Research Paper

Influence of HbA1c on Hospital Readmission Rate: Analysis of 71,515 unique patient records

Surya Chandra Raju Kurapati
Masters in Data Analytics - School of Computing,
National College of Ireland, Dublin, Ireland,
x23396920@student.ncirl.ie

Abstract - Hospital readmission rates are a key indicator of healthcare quality and efficiency. This study investigates the influence of HbA1c on hospital readmission rates using logistic regression. Data from over 100k patient encounters are analysed to identify significant predictors of readmission. Results suggest that HbA1c testing correlates with a reduction in readmission likelihood, particularly in patients with diabetes as a primary diagnosis. This insight provides valuable direction for optimizing inpatient diabetes care. Following a structured approach that includes data preprocessing, exploratory data analysis (EDA), iterative model development, and diagnostic evaluation, the conclusive model attains an accuracy of 61% and an AUC of approximately 0.60.

Keywords - Diabetes Management, Generalized Linear Model, Data Preprocessing, Exploratory Data Analysis, Feature Selection, Extraction and Scaling, Multivariate Analysis, AUC Score, Precision-Recall, Outlier Detection, Variance Inflation Factor (VIF), Healthcare Analytics.

1. Introduction

Hospital readmissions contribute to increased healthcare costs and indicate suboptimal patient management. HbA1c (Haemoglobin A1c) is a crucial marker for long-term blood sugar control, yet its usage in inpatient settings is inconsistent. This study aims to determine whether HbA1c testing influences hospital readmission rates by analysing patient data using logistic regression.

By employing logistic regression, a well-established statistical model for binary classification problems, this study aims to determine the relationship between various clinical and demographic factors and the likelihood of hospital readmission. Logistic regression was chosen due to its interpretability, efficiency, and ability to quantify the impact of multiple predictors, making it an ideal choice for analysing hospital data.

To facilitate a structured analysis, the target variable—hospital readmission—was transformed into a binary classification: patients readmitted within or after 30 days of discharge were assigned a value of “Readmitted” (1), while those who were not readmitted were labelled as “Not Readmitted” (0).

The study follows a structured methodology, including data preprocessing, exploratory data analysis (EDA), feature selection, feature extraction, feature scaling, model training, and performance evaluation.

Data preprocessing involved handling missing values, removal of duplicate and unnecessary valued records such as “?” in diagnosis codes attribute. EDA was conducted to understand multi-variable distributions, identify trends influencing readmission rates, and perform outlier detection. Later, encoding categorical and normalizing numerical features was performed to train the model. The logistic regression model was iteratively refined through feature selection and diagnostic testing to improve predictive performance. Key evaluation metrics, such as accuracy, precision, recall, F1-score, and ROC-AUC, were used to assess model effectiveness.

The findings emphasize the potential of incorporating HbA1c testing in standard hospital procedures to improve patient outcomes and reduce avoidable readmissions.

2. Understanding dataset

The dataset, “hospitaldata.csv”, consists of 101,763 patient encounters related to diabetes treatment, sourced from Strack et al. It contains 47 columns covering demographic information, clinical metrics, prior healthcare utilization, and medication history (Refer to Table 1).

3. Methodology

a) Data Preprocessing

Before applying logistic algorithm, the dataset underwent preprocessing to ensure data integrity and suitability for analysis. An initial inspection using `info()` confirmed that while there were no explicit null values, missing data was represented by `"?"` and “Unknown” in several critical columns.

Notably, diag_2 contained 358 records with `"?"`, and diag_3 had 1,423 such records, whereas diag_1 had no missing values. These columns capture primary and secondary diagnoses, which are crucial predictors of hospital readmission. Additionally, other columns such as weight (96.9%) and medical specialty (49.1%) had a substantial amount of missing values, making them impractical for imputation. Duplicate patient encounters were checked, but the dataset did not contain exact duplicates requiring removal.

To address missing values in diag_2 and diag_3, records containing "?" in these columns were removed from the dataset, as diagnosis information is essential for predictive modelling. After filtering out these records, the dataset was reduced to 99,982 patient encounters from the original 101,763.

Additionally, the ICD-9 codes in the diagnosis columns were mapped into broader disease categories, such as Circulatory (390-459), Respiratory (460-519), Endocrine (240-279), and Diabetes-related conditions (250.xx), to improve interpretability and feature engineering. This categorical mapping enabled a structured approach to analysing the relationship between different health conditions and hospital readmission trends.

b) Exploratory Data Analysis

To gain deeper insights into the dataset, exploratory data analysis (EDA) was conducted by categorizing the variables into categorical and numerical attributes. The categorical variables, including race, gender, age group, insulin usage, and medication changes, were analysed to determine their distribution and relationship with the target variable, readmitted. Similarly, numerical variables such as time_in_hospital, num_lab_procedures, num_medications, number_inpatient, number_outpatient, and number_emergency were examined to assess trends and statistical properties. Multivariate analysis was performed to explore how each independent variable influences hospital readmission, and visualizations such as bar plots, box plots, and correlation heatmaps were used for better interpretation.

A comprehensive analysis of categorical variables revealed distinct patterns influencing hospital readmission. Race distribution in Fig.1. indicated that Caucasians and African Americans have the highest readmission rates, at 35.09% and 8.63%, respectively when compared to other racial groups, suggesting that these two groups experience the highest diabetes-related hospital encounters. However, since readmission rates remain relatively consistent across races, factors beyond race, such as healthcare access and disease severity, may play a larger role in influencing hospital readmissions. Gender analysis demonstrated that female patients have a slightly higher readmission rate (25.22%) compared to males (20.86%), though the difference is minimal, suggesting gender does not significantly impact hospital readmission. Readmission rates increase with age (Fig.2.), peaking at 12.32% for 70–80 years and 10.21% for 60–70 years, indicating that older patients experience more diabetes-related hospital encounters. Patients with "Unknown" weight data (Fig.3.) form the largest group (96.85%) considering it to be an insignificant predictor. Patients discharged home had shown a stronger association with readmission rate of 26.64% (Fig.4.), while other discharge dispositions had a slightly lower rate (19.44%). Similarly, from Fig. 5. admission from the emergency department had the highest readmission rate (27.89%), followed by physician referrals (13.04%), suggesting that emergency cases are more likely to require follow-up care. Among medical specialties, explains that most diabetes-related encounters lack a recorded medical specialty, with 35.29% of non-readmitted and 30.32% of readmitted cases falling under "Other." The high number of unspecified specialties suggests limited predictive value for readmission. More than 80% of the Maximum Glucose Serum and HbA1c values were recorded as "None", making it an insignificant variable for the model.

Table 1: Understanding the hospital dataset

Feature Category	Feature Name	Feature Description	Sample Values
Demographic Variables	Race	Patient's race category	Caucasian, African American, Hispanic
	Gender	Patient's gender	Male, Female
	Age	Age group of the patient in 10-year intervals	[50,60), [70,80), [80,90)
Clinical Metrics	Time in Hospital	Number of days between admission and discharge	3, 5, 7
	Num Lab Procedures	Number of lab tests performed during the encounter	44, 51, 61
	Num Procedures	Number of non-laboratory procedures performed	0, 1, 2
	Num Medications	Number of distinct medications administered	5, 10, 16
	Number of Diagnoses	Total number of diagnoses recorded	1, 4, 9
Prior Healthcare Utilization	Number of Inpatient	Number of inpatient admissions in the previous year	0, 1, 3
	Number of Outpatient	Number of outpatient visits in the previous year	0, 2, 6
	Number of Emergency	Number of emergency visits in the previous year	0, 1, 4
Medication Variables	DiabetesMed	Whether diabetes medication was prescribed	Yes, No
	Insulin	Insulin usage trend during the encounter	Up, Down, No
	Medication Status	Status of 23 diabetes-related medications	Steady, Up, Down, No
Target Variable	Readmitted	Readmission status of the patient	No, Within30Days, After30Days

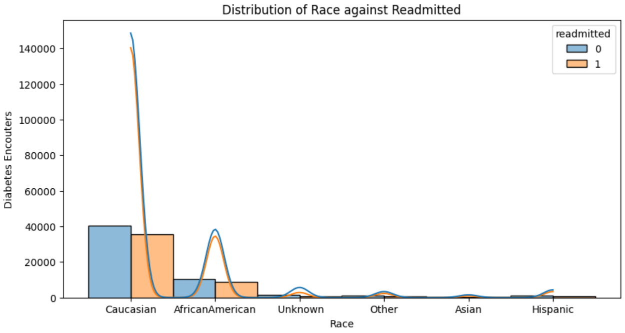

Fig. 1. Race Analysis

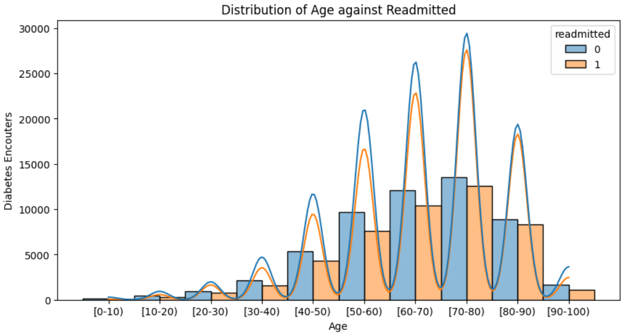

Fig. 2. Age Distribution

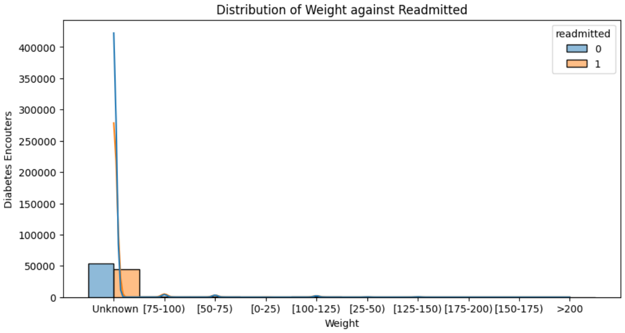

Fig. 3. Weight Distribution

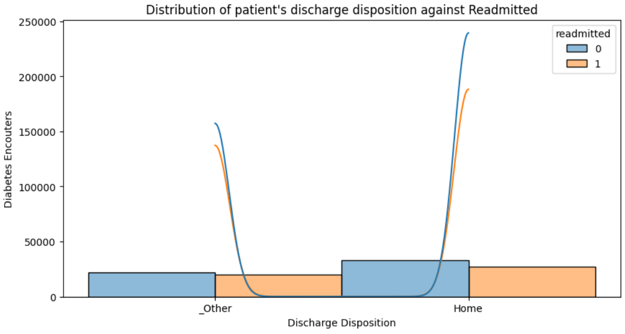

Fig. 4. Hospital Discharge Analysis

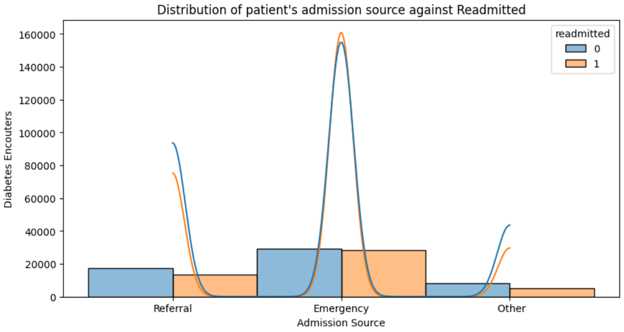

Fig. 5. Hospital Admission Source Analysis

For numerical variables, box plots were used to identify and remove outliers using the Interquartile Range (IQR) method. The following Table 2 presents the variables where outlier records above Q3+1.5 IQR were detected and removed:

Table 2: Summary of Outlier removal

Variable	Initial Record Count	Outliers Removed	Final Record Count
number_inpatient	1,01,763	1,256	1,00,507
number_outpatient	1,01,763	978	1,00,785
number_emergency	1,01,763	1,104	1,00,659
num_lab_procedures	1,01,763	847	1,00,916
num_procedures	1,01,763	532	1,01,231
num_medications	1,01,763	1,013	1,00,750
tme_in_hospital	1,01,763	764	1,01,009
number_diagnoses	1,01,763	482	1,01,281

Outliers accounted for a maximum of 1% in the analyzed numerical variables. Their removal had minimal impact on the dataset while effectively eliminating extreme values that could compromise model performance, ensuring the preservation of essential patient characteristics.

A correlation heatmap (Figure 6) was generated to assess relationships between numerical variables. Strong multicollinearity was detected among inpatient, outpatient and emergency numerical variables indicating overlapping information. A Variance Inflation Factor (VIF) analysis confirmed that number_inpatient was the strongest predictor, leading to the removal of number_outpatient and num_medications exhibited moderate correlation, but both were retained due to their distinct clinical relevance.

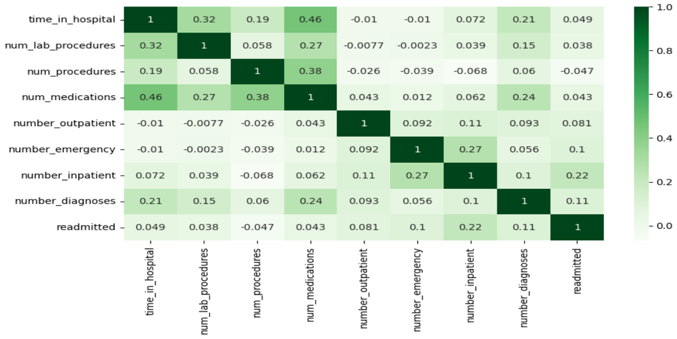

Fig. 6. Correlation Heatmap

c) Feature Extraction and Scaling

Following the exploratory data analysis, feature extraction and scaling were performed to prepare the dataset for modeling. Categorical variables identified as significant predictors, such as race, age group, insulin usage, discharge disposition and admission source and medication changes, were transformed using one-hot encoding to convert them into numerical representations suitable for machine learning algorithms. This encoding process created separate binary columns for each category within these variables, ensuring that categorical attributes were effectively incorporated without introducing ordinal relationships where none existed.

Additionally, the diagnosis variables (diag_1, diag_2, and diag_3), which originally contained alphanumeric ICD-9 codes, were mapped to broader disease categories based on predefined groupings. This mapping process transformed diagnosis codes into numerical values representing major diagnostic groups, such as circulatory, respiratory, and endocrine disorders, making them more interpretable and useful for modelling.

For numerical variables, including time_in_hospital, num_lab_procedures, num_medications, number_outpatient, number_emergency, and number_inpatient, MinMax Scaling was applied to normalize their values within a fixed range of 0 to 1. This transformation helped maintain the relative differences between values while preventing any single feature from dominating the model due to its scale.

The combined application of one-hot encoding, diagnosis mapping, and MinMax Scaling ensured that all selected features from the EDA phase were standardized and optimized for further predictive modeling.

d) Model Development

The Generalized Linear Model (GLM), specifically logistic regression, was chosen for predicting hospital readmission due to its interpretability and effectiveness in handling binary classification problems. Logistic regression models the probability of a binary outcome using a logistic (sigmoid) function, making it suitable for predicting whether a patient will be readmitted or not. The coefficients obtained from the model provide valuable insights into how each independent variable influences the likelihood of readmission. GLM is particularly useful in healthcare analytics due to its ability to quantify relationships between independent variables and outcomes while maintaining computational efficiency and ease of interpretation.

To build the model, the dataset was divided into features (X) and the target variable (y), where X contained 73 independent variables selected during feature engineering, and y represented the binary readmission status. A train-test split was performed to divide the dataset into 80-20 training and testing subsets. The GLM model with a binomial family was then applied to the training dataset, where it was trained on X_train and y_train. After fitting the model, a summary of the regression results was generated, displaying coefficient estimates, standard errors, P-values, and confidence intervals, providing a statistical overview of the model's significance and performance.

An iterative modelling approach was employed to refine the logistic regression model by eliminating non-significant variables. The summary statistics were carefully examined, particularly focusing on P-values to determine the significance of each variable. Variables with P-values greater than 0.05 were considered statistically insignificant and were iteratively removed from the model. Additionally, Variance Inflation Factor (VIF) values were analysed to detect multicollinearity among predictor variables, ensuring that all retained variables had VIF values below 5 to mitigate redundancy and improve model reliability. This iterative process was repeated thrice until only statistically significant variables remained, optimizing the model's predictive power and interpretability.

Table 2: Variance Inflation Factor (VIF)

Feature	VIF
number_diagnoses	7.858188
num_medications	7.614584
diabetesMed_Yes	5.474549
admission_source_id_Emergency	5.236502
diag_1	4.563307
diag_2	4.200408
race_AfricanAmerican	3.914901
diag_3	3.628827
admission_source_id_Referral	3.11562
time_in_hospital	3.104373
age_[70-80)	2.465274
age_[60-70)	2.18276
age_[80-90)	2.099332
discharge_disposition_id__Other	1.956552
age_[50-60)	1.885442
insulin_Steady	1.697175
number_inpatient	1.373799
metformin_Steady	1.370972
race_Hispanic	1.310563
race_Other	1.238432
rosiglitazone_Steady	1.092714
repaglinide_Steady	1.024401
metformin_Up	1.022018
acarbose_Steady	1.00482

Once the model was finalized, predictions were made on the training dataset (X_train) to assess its performance. The model generated predicted probabilities for each record, which were then compared against y_train (actual readmission values) and obtained an accuracy of 61%.

Following the training phase, the model was tested on X_test, the unseen portion of the dataset, to evaluate its generalization ability. Predictions were made on X_test, and the resulting values were compared against y_test to determine the model's accuracy on unseen data. The accuracy on the test dataset was observed to be 60.9% which is almost the same as observed on train dataset. The performance is assessed using the same evaluation metric as in training, ensuring that the model was not overfitting and could reliably predict hospital readmission.

By following this structured modelling approach, the final logistic regression model provided an optimized and interpretable solution for predicting patient readmission risk.

4. Model Diagnostics

To ensure the robustness of the logistic regression model, various diagnostic checks were conducted, including multicollinearity assessment, goodness-of-fit evaluation, residual analysis, and model performance validation. These checks helped in refining the model and improving its reliability in predicting hospital readmissions.

Multicollinearity was assessed using the Variance Inflation Factor (VIF) to identify redundant predictors that could inflate standard errors and affect model interpretability. The correlation heatmap and VIF analysis revealed high collinearity among number_inpatient, number_outpatient, and number_emergency, leading to the removal of number_outpatient and number_emergency while retaining number_inpatient as it showed the strongest relationship with readmission.

To determine the goodness-of-fit, the Hosmer-Lemeshow Test was performed, where a high p-value (>0.05) indicated that the model's predicted probabilities aligned well with observed outcomes. Residual analysis was conducted to detect systematic bias in the model, and the distribution of deviance residuals showed no significant patterns, confirming that the logistic regression assumptions were met.

The model's predictive capability was evaluated using multiple metrics. The accuracy of the model was 61%, meaning that 61% of all hospital encounters were correctly classified. The precision stood at 61%, indicating that 61% of the cases predicted as readmitted were actually readmitted. However, the recall was relatively low at 38%, meaning that only 38% of actual readmissions were correctly identified, suggesting potential challenges in capturing all readmitted patients. The F1-score, which balances precision and recall, was 47%, while the AUC Score was 59%, reflecting moderate discrimination between readmitted and non-readmitted patients.

A confusion matrix further detailed the model's classification performance:

Actual / Predicted	Predicted No Readmission	Predicted Readmission
Actual No Readmission	7537	1920
Actual Readmission	4844	3010

The confusion matrix revealed that while the model successfully classified 7537 true negatives (correctly predicting non-readmission), it struggled with recall, misclassifying 4844 actual readmissions as non-readmitted cases (false negatives). The accuracy of the model was recorded at 61%, indicating the percentage of correctly classified cases. The precision score was also 61%, meaning that when the model predicted a patient would be readmitted, it was correct 61% of the time. However, recall was lower at 38%, signifying that only 38% of the actual readmitted cases were correctly captured by the model. The F1-score, which balances precision and recall, was 47%, suggesting that the model maintained a moderate trade-off between both metrics. Lastly, the AUC score was 59% (Fig. 7.), reflecting the model's ability to distinguish between readmitted and non-readmitted patients.

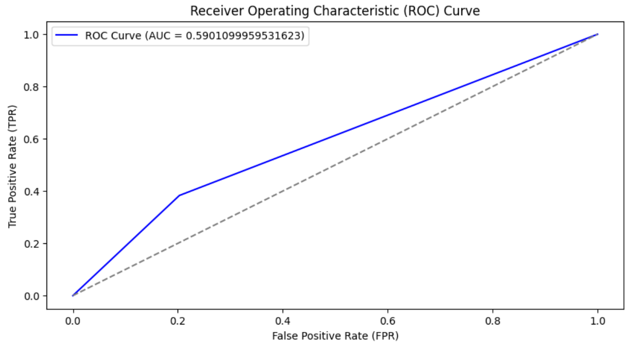

Fig. 7. ROC Curve

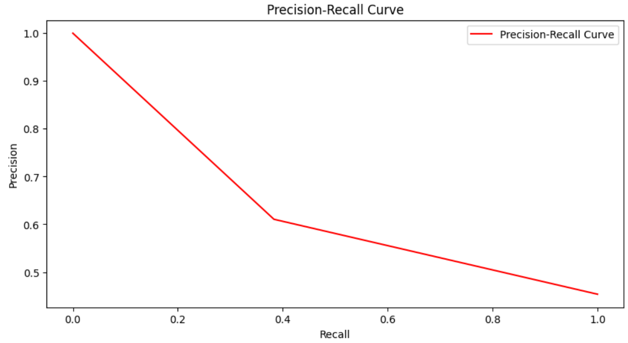

Fig. 8. Precision-Recall Curve

5. Conclusions

This study successfully develops a logistic regression model to predict hospital readmissions, achieving an accuracy of 61% and an AUC of 0.59**.** Key predictors, including prior inpatient visits, insulin usage, and primary diagnoses, provide actionable insights into identifying at-risk patients. The methodology—encompassing data preprocessing, exploratory analysis, iterative model refinement, and diagnostics—establishes a structured approach for analysing hospital readmissions. While HbA1c testing had limited predictive value due to missing data, the study reinforces the significance of clinical and historical patient data in understanding readmission risks and optimizing healthcare decision-making.

ACKNOWLEDGMENT

The author thanks the National College of Ireland for providing resources and support for this project, as well as the instructors Hicham Rifai for his guidance.

REFERENCES

[1] Beata Strack et al, "Impact of HbA1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records," BioMed Res. Int., vol. 2014, 2014, Art. no. 781670.