AI/ML Pipeline for Predicting Diabetic Readmissions from CMS OASIS dataset

Mirna Elizondo
Department of Computer Science
Texas State University
San Marcos, Texas
m_e172@txstate.edu

Dr. Denise Gobert
Department of Physical Therapy
Texas State University
Round Rock, Texas
dg46@txstate.edu

Dr. Jelena Tešić
Department of Computer Science
Texas State University
San Marcos, Texas
jtesic@txstate.edu

***Abstract*—This study examines how different feature encoding methods and data balancing techniques affect the performance of various machine learning approaches for varying feature importance and predicting hospital readmission from the CMS OASIS dataset. For the first task of attribute importance, we compare two feature encoding methods, Binary and Frequency encoding, and their interaction with balanced and unbalanced datasets using a dozen attribute importance methods and feature importance from four gradient boosting methods. RidgeClassifier selected the most features, followed by LogisticRegression, while methods like GradientBoostingClassifier and XGBClassifier selected the fewest features. For the second task of predicting rehospitalization, Binary Encoding consistently outperforms Frequency Encoding, leading to fewer false rehospitalization predictions (FP) and rehospitalization omissions (FN) when combined with data balancing. While Frequency Encoding reduced rehospitalization omissions (FN) with balanced data, it simultaneously increased false rehospitalization predictions (FP). Data balancing improved correct rehospitalization prediction (TP) across all models and encoding methods, highlighting its importance in improving classification accuracy. The correct false rehospitalization predictions (TN) remained relatively stable. Overall, the findings suggest that Binary Encoding with data balancing is the most effective approach for minimizing classification errors, offering valuable insights for feature engineering and selection in healthcare and similar fields.**

***Index Terms*—gradient boosting, predictive modeling, high-cardinality, diabetes**

I. Introduction

Hospital readmission is a significant public health issue, impacting both care quality and healthcare costs [1]. While some readmissions, such as those related to chronic conditions like cancer, are expected, they remain a concern. Readmission, defined as an inpatient stay within 30 days of discharge, varies across demographics. A systematic review found higher 30-day readmission rates among non-Hispanic Black (19.4%), Hispanic (16.0%), and American Indian/Alaska Native (15.9%) beneficiaries compared to non-Hispanic White beneficiaries (13.8%). Additionally, individuals with potentially disabling conditions had higher rates (18.3%) compared to those without (11.9%) [2]. The Outcomes and Assessment Information Set (OASIS), developed by CMS, is a comprehensive tool for evaluating home healthcare quality in the U.S. and is mandatory for all Medicare-certified agencies. [3]. CMS OASIS dataset includes patient demographics, functional and psychosocial status, health services utilization, and diagnoses. In this study, we analyze the OASIS dataset, which contains 15,754,330 rows and 271 columns, to assess the prevalence of diabetes among 4,851,692 beneficiaries, of whom 333,943 have been diagnosed with diabetes. The dataset features 306 diabetes-related ICD-10 codes across six diagnosis columns, enabling a detailed analysis of health outcomes, risk factors, and predictive model development for Medicare beneficiaries. The dataset's comprehensive nature supports the reliability of our research methodology and findings.

II. Methodology

A. Data Collection

The Outcome and Assessment Information Set (OASIS) by the Centers for Medicare and Medicaid Services (CMS) is a standardized set of data elements developed over two decades to enhance home healthcare quality measurement and improvement. The CMS OASIS dataset provides a comprehensive view of home healthcare services and outcomes. By analyzing this data, we can assess the quality of care provided by home health agencies (HHAs) and identify areas for improvement. The CMS OASIS dataset provides a comprehensive view of home healthcare services and outcomes. By analyzing this data, we can assess the quality of care provided by home health agencies (HHAs) and identify areas for improvement.

B. Preprocessing

The preprocessing phase of this study focuses on preparing the dataset for analysis using Apache Spark, specifically targeting various data quality issues common in noisy tabular data. The OASIS dataset includes records from all home health agencies certified to receive Medicare and Medicaid payments, containing 306 distinct condition codes. However, our analysis concentrates solely on diabetic conditions. The relevant ICD codes for diabetes are those beginning with E08, E09, E10, E11, and E13 [4]. Among the beneficiaries, 333,943 have been identified with diabetes, while 4,827,847 have another condition. Examples of specific codes include E11.61 for Type 2 diabetes mellitus with diabetic arthropathy, E08.3593 for diabetes mellitus due to an underlying condition, and E13.6 for other specified diabetes mellitus. We identify any missing information in the dataset. Columns with more than 70% of

missing values are dropped to enhance the overall quality of the data.

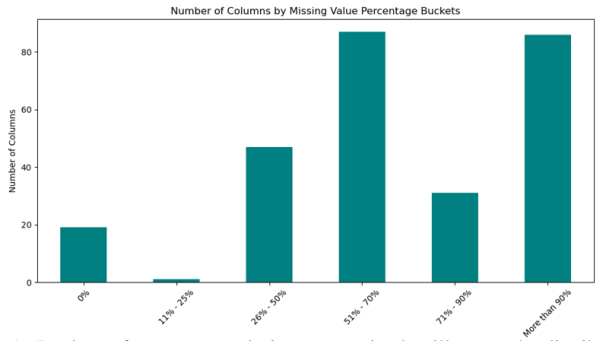

Fig. 1. Buckets of percentage missing, categorized to illustrate the distribution of *absence* across different ranges; each bucket represents a distinct range of missing values.

The CMS OASIS data revealed that when a patient had one missing test result, other tests, too, have a higher rate of *absence*. This monotone pattern of *absence* was first captured in [5], and our analysis shows that CMS OASIS data captures the same behavior for the group of patients who do not seek healthcare regularly. Figure 1 shows the distribution of *absence* of values across different ranges.

C. Clustering ICD Codes

- **Hamming Similarity**
- **LDA**
- **KMeans Clustering**

D. Encoding High-Cardinality Columns

• **Binary Encoding** is commonly used to convert categorical variables into a binary format. That has proven to be more efficient for machine learning algorithms. Thus, each category is assigned a unique integer and then converted into its binary representation. For example, if there are eight categories, they can be represented as three binary digits, e.g., category 0 = 000, category 1 = 001, category 2 = 010, etc.

• **Frequency Encoding** replaces each category with its frequency of occurrence in the dataset. For instance, if a certain category appears 100 times in a dataset of 1,000 rows, it would be replaced by the value 0.1 (100/1000). This technique preserves some information about the distribution of the values in the dataset.

• **One-Hot Encoding** is commonly used for categorical variables, it creates new binary columns, indicating the presence of each possible category in a column but is impractical for high-cardinality features as it creates a binary column for each category, significantly increasing dimensionality and sparsity in the dataset representation.

In this section, we propose to compare frequency encoding and binary encoding for the CMS OASIS dataset. The CMS OASIS dataset transformed with the frequency encoding consists of 17,221 rows and 354 columns, whereas binary encoding expands the dataset to 389,688 rows and 547 columns. For numeric columns, frequency encoding includes 345 columns (149 floats and 196 integers), while binary encoding includes 538 columns (7 floats and 531 integers). Both methods have nine object columns. The frequency encoding supports a broad range of values based on frequency counts, making it easier to handle due to its numerical nature. In contrast, binary encoding is limited to binary values (0 or 1) and may display sparse relationships. Frequency encoding results in a compact dataset that retains numerical meaning, offering lower dimensionality and more meaningful correlations for numeric features. Binary encoding, on the other hand, preserves categorical relationships but results in higher dimensionality due to the expansion of features, with correlations often being sparse.

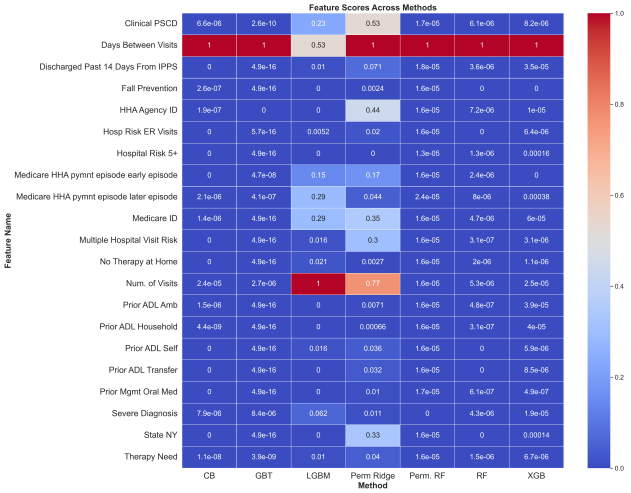

Fig. 2. Normalized feature importance scores across a dozen methods.

Task 1: Attribute Selection The feature importance scores in Figure 4 were normalized to facilitate comparison across different machine learning models. For each feature, the scores assigned by all models were summed to obtain a total score. Then, the normalized score for a specific feature and model was calculated by dividing the model's score for that feature by the total score. This normalization ensures that the scores are relative to the overall importance of the feature, allowing for a fair comparison between models. RidgeClassifier selected the most features, with a total of 322, followed by LogisticRegression, which selected 172 features. In contrast, CatBoostClassifier, LGBMClassifier, and RandomForestClassifier selected 3, 5, and 6 features, respectively. The models GradientBoostingClassifier and XGBClassifier each selected only one feature.

Task 2: Rehospitalization Modeling The Figure 4 presents bar plots for the confusion matrix metrics—correct rehospitalization prediction (TP),correct false rehospitalization prediction (TN), false rehospitalization prediction (FP) and rehospitalization omission (FN) of the model —highlighting variations by encoding type (Binary vs. Frequency) and data balancing (balanced vs. unbalanced). The total count of TP, TN, FP, FN represents the total number of instances in the dataset, as the confusion matrix sums to the total dataset size. *TN*: TN counts remain high for both encoding types, with slightly higher values in unbalanced data. Data balancing

TABLE I
TOP 7 FEATURES SELECTED BY MORE THAN NINE ATTRIBUTE SELECTION METHODS, AND THEIR IMPORTANCE.

Feature	Relevance	High Value	Low Value
DAYS_BETWEEN _VISITS	Highlights consistency in care	Frequent visits indicate active management	Long intervals suggest gaps in care
Number_of_visits	Tracks total healthcare visits	High visit count indicates intensive management	Few visits may reflect missed interventions
M1033_HOSP_RISK _MLTPL_HOSPZTN	High risk for multiple hospitalizations	High risk indicates challenges in disease management	Low risk suggests better disease control
M0110ETM_02 - Medicare HHA payment Episode	Reflects timing of home health care	Early episodes may indicate acute care needs	Later episodes may reflect chronic disease management
M0010 - Agency Medicare Number Frequency	Measures care frequency by Medicare agencies	High frequency suggests coordinated or specialized services	Low frequency may indicate limited access to care
M1000_DC_IPPS _14_DA - Discharged Past 14 Days	Recent discharge indicates a recent health event	Recent discharge requires prompt follow-up to ensure recovery	No recent discharge implies stable condition
M1900_PRIOR _ADLIADL_SELF - Prior ADL/IADL	Assesses prior self-care and independence	High independence indicates better self-care capability	Low independence suggests the need for additional support

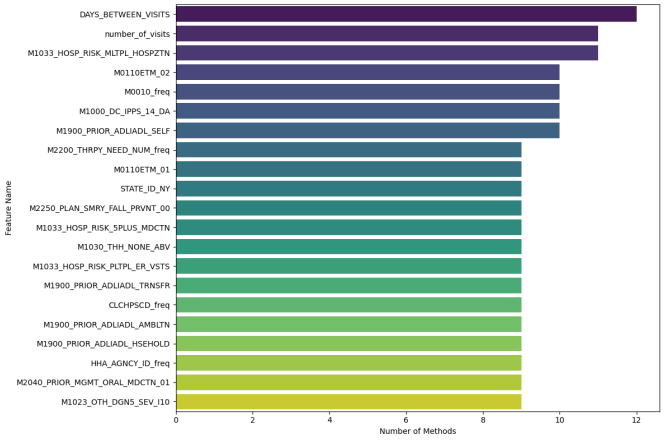

Fig. 3. Attributes selected by more than nine selection methods.

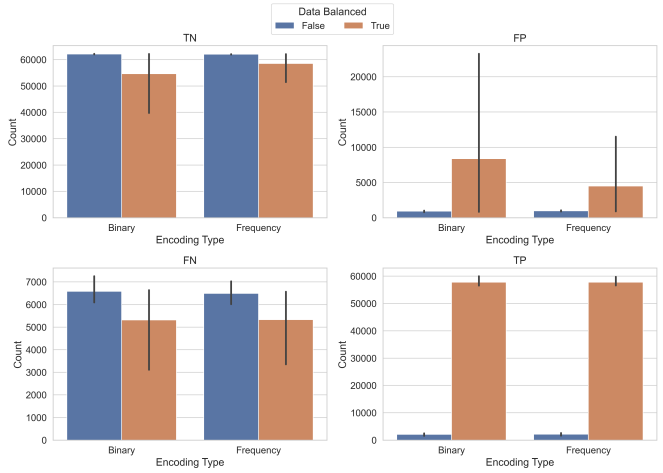

Fig. 4. Confusion matrix metrics by encoding type and data balancing.

has a minor impact on reducing TN counts, suggesting the limited influence of encoding type in unbalanced settings. *FP*: Frequency encoding with balanced data shows a significant increase in FP counts compared to unbalanced data. In contrast, Binary encoding maintains low rehospitalization miss values regardless of balancing. The finding suggests that frequency encoding may introduce more wrong rehospitalization predictions when the data is balanced. *FN*: is the count of the time model missed to predict the actual rehospitalization. The omission counts are highest with unbalanced frequency encoding, but balancing the data reduces them. Binary encoding consistently shows lower FN counts, with balancing further improving results, indicating its effectiveness in minimizing FNs. *TP*: Data balancing greatly increases correct rehospitalization prediction counts for both encoding types, while the encoding type itself has minimal impact when data is balanced.

IV. CONCLUSION AND FUTURE WORK

This study developed a machine learning model to predict high-risk patients using a comprehensive healthcare dataset. Key predictors, such as visit frequency, discharge history, and functional status, were identified. The models performed well in distinguishing between high-risk and low-risk patients. Binary Encoding outperformed other methods by maintaining lower false predictions and false omissions of rehospitalization in the model—data balancing further enhanced performance, particularly in increasing correct prediction. Future research will focus on incorporating temporal data, expanding features, improving model interpretability, and deploying the model in clinical settings to enhance patient care.

Acknowledgments: This work was sponsored in part by the NSF Expand AI Grant No. 2334268 and the Texas State Center for Analytics and Data Science (TXST CADS).

REFERENCES

[1] A. J. Weiss and H. J. Jiang, "Overview of clinical conditions with frequent and costly hospital readmissions by payer, 2018," Jul 2021. [Online]. Available: https://hcup-us.ahrq.gov/reports/statbriefs/sb278-Conditions-Frequent-Readmissions-By-Payer-2018.jsp

[2] C. for Medicare & Medicaid Services, Sep 2020. [Online]. Available: https://www.cms.gov/files/document/impact-readmissions-reduction-initiatives-report.pdf

[3] ——, "Oasis data sets," https://www.cms.gov/medicare/quality/home-health/oasis-data-sets, May 2 2024.

[4] J. Dugan and J. Shubrook, "International classification of diseases, 10th revision, coding for diabetes," *Clinical Diabetes: A Publication of the American Diabetes Association*, vol. 35, no. 4, pp. 232–238, 2017.

[5] J. Li, X. S. Yan, D. Chaudhary, V. Avula, S. Mudiganti, H. Husby, S. Shahjouei, A. Afshar, W. F. Stewart, M. Yeasin, and et al., "Imputation of missing values for electronic health record laboratory data," *npj Digital Medicine*, vol. 4, no. 1, Oct 2021.