Preprocessing Documentation

Summary

The main process was to consolidate raw **Device**, **Manufacturer**, and **Event** tables into a single cleaned dataset with a usable **Failure Severity Class** column. This serves as a foundation for downstream modeling and analysis.

1. Dataset Integration

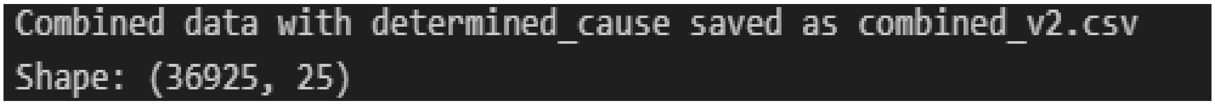
Three source datasets were used:

- **Devices Table** – device details (ID, type, manufacturer reference).
- **Events Table** – adverse events, recall information, actions, causes.
- **Manufacturers Table** – manufacturer information.

Integration approach:

- Performed **INNER JOINs** on:
    - `devices.id = events.device_id`
    - `devices.manufacturer_id = manufacturers.id`

This ensured only consistent, valid records were kept, yielding one unified dataset.



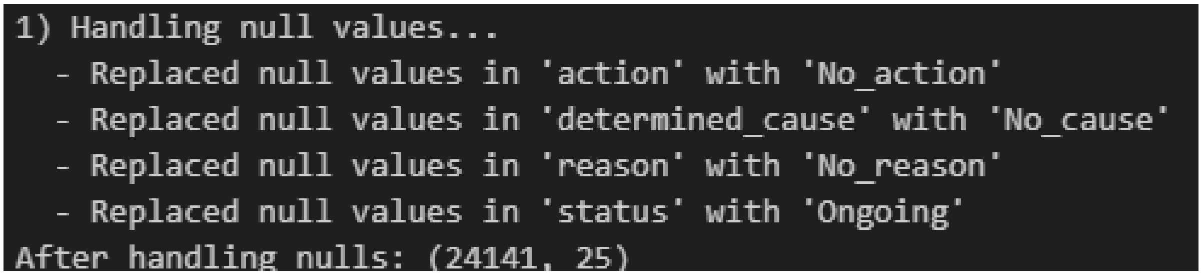
```
Combined data with determined_cause saved as combined_v2.csv
Shape: (36925, 25)
```

2. Data Cleaning & Preprocessing

- **Handling Missing Values**
    - Inspected all columns for nulls.
    - Removed records missing essential keys or event details.
- **Duplicate Removal**

- ○ Identified and dropped redundant rows to prevent bias.

- **Column Standardization**
  - ○ Normalized categorical fields: `action_classification`, `recall_level`, `risk_class`, `type`, `reason`.
  - ○ Trimmed string whitespace and corrected inconsistent labels.

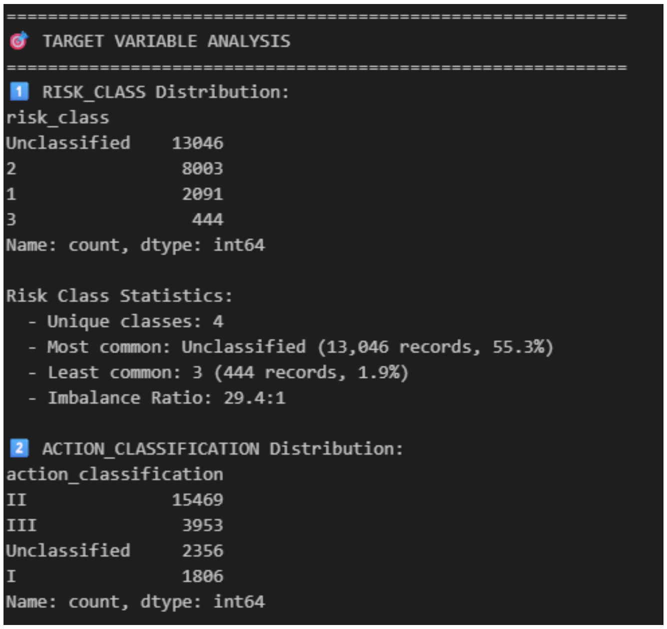


```
1) Handling null values...
  - Replaced null values in 'action' with 'No_action'
  - Replaced null values in 'determined_cause' with 'No_cause'
  - Replaced null values in 'reason' with 'No_reason'
  - Replaced null values in 'status' with 'Ongoing'
After handling nulls: (24141, 25)
```

3. Feature Engineering

- **recall_level** – standardized regulatory recall levels (Class 1–3).
- **risk_class** – derived risk/severity category.
- **action_classification** – normalized manufacturer actions (e.g., FSN, Recall, Safety Alert).
- **determined_cause & reason** – cleaned text fields describing root causes.
- **type** – standardized device type categories.

Intial imbalanced:



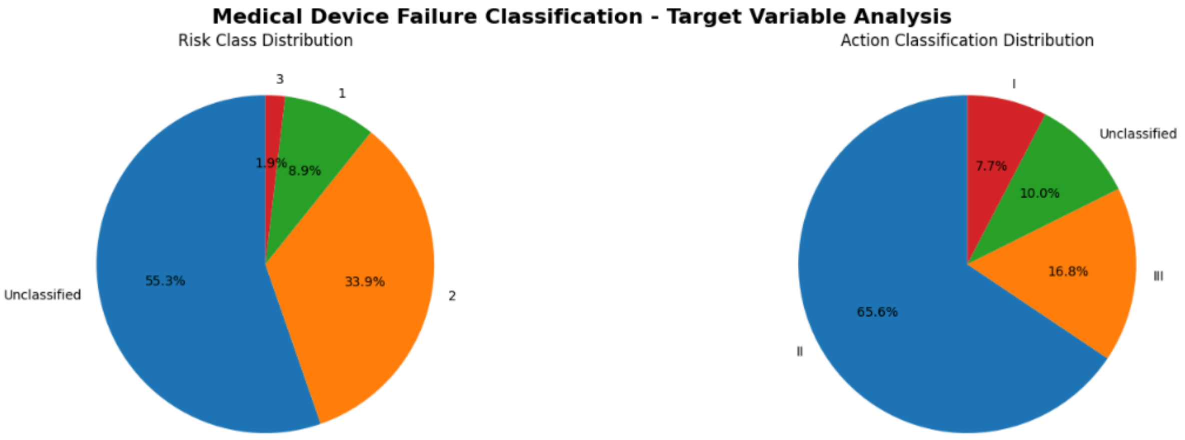
```
============================================================
🎯 TARGET VARIABLE ANALYSIS
============================================================
1️⃣ RISK_CLASS Distribution:
risk_class
Unclassified    13046
2                8003
1                2091
3                 444
Name: count, dtype: int64

Risk Class Statistics:
  - Unique classes: 4
  - Most common: Unclassified (13,046 records, 55.3%)
  - Least common: 3 (444 records, 1.9%)
  - Imbalance Ratio: 29.4:1

2️⃣ ACTION_CLASSIFICATION Distribution:
action_classification
II              15469
III              3953
Unclassified     2356
I                1806
Name: count, dtype: int64
```

5. Exploratory Data Analysis (EDA)

- Computed frequency distributions of `recall_level`, `risk_class`, and `action_classification`.
- Visualized class distributions to assess imbalance.
- Reduced dataset size: from **~124k raw rows → ~23k valid, cleaned rows**.



4. Target Variable Definition

Defined **Failure Severity Class** as the supervised learning target:

- **Class 1** – Most severe (life-threatening or serious injury).
- **Class 2** – Moderate severity.
- **Class 3** – Low severity.
- **FSN (Field Safety Notice)** – manufacturer advisory.
- **SA (Safety Alert)** – regulatory/public safety communication.

6. Final Output

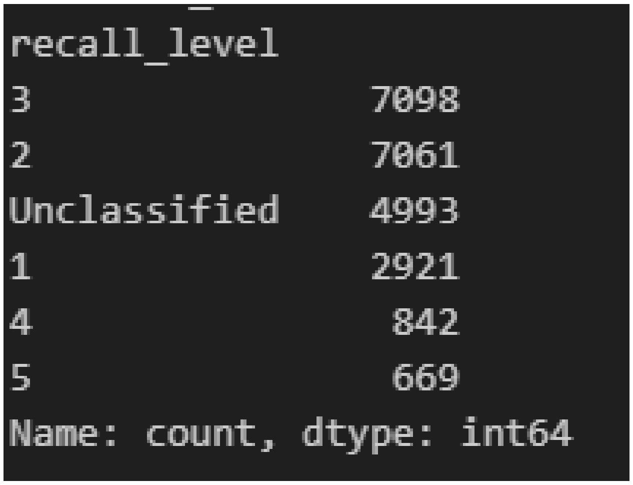
The preprocessed dataset includes the following key fields:

`id, action, action_classification, action_summary, country, determined_cause, reason, status, type, device_id, manufacturer, recall_level, risk_class, Failure Severity Class`.

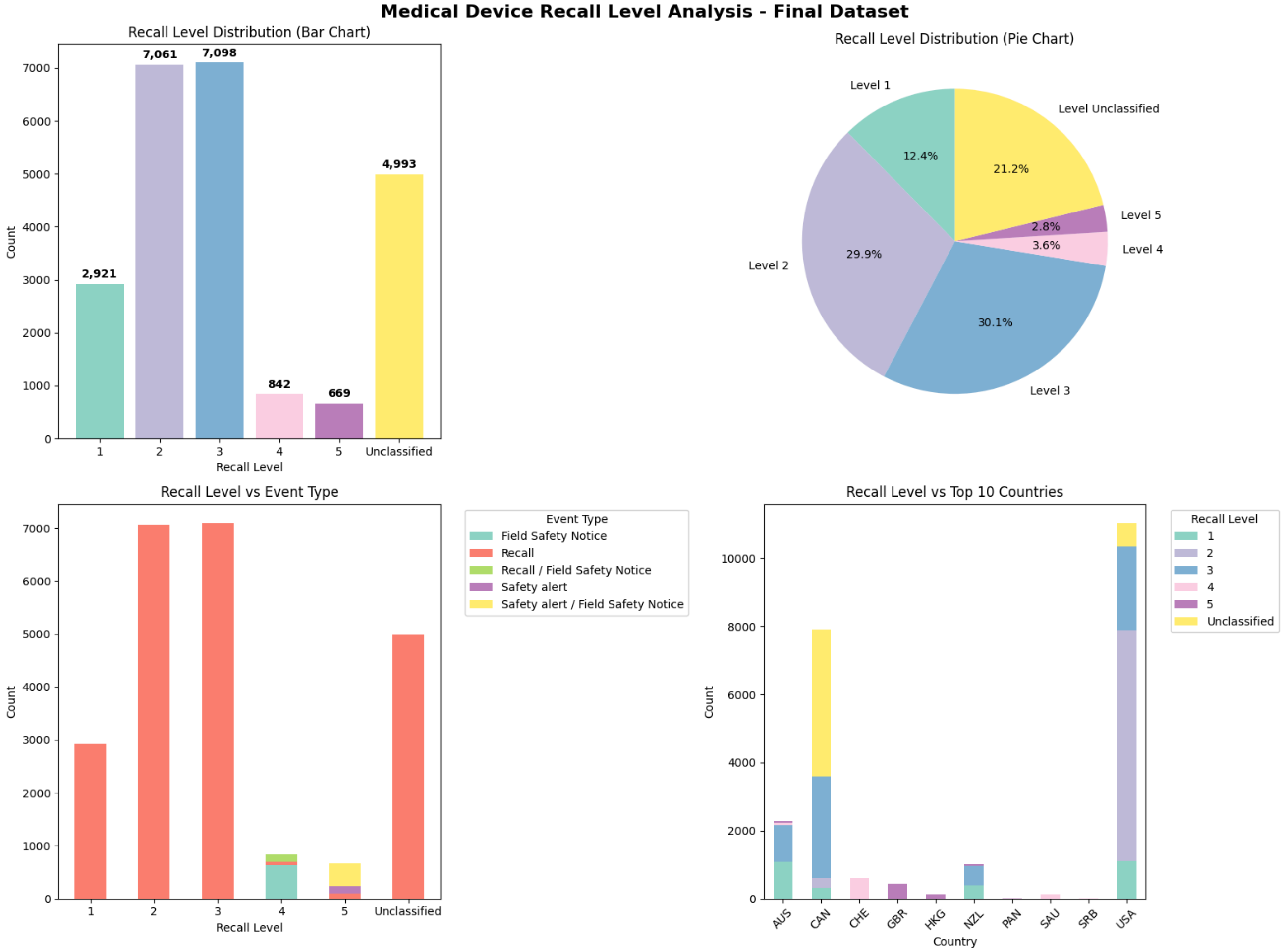
This dataset is suitable for:

- **Severity Classification** (multi-class: Class 1, 2, 3, FSN, SA).
- Future predictive modeling of medical device failure severity.

- Final Cleaned without nlp;



```
recall_level
3               7098
2               7061
Unclassified    4993
1               2921
4                842
5                669
Name: count, dtype: int64
```


Medical Device Recall Level Analysis - Final Dataset
Recall Level Distribution (Bar Chart)
Count
Recall Level
Unclassified
Recall Level Distribution (Pie Chart)
Level 1
Level Unclassified
Level 5
Level 4
Level 3
Level 2
Recall Level vs Event Type
Count
Recall Level
Unclassified
Event Type
Field Safety Notice
Recall
Recall / Field Safety Notice
Safety alert
Safety alert / Field Safety Notice
Recall Level vs Top 10 Countries
Count
Country
AUS
CAN
CHE
GBR
HKG
NZL
PAN
SAU
SRB
USA
Recall Level
Unclassified