Data Encoding Rule

for *diabetes* project

by Shuiliang (Leon) Wu

July 12, 2024

1 Dataset Description

The dataset for diabetes project shall have 9 columns related to patients with diabetes. The **readmitted** variable, which is the target for predictions, indicates whether patients were readmitted to the hospital after their initial visit. This dataset is used for analyzing factors that influence hospital readmission rates in diabetic patients. Below is the details of each column. **The order of feature name in *site_name_header.csv* follows the order of the list below.**

1. **readmitted**: Categorical target variable indicating if the patient was readmitted to the hospital (*NO*, *>30*, *<30*).
2. **race**: Categorical data representing the race of the patient, with 6 unique categories which are *?*, *AfricanAmerican*, *Asian*, *Caucasian*, *Hispanic*, *Other*. There may be some missing values with race of *?* in this column.
3. **gender**: Categorical data with two categories (*Male*, *Female*).
4. **age**: Categorical data divided into age groups (10 unique groups).
5. **time_in_hospital.x**: Numeric data indicating the duration of the hospital stay.
6. **HbA1c**: Categorical data describing the hemoglobin A1c test results, with three categories (*>7*, *>8*, *Norm*).
7. **diabetesMed**: Categorical data indicating whether the patient was on diabetes medication (*Yes* or *No*).
8. **admission_source_id**: Categorical data describing the source of admission with 3 unique categories which are *Emergency room*, *Other*, *Physician/clinic referral*.
9. **number_outpatient_emerg_inpatient**: Numeric data representing the total number of outpatient, emergency, and inpatient visits in the year before the encounter.

2 Feature Engineering

Feature engineering is a critical step in preparing the dataset for modeling. This section outlines the strategies employed to handle missing data and the treatment of categorical data.

2.1 Handling Missing Data

The *race* column may contain instances marked as *?*. The following approach is used to handle these missing values:

Label Encoding with Separate Category Instead of imputing these values, a separate category was created for unknown race values during the label encoding process. This allowed the model to recognize and handle these entries explicitly.

2.2 Handling Categorical Data

Categorical data in the dataset were handled using various encoding techniques to convert them into numerical values suitable for modeling. The following approaches were employed:

Label Encoding This technique was used to convert categorical text data into numerical values. The mappings created during label encoding are as follows:

- *race*: {*?*: 0, *AfricanAmerican*: 1, *Asian*: 2, *Caucasian*: 3, *Hispanic*: 4, *Other*: 5}
- *gender*: {*Female*: 0, *Male*: 1}
- *diabetesMed*: {*No*: 0, *Yes*: 1}
- *admission_source_id*: {*Emergency room*: 0, *Other*: 1, *Physician/clinic referral*: 2}

Custom Mappings Some features required custom mappings to convert their categories into numerical values:

- *age*: {*[0-10)*: 0, *[10-20)*: 1, *[20-30)*: 2, *[30-40)*: 3, *[40-50)*: 4, *[50-60)*: 5, *[60-70)*: 6, *[70-80)*: 7, *[80-90)*: 8, *[90-100)*: 9}
- *HbA1c*: {*Norm*: 0, *>7*: 1, *>8*: 2}

Binary Encoding The *readmitted* column, which indicates whether a patient was readmitted, was encoded as a binary feature:

- *readmitted*: 0 for *NO* and 1 for *YES*