Analysis of Specialty Drugs in the Dataset

Objective

The goal of this analysis was to identify and validate specialty drugs from a given dataset containing Medicare and Medicaid spending data. We aimed to ensure the accuracy of specialty drug classification by leveraging authoritative sources.

1. Data Exploration and Loading

The dataset was provided as a CSV file containing brand names of drugs and their respective Medicare and Medicaid spending from 2018 to 2022.

We loaded the dataset into a Pandas DataFrame for further analysis.

2. Identifying Specialty Drugs

Specialty drugs are typically high-cost medications used to treat chronic, complex, or rare diseases.

We established criteria based on high spending thresholds and drug categorization from specialty pharmacy sources.

3. Cleaning and Normalizing Drug Names

To avoid inconsistencies in drug name matching, we:

Converted all drug names to lowercase.

Removed special characters like *, -, and spaces.

Applied normalization techniques to make name comparison more effective.

4. Implementing Fuzzy Matching

To improve accuracy, we used fuzzy string matching with the fuzz package to find close matches.

A threshold was set to match drug names in the dataset to known specialty drug lists with high confidence.

5. Validating Against Authoritative Sources

We cross-referenced the identified drugs with specialty drug lists from:

CVS Specialty Pharmacy https://www.cvsspecialty.com

Only drugs confirmed by these sources were classified as specialty drugs.

6. Generating the Final List of Specialty Drugs

After validation, we compiled the top 50 specialty drugs based on Medicare and Medicaid spending.

A final CSV file was created and made available for download.

7. Confirming Results

A total of zero non-specialty drugs were found in the provided dataset.

All drugs identified were validated as specialty medications through cross-referencing with authoritative sources.

Conclusion

This analysis successfully identified and validated specialty drugs using data normalization, fuzzy matching, and authoritative sources. The approach ensured accuracy in classification and provided an efficient way to analyze large-scale pharmaceutical spending data.