Data Preprocessing Report

A comprehensive overview of data cleaning and preprocessing techniques applied to the cardio dataset.

Author: Anshu Gondi

Date: 14 May 2025

Table of Contents

1. Introduction

This report provides an overview of the data preprocessing steps taken on the cardio dataset. Data preprocessing is a crucial step in data science that involves cleaning and transforming raw data into a format suitable for analysis.

Data Preprocessing Report

2. Importing Required Libraries

The following libraries were used:

- pandas
- numpy
- scipy
- matplotlib
- seaborn
- sklearn

3. Loading the Dataset

The dataset was loaded using pandas with the following command:

data = pd.read_csv("../data/raw/cardio_train.csv", sep=";")

4. Data Cleaning

Age Column

Issue: The age was recorded in days, leading to unusually high mean values.

Solution: Converted age from days to years by dividing the values by 365.

Technique Used: Unit transformation.

Blood Pressure (ap_hi, ap_lo)

Issue: Extreme and erroneous values (e.g., negative values for blood pressure).

Solution: Removed rows with systolic blood pressure (ap_hi) outside 90-200 mmHg and diastolic blood pressure (ap_lo) outside 60-120 mmHg.

Technique Used: Threshold-based filtering.

Cholesterol and Glucose (gluc)

Issue: Potential invalid entries; these columns should only contain categorical values (1, 2, 3).

Solution: Filtered rows where values in these columns were outside the valid range.

Technique Used: Value verification.

Binary Variables (smoke, alco, active)

Issue: Binary variables should only contain 0 and 1; potential for incorrect values.

Solution: Verified and removed rows where these variables had values other than 0 or 1.

Technique Used: Value validation.

5. Data Summary

The cleaned dataset contains the following summary statistics:

Count: 68418

Mean Age: 52.83

Mean Systolic Blood Pressure (ap_hi): 126.64

Mean Diastolic Blood Pressure (ap_lo): 81.32

Mean Cholesterol: 1.36

Mean Glucose: 1.23

Mean Smoke: 0.09

Mean Alcohol: 0.05

Mean Active: 0.80

Mean Cardio: 0.49

6. Visualizations

A boxplot for cleaned numerical features was created to visualize the distribution of age, systolic blood pressure (ap_hi), and diastolic blood pressure (ap_lo).

7. Conclusion

The preprocessing steps taken were essential in ensuring the quality of the dataset. By addressing issues such as outliers and invalid entries, the dataset is now ready for further analysis.

8. References

1. Jupyter Notebook Documentation
2. Pandas Documentation
3. Scikit-learn Documentation
4. Matplotlib Documentation
5. Seaborn Documentation