Elevate Labs – Day 1 Assignment Report

Data Analytics Internship

Prepared by: Busireddy Harshitha
Assignment: Data Cleaning & Preprocessing

1. Introduction

This report presents the detailed work completed as part of Day 1 in the Data Analytics Internship at Elevate Labs. The objective of this assignment was to clean and preprocess the raw dataset titled Medical Appointment No Shows. Data cleaning ensures accuracy, consistency, reliability, and readiness for exploratory analysis and modeling.

2. Dataset Overview

The dataset contains medical appointment information, including patient demographics, health indicators, scheduling details, and appointment outcomes. The key target variable, no_show, indicates whether a patient missed their appointment.

3. Issues Identified in Raw Data

The following issues were identified during initial exploration:

• Non-standard column names
• Invalid/negative age values
• Missing and malformed date fields
• Float-formatted IDs
• Duplicate appointment records
• Text inconsistencies (gender, neighbourhood)
• Binary columns with mixed formats

4. Cleaning and Preprocessing Steps

The following steps were executed to clean the dataset:

4.1 Column Standardization: Clean snake_case naming applied.

4.2 Data Type Correction: IDs converted to integers, gender to categories, binary columns normalized to 0/1.

4.3 Handling Invalid Values: Negative/unrealistic ages replaced with median.

4.4 Duplicate Removal: Removed duplicates based on appointment_id and full-row comparison.

4.5 Date Parsing: Standardized and extracted scheduled_date, appointment_date, and wait_days.

4.6 Target Column Fix: Clean binary no_show column created (1 = no-show, 0 = showed).

4.7 Text Correction: Cleaned gender and neighbourhood text formatting.

5. Output Files Generated

Final cleaned datasets:

• medical_appointments_cleaned.csv
• medical_appointments_cleaned.xlsx

Files contain corrected data types, standardized fields, and engineered features.

6. Summary of the Cleaned Dataset

The cleaned dataset is:

• Structurally consistent
• Free of duplicates
• Standardized in naming and formatting
• Enhanced with meaningful features
• Ready for analysis and modeling

7. Conclusion

Day 1 of the internship at Elevate Labs has been successfully completed. The dataset is now fully prepared for Day 2 tasks, including exploratory data analysis and visualization.

Prepared By: Busireddy Harshitha
Date: Day 1 Submission