Deliverable 1: Data Collection, Cleaning, and Exploration

Prafulla Man Singh Pradhan

Student ID: 005022638

MSCS634 - B01 Advanced Big Data and Data Mining

Dr. Satish Penmatsa

12th July 2025

Dataset selection and description

For this project, we selected a Healthcare Dataset obtained from Kaggle.

This dataset consists of over 500 records and includes 8–10 attributes such as age, gender, symptoms, and medical history.

We chose this dataset because:

- It provides real-world, patient-level data relevant to healthcare analytics.
- The mix of numerical and categorical variables makes it suitable for multiple data mining techniques, including classification, clustering, regression, and association rule mining.
- Predictive modeling on such data can potentially help discover patterns useful for improving healthcare decision-making and patient outcomes.

Loading and inspecting the dataset

The dataset was loaded into Python using the Pandas library:

```
import pandas as pd

df = pd.read_csv('healthcare_dataset.csv')
```

We then inspected its structure using:

- df.head() to view the first few rows
- df.info() to see data types and missing values

- df.describe() to get statistical summaries of numerical columns

This step helped us understand the composition of the dataset and identify initial data quality issues.

Data cleaning steps

To ensure the dataset is ready for analysis, we performed several cleaning steps:

1. Handling missing values
    - For numerical columns, missing values were imputed using the median. Median was chosen because it is less sensitive to outliers than the mean.
    - For categorical columns, missing values were filled using the mode (most frequent value).
2. Removing duplicates
    - Duplicate records were identified and removed using:
      df = df.drop_duplicates()
    - This helps prevent bias and overrepresentation of certain data points.
3. Standardizing data and correcting inconsistencies
    - Converted categorical text columns (e.g., Gender) to lowercase to ensure consistent formatting.
    - Verified and corrected data types (e.g., ensuring age is numerical, gender is categorical).

    - Cleaned inconsistent strings and typos in categorical data.

4. Identifying and addressing noisy data
    - Detected outliers in numerical variables like Age (e.g., extremely high values).
    - Flagged these outliers for further treatment or analysis in later modeling phases.

Exploratory Data Analysis (EDA)

To better understand the data, we conducted EDA using Seaborn and Matplotlib:

- Univariate analysis:
    - Histogram of Age: Showed that the age distribution is skewed, with a concentration in certain age ranges and presence of outliers.
    - Bar chart of Gender: Revealed a mild imbalance between male and female records.
- Bivariate analysis:
    - Boxplot of Age by Gender: Helped visualize the spread and potential differences in age distribution across genders.
    - Correlation heatmap: Explored relationships among numerical variables and identified moderate correlations that may influence predictive modeling.

Insights gained and next steps

From this analysis, we gained several insights:

- There is some imbalance in gender distribution, which we need to consider when building models to avoid bias.

- Certain numerical variables have moderate correlations, which can guide feature selection for regression or classification tasks.
- Outliers in Age may affect model accuracy, so we will decide whether to remove or transform these values later.

These insights will directly inform our feature engineering, choice of models, and how we handle imbalanced data or outliers in Deliverable 2 and beyond.