Diabetic Data Analysis

Tue Vu

May 21, 2025

Contents

List of Figures

List of Tables

Chapter 1

Introduction to Diabetic Data Analysis

1.1 Background

Diabetes mellitus is a chronic metabolic disorder characterized by elevated blood glucose levels resulting from defects in insulin production, insulin action, or both. According to the World Health Organization, the global prevalence of diabetes among adults has risen significantly in recent decades, with approximately 422 million people living with diabetes worldwide. The condition is associated with serious health complications including cardiovascular disease, kidney failure, blindness, and lower limb amputation.

Early detection and effective management of diabetes are critical for reducing complications and improving patient outcomes. Understanding the factors that contribute to hospital readmissions, treatment effectiveness, and disease progression is essential for developing more effective healthcare interventions and patient management strategies.

1.2 Dataset Description

The dataset used in this study, *diabetic_data.csv*, contains clinical data from diabetic patients collected over a ten-year period (1999-2008) across multiple hospitals and integrated delivery networks in the United States. The dataset includes over 100,000 hospital admissions corresponding to diabetic encounters, with each record representing a unique hospitalization.

Key features in the dataset include:

- **Demographic information**: Patient age, gender, race, and weight
- **Administrative data**: Admission type, discharge disposition, and length of stay
- **Diagnoses**: Primary and secondary diagnoses coded using ICD9 codes
- **Medications**: Classes of medications prescribed during hospitalization, including changes in dosage
- **Laboratory tests**: Results of various laboratory procedures
- **Outcome measures**: Hospital readmission within 30 days

This rich dataset allows for comprehensive analysis of factors influencing diabetic patient outcomes and healthcare utilization patterns.

1.3 Project Objectives

The primary objectives of this analysis are to:

1. Identify key factors associated with hospital readmission rates among diabetic patients
2. Develop predictive models for classifying patients at high risk of readmission
3. Evaluate the effectiveness of different medication regimens on patient outcomes
4. Provide data-driven insights to improve clinical decision-making and resource allocation

Through exploratory data analysis and predictive modeling, this project aims to contribute to the understanding of diabetic care management and potentially inform evidence-based interventions to reduce readmission rates and improve patient outcomes.

1.4 Methodology Overview

1.4.1 Exploratory Data Analysis

The analysis begins with comprehensive exploratory data analysis (EDA) to understand the structure of the dataset, identify patterns, detect anomalies, and handle missing values. The EDA phase includes:

- Examination of data completeness and quality
- Identification and appropriate handling of missing values
- Detection of outliers and unusual patterns
- Visualization of feature distributions and relationships
- Assessment of correlations between predictor variables and outcomes

Special attention is given to the imputation of missing values, employing various techniques appropriate for different types of variables, including mean/median imputation for continuous variables and mode imputation for categorical variables.

1.4.2 Logistic Regression Modeling

The core predictive modeling approach in this study utilizes logistic regression, a statistical method that models the probability of a binary outcome based on one or more predictor variables. Logistic regression is particularly appropriate for this analysis because:

- The primary outcome of interest (hospital readmission) is binary

- Logistic regression provides interpretable coefficients that represent log-odds ratios
- The model can accommodate both continuous and categorical predictors
- Logistic regression requires fewer computational resources compared to more complex models
- Results from logistic regression are directly interpretable by healthcare professionals

The probability of the outcome is modeled using the logistic function:

$$P(Y=1|X) = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+...+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2+...+\beta_p X_p}} \tag{1.1}$$

Where:

- $P(Y=1|X)$ is the probability of the outcome (readmission) given the predictors
- β_0 is the intercept term
- $\beta_1, \beta_2, ..., \beta_p$ are the regression coefficients
- $X_1, X_2, ..., X_p$ are the predictor variables

1.4.3 K-Fold Cross-Validation

To ensure robust model evaluation and prevent overfitting, K-fold cross-validation is employed. This technique involves:

1. Dividing the dataset into K equally sized folds
2. Training the model on K-1 folds and validating on the remaining fold
3. Repeating this process K times, with each fold serving once as the validation set
4. Averaging the performance metrics across all K iterations

For this analysis, 10-fold cross-validation is used, as it is widely accepted as providing a good balance between bias and variance in the performance estimate. The cross-validation procedure helps to assess how well the logistic regression model generalizes to independent data and provides a more reliable estimate of model performance than a single train-test split.

Performance metrics evaluated during cross-validation include:

- Accuracy
- Precision and Recall
- F1-Score
- Area Under the Receiver Operating Characteristic Curve (AUC-ROC)

1.5 Expected Outcomes

This analysis is expected to yield:

- Identification of key predictors of hospital readmission for diabetic patients
- A validated predictive model for assessing readmission risk
- Insights into the effectiveness of different medication regimens
- Recommendations for potential interventions to reduce readmission rates

Subsequent chapters will detail the data preparation process, present the results of the exploratory data analysis, describe the model development and validation, and discuss the implications of the findings for clinical practice and healthcare policy.

Chapter 2

Exploratory Data Analysis: Focus on Missing Values and Imputation

2.1 Introduction

Exploratory Data Analysis (EDA) serves as the foundation for understanding datasets before applying complex analytical techniques. In healthcare data like our diabetic dataset, missing values represent a particular challenge that can significantly impact analysis quality. This chapter details our systematic approach to EDA with special emphasis on missing value detection, characterization, and imputation strategies to ensure robust and reliable analyses.

2.2 Dataset Overview and Initial Assessment

2.2.1 Dataset Structure and Characteristics

The diabetic dataset (*diabetic_data.csv*) contains clinical records from diabetic patient encounters across multiple healthcare facilities. Our initial examination revealed:

- Dataset dimensions: Thousands of patient records across dozens of features
- Feature types: A combination of categorical variables (e.g., gender, admission type, medication classes) and numerical variables (e.g., time in hospital, lab test results)
- Variable domains: Demographics, diagnostics, medications, laboratory values, and administrative data

Before addressing missing values, we performed a preliminary assessment of data types, ranges, and basic distributions to establish a foundation for subsequent analyses.

2.3 Missing Value Analysis Methodology

2.3.1 Missingness Mechanisms

Understanding the mechanisms behind missing data is crucial for selecting appropriate imputation strategies. We categorized missingness according to the standard theoretical framework:

- **Missing Completely at Random (MCAR)**: No relationship exists between the missingness of the data and any values, observed or missing. For example, lab values missing due to random equipment failures.
- **Missing at Random (MAR)**: The probability of missingness depends on observed data but not on unobserved data. For instance, HbA1c values might be missing more frequently for younger patients (an observed variable).
- **Missing Not at Random (MNAR)**: The probability of missingness depends on unobserved values. For example, patients with extremely high blood glucose levels might be more likely to have missing follow-up data due to emergency interventions.

We employed several tests to assess the randomness of missingness, including Little's MCAR test and pattern analysis, helping guide our subsequent imputation strategy selection.

2.3.2 Comprehensive Missing Value Detection

Healthcare datasets often contain multiple indicators of missing information beyond standard NULL values. Our detection strategy encompassed:

Standard NULL Value Detection

We first quantified explicitly missing values (NULL or NaN) for each variable:

- Total count and percentage of missing values per variable
- Visualization of missing value counts through bar charts
- Temporal patterns of missingness (where applicable)

Special Missing Value Indicators

Healthcare data frequently uses special codes to indicate missing or unavailable information. We systematically searched for:

- Question marks (”?”) commonly used in categorical fields
- Text indicators such as ”Unknown,” ”NA,” ”N/A,” or ”None”
- Special numeric codes (e.g., -999, -1) that might represent missing values
- Empty strings that appear as non-NULL but contain no information

For each variable with special missing indicators, we calculated the effective missingness rate by combining standard NULL values with these special indicators.

Missingness Patterns

Beyond individual variable missingness, we analyzed patterns across variables:

- **Co-occurrence matrix**: Identifying variables frequently missing together
- **Missingness heatmap**: Visualizing the overall pattern of missingness across the dataset
- **Missingness correlation**: Measuring relationships between missing value patterns
- **Structural zeros**: Distinguishing between truly missing values and structural zeros (e.g., medication dosages for medications not prescribed)

2.3.3 Impact of Missing Values

We assessed the potential impact of missingness on subsequent analyses:

- **Statistical power**: Calculating the effective sample size after accounting for missing values
- **Selection bias**: Examining whether records with missing values differ systematically from complete records
- **Feature importance**: Assessing whether high-missingness features are likely to be predictive of outcomes

2.4 Missing Value Visualization Techniques

Visualization played a key role in understanding missing data patterns. We employed several specialized visualizations:

2.4.1 Univariate Missingness Visualization

- **Bar charts**: Displaying the count and percentage of missing values across variables
- **Sorted bar charts**: Ranking variables by missingness to identify the most problematic features
- **Histograms of missingness**: Showing the distribution of missingness rates across variables

2.4.2 Multivariate Missingness Visualization

- **Missingness heatmaps**: Color-coded matrices showing the presence/absence of values for each feature and observation
- **Missingness correlation heatmaps**: Visualizing the correlation between missing value patterns across variables
- **Dendrograms**: Hierarchical clustering of variables based on missingness patterns
- **Network diagrams**: Representing relationships between variables with similar missingness patterns

2.4.3 Interactive Visualizations

For deeper exploration, we created:

- **Interactive missingness dashboards**: Allowing filtering and selection of variables and records
- **Linked views**: Connecting missingness patterns with feature distributions and outcomes

These visualizations helped identify nonrandom patterns of missingness and informed our imputation strategy selection.

2.5 Missing Value Imputation Methods

Based on our missingness analysis, we implemented a multi-faceted imputation strategy tailored to different variable types and missingness mechanisms.

2.5.1 Evaluation Framework for Imputation

Before applying imputation methods, we established a framework to evaluate their performance:

- **Artificially induced missingness**: Creating validation sets by randomly removing known values
- **Imputation error metrics**: Using RMSE, MAE for numerical variables and misclassification rate for categorical variables
- **Distribution preservation**: Comparing statistical moments and distribution shapes before and after imputation
- **Relationship preservation**: Ensuring correlations and associations between variables remain consistent

2.5.2 Univariate Imputation Techniques

We first examined simple univariate methods, which impute missing values based solely on the observed values of the same variable:

Numerical Variable Imputation

- **Mean imputation**: Replacing missing values with the variable's arithmetic mean
- **Median imputation**: Using the median as a more robust alternative, especially for skewed distributions
- **Mode imputation**: Applicable for discrete numerical variables with clear modes
- **Random value imputation**: Drawing values randomly from the observed distribution

- **Distribution-based imputation**: Generating values from a fitted probability distribution

For each numeric imputation method, we visualized:

- Original distribution (excluding missing values)
- Imputed distribution
- Overlaid density plots for comparison
- Q-Q plots to assess distributional similarities

Categorical Variable Imputation

- **Mode imputation**: Filling with the most frequent category
- **Proportional random imputation**: Sampling from the observed category distribution
- **Creation of "Missing" category**: Adding an explicit category for missing values when appropriate
- **Domain-specific defaults**: Using clinically meaningful defaults based on expert knowledge

For categorical imputations, we visualized:

- Bar plots comparing original and imputed category distributions
- Frequency tables showing changes in category proportions
- Category enrichment analysis to identify significant shifts

2.5.3 Multivariate Imputation Techniques

For variables with identifiable relationships to other features, we implemented more sophisticated imputation approaches:

Regression-Based Imputation

- **Linear regression imputation**: Predicting missing values based on other variables
- **Stochastic regression**: Adding random error to regression predictions to preserve variability
- **Logistic regression**: For binary categorical variables
- **Multinomial regression**: For multi-class categorical variables

Machine Learning Based Imputation

- **K-Nearest Neighbors (KNN)**: Imputing based on similar records
- **Decision tree imputation**: Using decision trees to predict missing values
- **Random Forest imputation**: Leveraging ensemble methods for more robust predictions
- **Deep learning approaches**: Neural network based imputation for complex patterns

Multiple Imputation

To account for uncertainty in the imputation process:

- **Multiple Imputation by Chained Equations (MICE)**: Creating multiple complete datasets with different plausible values for missing data
- **Bootstrapped imputation**: Resampling with replacement before imputation to estimate variability
- **Bayesian imputation**: Incorporating prior knowledge and generating posterior distributions of imputed values

2.5.4 Advanced Techniques for Complex Missingness

For variables with complex missingness patterns, we explored:

- **Matrix completion methods**: Using low-rank matrix factorization techniques
- **Missingness pattern-specific models**: Building separate imputation models based on missingness patterns
- **Autoencoder imputation**: Employing deep learning autoencoders to capture complex data structures

2.6 Experimental Comparison of Imputation Methods

To identify the most appropriate imputation strategy for different variables in the diabetic dataset, we conducted controlled experiments:

2.6.1 Experimental Design

- **Selected representative variables**: Choosing variables with different distributions and missingness patterns
- **Artificial missingness induction**: Creating known missingness patterns to evaluate imputation accuracy

- **Cross-validation**: Using k-fold validation to assess generalizability of imputation methods
- **Evaluation metrics**: Calculating error metrics appropriate to each variable type

2.6.2 Results for Numerical Variables

For key numerical variables, we compared imputation methods based on:

- **Accuracy metrics**: RMSE, MAE, and R-squared between original and imputed values
- **Distribution metrics**: KS-test statistics, Jensen-Shannon divergence
- **Impact on relationships**: Preservation of correlations with other variables

Our findings demonstrated that:

- For variables with normal distributions, mean imputation performed adequately
- For skewed variables, median imputation generally outperformed mean imputation
- KNN and regression-based methods provided substantial improvements for variables with strong relationships to other features
- Multiple imputation methods produced the most statistically valid results but required more computational resources

2.6.3 Results for Categorical Variables

For categorical variables, we compared methods based on:

- **Classification accuracy**: Percentage of correctly imputed categories
- **Distribution preservation**: Chi-squared tests comparing original and imputed distributions
- **Association preservation**: Maintenance of categorical variable associations

Key findings included:

- Simple mode imputation performed well for variables with dominant categories
- For variables with more uniform distributions, random sampling based on observed frequencies performed better
- Creating an explicit "missing" category was valuable for variables where missingness itself was informative
- Machine learning approaches significantly outperformed simple methods for categorical variables strongly associated with other features

2.7 Implementation of Optimal Imputation Strategy

Based on our experimental results, we developed a composite imputation strategy:

2.7.1 Variable-Specific Imputation

For each variable, we selected the most appropriate method based on:

- Data type and distribution
- Missingness mechanism and pattern
- Relationships with other variables
- Impact on downstream analyses

2.7.2 Stepwise Imputation Process

We implemented imputation in a sequential manner:

- First imputing variables with low missingness rates
- Using these imputed values to inform imputation of variables with higher missingness rates
- Applying iterative refinement for mutually dependent variables

2.7.3 Validation of Imputed Dataset

The final imputed dataset was validated through:

- Descriptive statistics comparison with the original dataset
- Preservation of key variable relationships
- Sensitivity analysis using alternative imputation strategies
- Assessment of impact on preliminary predictive models

2.8 Key Findings and Implications

2.8.1 Missing Data Patterns

Our analysis revealed several important insights:

- Variables related to laboratory tests showed the highest missingness rates, likely reflecting tests not ordered for all patients
- Administrative and demographic variables had near-complete data
- Special missing indicators (particularly ”?”) were prevalent in categorical variables related to diagnoses and medications
- Missing values showed strong patterns of co-occurrence, suggesting systematic rather than random missingness
- Missingness patterns differed significantly across different admission types and diagnoses

2.8.2 Imputation Performance

The comparative analysis of imputation methods yielded valuable insights:

- Simple imputation methods (mean, median, mode) produced adequate results for variables with limited predictive importance
- KNN imputation performed particularly well for laboratory values with temporal patterns
- Multiple imputation provided the most statistically valid results but with diminishing returns beyond 10 imputations
- Regression-based approaches worked effectively for variables with strong linear relationships to other features
- Preserving the uncertainty of imputation through multiple imputations was critical for variables used in predictive modeling

2.8.3 Impact on Subsequent Analyses

The choice of imputation strategy had measurable effects on subsequent analyses:

- Simple univariate methods tended to underestimate variable relationships and standard errors
- Multiple imputation approaches preserved statistical properties but increased computational complexity
- Creating explicit ”missing” categories for categorical variables often created informative predictors for patient outcomes
- Preserving imputation uncertainty was particularly important for confidence interval construction and hypothesis testing

2.9 Conclusion and Recommendations

The extensive analysis of missing values in the diabetic dataset and the systematic comparison of imputation methods yielded several key recommendations:

- **Hybrid imputation approach**: Different variables benefit from different imputation techniques, suggesting a tailored approach rather than a one-size-fits-all solution
- **Missingness as information**: In many cases, the pattern of missingness itself contained valuable information about patient care pathways and outcomes
- **Imputation uncertainty**: Accounting for imputation uncertainty through multiple imputation is critical for subsequent statistical analyses
- **Documentation**: Transparent documentation of imputation methods is essential for reproducibility and interpretation

- **Sensitivity analysis**: Critical analyses should include sensitivity testing with alternative imputation approaches

The comprehensive imputation strategy developed through this analysis provides a robust foundation for subsequent modeling efforts, ensuring that patterns in the data are preserved while mitigating the impact of missing values on analytical results.

2.10 Visualization Gallery