chiara e omogenea, indispensabile per la comprensione e l’uso corretto del dataset.

**2. Pulizia dei Valori Testuali**

È stata effettuata un’accurata pulizia dei dati testuali presenti nel dataset. Questo intervento ha previsto:

* Rimozione di spazi bianchi iniziali e finali
* Eliminazione di caratteri speciali indesiderati che potrebbero compromettere le analisi o la visualizzazione

Dall’analisi automatica delle colonne testuali è emerso che nessun valore presentava spazi superflui o anomalie. Sebbene non siano state necessarie modifiche sostanziali, questa fase ha consolidato la qualità complessiva del dataset, prevenendo possibili errori invisibili che avrebbero potuto influire sulle elaborazioni future.

**3. Uniformazione dei Valori Testuali**

Per migliorare la **coerenza visiva** e la **leggibilità** dei dati testuali, è stata applicata l’uniformazione dei valori nel formato **Title Case** (prima lettera maiuscola di ogni parola).

Questa scelta:

* Riduce le discrepanze causate da formattazioni incoerenti (es. “ricoverato” vs “Ricoverato”)
* Favorisce una migliore esperienza d’uso in consultazioni e report
* Facilita l’identificazione univoca dei valori categoriali

Dall’analisi è emerso che soltanto una colonna ha richiesto questa trasformazione, mentre le altre risultavano già coerenti o prive di interventi necessari.

Pur essendo un intervento quantitativamente minimale, contribuisce significativamente a migliorare la qualità semantica e l’ordine del dataset.

**4. Rimozione di Righe Duplicate**

È stata effettuata una verifica approfondita per individuare eventuali righe duplicate, al fine di evitare distorsioni nelle analisi statistiche, aggregazioni o modelli predittivi.

La scansione ha esaminato l’intero dataset, confrontando ogni riga per rilevare duplicazioni esatte.  
Risultato: nessuna riga duplicata è stata trovata, confermando l’unicità di ogni osservazione e garantendo così l’integrità del dataset.

Questa verifica, pur senza richiedere azioni correttive, è un passaggio chiave per assicurare la qualità e l’affidabilità complessive dei dati.

**5. Gestione dei Valori Nulli**

Un’analisi completa è stata svolta per identificare la presenza di eventuali valori nulli o mancanti, operazione essenziale per evitare lacune che potrebbero compromettere analisi o modellazioni future. L’indagine ha riguardato tutte le colonne, sia numeriche che testuali quindi il dataset risulta completo e consistente, pronto per le fasi successive di analisi**.**

**6. Arricchimento con Variabili Derivate (Feature Engineering)**

È stata introdotta una nuova variabile denominata **Frequenza Ricovero**, derivata dalla colonna **Numero Di Ricoveri** mediante un processo di binning.  
La logica adottata prevede la suddivisione dei valori numerici in tre fasce di frequenza, definite sulla base della distribuzione dei dati (minimo: 1, massimo: 2973), con particolare attenzione ai quartili:

**Intervalli e Categorie:**

* **Bassa**: da 1 a 33 ricoveri
* **Media**: da 34 a 99 ricoveri
* **Alta**: da 100 a 2973 ricoveri

**Obiettivi e Benefici:**

* Riduzione della complessità numerica nella variabile originaria
* Facilitazione nell’analisi comparativa tra gruppi omogenei di strutture o pazienti
* Introduzione di una dimensione qualitativa utile per l’esplorazione dei dati e la modellazione predittiva
* L’arricchimento si è concluso con successo, aggiungendo la colonna **Frequenza Ricovero** al dataset

**8. Esportazione Finale**

Al termine delle operazioni di pulizia, trasformazione e arricchimento, il dataset è stato esportato nel formato CSV con il nome: ‘Schede Di Dimissioni Ospedaliere 2024.csv’.

Motivazioni della scelta del formato CSV:

* Massima compatibilità con software di analisi dati, fogli di calcolo (Excel, Google Sheets), strumenti di visualizzazione e ambienti di programmazione (Python, R, etc.)
* File leggero e facilmente leggibile anche da utenti non esperti
* Formato standard aperto, ideale per la pubblicazione come Open Data, senza dipendenza da software proprietari

Questa operazione rappresenta un passaggio cruciale nella pipeline di gestione dati, in quanto:

• Consolida tutte le modifiche effettuate in una versione definitiva e documentata.

• Rende il dataset pronto per la distribuzione, l’analisi o l’integrazione in portali open data e sistemi informativi.

• Garantisce trasparenza e tracciabilità dei dati trattati, supportando la riproducibilità delle analisi future.

Il dataset così processato è stato esportato in un file finale, pronto per la pubblicazione come Open Data. Il formato è stato scelto per garantire ampia compatibilità e facilità di accesso, accompagnato da documentazione esplicativa per agevolare la comprensione da parte degli utenti.

**Conclusioni**

Il processo di pre-processing ha apportato un miglioramento sostanziale alla qualità, leggibilità e coerenza del dataset *‘SDO\_2024\_per\_OpenDATA.csv’.*  
Ogni fase è stata progettata per garantire:

* Trasparenza nelle trasformazioni
* Affidabilità dei dati
* Usabilità per utenti eterogenei

Il risultato è un dataset pulito, arricchito, documentato e pronto per essere pubblicato come Open Data.  
Può essere ora utilizzato in contesti di analisi statistica, ricerca, monitoraggio istituzionale e nello sviluppo di strumenti decisionali data-driven.