**Exploratory Data Analysis of Diabetes 130-US Hospital Dataset**

Prepared by

SHI HEPING

KOH WEI MING, LOUIS

WONG CHEE SENG

National University of Singapore

IT5006 Fundamentals of Data Analytics

Prof Prakash Chandra Sukhwal

Prof Ashish Deepak Dandekar

Updated on September 2025

**Dataset Introduction & Overview**

This report presents an exploratory data analysis (EDA) of the "Diabetes 130-US Hospitals" dataset, sourced from the UCI Machine Learning Repository. The dataset contains records of diabetic patient encounters in 130 US hospitals between 1999 and 2008. The primary objective of this analysis is to uncover patterns in patient demographics, hospitalization characteristics, medication usage, and factors associated with hospital readmission.

The dataset is composed of 101,766 patient encounters with 48 distinct features. These features provide a comprehensive view of each hospital stay, encompassing patient information and clinical outcomes. The feature breakdown is as follows:

* Numerical Features: 11 variables, including time\_in\_hospital and num\_medications.
* Categorical Features: 37 variables, including patient demographics (race, gender, age) and various medication prescriptions.

**Data Preprocessing and Cleaning**

To prepare the data for analysis, several preprocessing steps were performed to address quality issues and standardize formats.

1. Missing Values Analysis

A significant challenge in this dataset is the presence of missing values. The analysis identified nine columns with missing data, with five exhibiting substantial gaps:

* weight: 96.9% missing values
* max\_glu\_serum: 94.7% missing values
* A1Cresult: 83.3% missing values
* medical\_specialty: 49.1% missing values
* payer\_code: 39.6% missing values

The high percentage of missing data in columns like weight and key lab results (max\_glu\_serum, A1Cresult) suggests these tests were not routinely performed or recorded for all patients. In addition to standard nulls, several categorical columns contained ? as a placeholder for missing data. These were systematically identified and converted to NaN (Not a Number) for consistent handling.

1. Data Type Correction

To optimize memory usage and facilitate analysis, the primary, secondary, and additional diagnosis columns (diag\_1, diag\_2, diag\_3) were converted from object types to categorical data types.

1. Variable Reduction

The initial dataset included variables that offered no analytical value. Two columns, examide and citoglipton, were found to have only a single, constant value across all patient records. As constant variables provide no variance, they were identified for removal in subsequent modeling phases.

**Exploratory Data Analysis**

This section details the key findings from the exploration of patient demographics, clinical characteristics, and readmission patterns.

1. Target Variable: Readmission Status

The primary outcome of interest is the readmitted variable, which is distributed across three categories. The analysis reveals a significant class imbalance:

* Not Readmitted (NO): 53.9% of patients
* Readmitted after >30 days (>30): 34.9% of patients
* Readmitted within <30 days (<30): 11.2% of patients

The low representation of the <30 days category is a critical finding that must be addressed in predictive modeling, for instance, by using techniques like SMOTE (Synthetic Minority Over-sampling Technique) to prevent model bias.

1. Patient Demographics

* Age: The patient population is predominantly older, with the [70-80) age group being the most common, followed by [60-70) and [50-60).
* Gender: The distribution is relatively balanced, with a slight majority of females (53.8%) compared to males (46.2%). A negligible number of records (3) were marked as 'Unknown/Invalid'.
* Race: The dataset is dominated by patients identified as Caucasian (74.8%), followed by African American (18.9%). Other groups like Hispanic (2.0%) and Asian (0.6%) are significantly underrepresented relative to their proportions in the general U.S. population.

1. Hospitalization and Diagnostic Patterns

* Length of Stay: The average hospital stay is 4.4 days, with a standard deviation of 3.0 days. The distribution is right-skewed, indicating that most stays are short, but a small number of patients have significantly longer hospitalizations.
* Number of Medications & Diagnoses: On average, a patient received 16.0 medications and had 7.4 diagnoses recorded during their stay. This highlights the complexity of managing diabetic patients, who often present with multiple comorbidities.
* Correlation of Numerical Features: A correlation analysis was performed on the 11 numerical variables. The results showed no highly correlated pairs (defined as a correlation coefficient > 0.5). This suggests that numerical features like time\_in\_hospital, num\_medications, and num\_lab\_procedures are largely independent of each other.

1. Correlation Analysis

Among 11 numeric columns, a Pearson correlation matrix was computed. Using a threshold of |correlation| > 0.5 to define a significant correlation, no pairs of variables met this criterion. The highest observed correlations were weak to moderate at best, such as the relationship between the number of medications and the time spent in the hospital.

This finding has two important implications for subsequent modeling:

* Low Multicollinearity: The lack of strong correlations suggests that multicollinearity among numerical predictors is not a significant concern. This allows for the inclusion of all these features in linear-based models without risking model instability.
* Complex Relationships: The absence of simple linear trends indicates that the relationships between these variables and patient outcomes (like readmission) are likely complex and non-linear. This suggests that more sophisticated models, such as tree-based algorithms (e.g., Random Forest, Gradient Boosting) or neural networks, may be more effective at capturing the underlying patterns than traditional linear regression models. The predictive power in this dataset may be more heavily influenced by categorical features or interactions between variables rather than by individual numerical predictors alone.

1. Readmission Pattern Analysis

The relationship between patient characteristics and readmission status was explored:

* Readmission by Age: The rate of readmission within 30 days (<30) generally increases with age, peaking at 12.1% for the [80-90) age group. Older patients appear to be at a higher risk for early readmission.
* Readmission by Length of Stay: Patients readmitted within 30 days tend to have slightly longer initial hospital stays compared to those not readmitted or readmitted after 30 days. This may indicate that patients with more severe initial conditions are more likely to be readmitted quickly.
* Readmission by Gender: There is no significant difference in readmission rates between male and female patients.

1. Medication Analysis

* Common Medications: Beyond insulin, Metformin is the most commonly prescribed oral diabetes medication, followed by Glipizide and Glyburide.
* Insulin Usage: Insulin is a key medication, with its dosage being adjusted ('Up' or 'Down') in 21.2% of encounters. In 52.8% of cases, the dosage remained 'Steady', and for 26.0% of patients, 'No' insulin was administered.
* Change in Medication: A change in the overall medication regimen occurred in 46.6% of hospital stays, indicating active management of patient conditions during hospitalization.

1. Outlier Detection

Boxplots of numerical variables revealed the presence of a significant number of outliers, particularly in variables related to hospital utilization:

* number\_outpatient: 16.5% of records are outliers.
* number\_emergency: 11.2% of records are outliers.
* number\_inpatient: 6.9% of records are outliers.

These outliers represent patients with exceptionally high hospital visit frequencies and may warrant special attention or separate analysis.

**Summary and Key Findings**

This exploratory analysis provides several key insights into the diabetes patient dataset:

* Data Quality: The dataset has significant data quality issues, including high percentages of missing values for critical features like weight and A1Cresult, as well as a large number of outliers in hospital utilization metrics. These issues must be carefully managed during model development.
* Patient Profile: The typical patient in this dataset is an older adult (70-80 years), Caucasian, and slightly more likely to be female. They are often managing multiple health conditions, as indicated by the high average number of diagnoses and medications.
* Readmission Insights: Hospital readmission within 30 days is a key concern, though it represents the minority outcome (11.2%), creating a class imbalance problem for predictive modeling. Factors like older age and longer initial hospital stays are associated with a higher risk of early readmission.
* Feature Relationships: There are no strong linear correlations among the main numerical features, suggesting that more complex, non-linear models or feature interaction analysis may be required to uncover predictive relationships.