# Analysing and Predicting Diabetes Readmission

**Akshit Patel**
Faculty of Computer Science
Dalhousie University
akshit.patel@dal.ca

**Ashutosh Sagar**
Faculty of Computer Science
Dalhousie University
as890306@dal.ca

**Parthil Patel**
Faculty of Computer Science
Dalhousie University
parthilpatel66@dal.ca

**Preet Dudhat**
Faculty of Computer Science
Dalhousie University
preet.dudhat@dal.ca

## 1 Abstract

"Analysing and Predicting Diabetes Readmission," aims to address the pervasive issue of hospital readmissions among diabetic patients. These readmissions can impose a significant strain on healthcare systems and often reveal gaps in patient care and treatment procedures. The project endeavours to scrutinise the causes that contribute to readmission and to create predictive models employing machine learning methodologies. The specific objectives include the identification of crucial readmission predictors, the development of accurate predictive models, an exploration into the effects of patient characteristics on readmission rates, and the provision of actionable insights to healthcare professionals. The methodology incorporates data preprocessing, dataset balancing, and the application of various classification models such as Random Forest, XGBoost, Linear Regression, kNN and Gradient Boosting. Remarkable results were achieved, with the Gradient Boosting algorithm displaying an accuracy exceeding 90%, and other models also performing commendably with an approximate accuracy of 88%. The findings from this project have significant potential to enhance patient care, decrease healthcare costs, and improve industry practices by promoting proactive interventions to prevent readmissions among diabetic patients.

## 2 Introduction

The management of diabetes is a critical challenge for healthcare professionals, as patients with diabetes often face a higher risk of hospital readmission. Understanding the factors contributing to readmission and developing predictive models can significantly improve patient care, reduce healthcare costs, and enhance industry practices. This project aims to analyze and predict diabetes readmission using machine learning techniques, providing valuable insights for healthcare professionals, and contributing to advancing knowledge in the field.

### 2.1 Problem Statement

The specific problem addressed in this project is the high rate of readmissions among patients with diabetes. Hospital readmissions burden healthcare systems and indicate potential gaps in the quality of care and treatment plans for diabetic patients. By identifying the factors associated with readmission and developing predictive models, healthcare professionals can proactively address these issues and reduce readmission rates.

### 2.2 Objectives

- Analyze the factors contributing to readmission among diabetic patients and identify the critical predictors of readmission.
- Develop a predictive model using machine learning algorithms to predict the readmission likelihood accurately.

# 3 Literature Review

Several studies have been conducted to address the problem of readmissions among patients with diabetes. These studies have focused on identifying the factors contributing to readmission and developing predictive models to assess the likelihood of readmission. The following is a summary of previous related works and their strengths and weaknesses:

## 3.1 Identifying Diabetic Patients with High Risk of Readmission[1]

This study aimed to identify high-risk readmission patients with diabetes by developing a risk stratification framework incorporating demographic, clinical, and medication-related features [1]. The study's strengths lie in its comprehensive feature set and the creation of a risk stratification framework [1]. However, limitations include a relatively small sample size and no comparisons with other predictive models [1].

## 3.2 Predicting the Risk of Readmission of Diabetic Patients using MapReduce[2]

In this study, the authors utilized the MapReduce framework to predict readmission risk in diabetic patients [2]. They employed feature selection and trained a large dataset on a support vector machine (SVM) model [2]. The strengths of their work include the use of distributed computing for scalable analysis and the adoption of an SVM-based predictive model [2]. However, limitations include the need for more detailed feature selection discussions and the absence of performance comparisons with other algorithms [2].

## 3.3 The 30-day hospital readmission risk in diabetic patients: predictive modeling with machine learning classifiers[3]

This study aimed to create predictive models for the risk of 30-day hospital readmission in diabetic patients using machine learning classifiers [3]. The authors evaluated the performance of different classifiers such as decision trees, random forests, and support vector machines [3]. The study's strengths lie in comparing multiple classifiers and including a comprehensive set of features [3]. However, the study did not address the interpretability aspect of the predictive models [3].

## 3.4 An improved support vector machine-based diabetic readmission prediction[4]

This study proposed an improved support vector machine (SVM)-based approach for predicting diabetic readmissions [4]. The authors enhanced the SVM algorithm by incorporating a particle swarm optimization (PSO) algorithm to optimize the SVM parameters [4]. The strengths of this work include the optimization of SVM using PSO and the consideration of multiple features. However, the study did not compare the performance of their approach with other state-of-the-art models [4].

## 3.5 Predicting 30-Day Hospital Readmission for Diabetes Patients using Multilayer Perceptron by Ti'jay Goudjerkan and Manoj Jayabalan [5]

In this paper authors use predict the readmission of diabetes patients. Author uses SMOTE for balancing dataset[5]. They use four different types of models which are Random Forest Gini, CNN, RNN and Proposed MLP for appropriate prediction[5]. However author could not solve black box problem.[5]

# 4 Methodology

## 4.1 Dataset

The dataset used in this study is the "Diabetes 130-US hospitals for years 1999-2008"[6]. It comprises clinical care data spanning a 10-year timeframe, from 1999 to 2008, collected from 130 US hospitals and integrated delivery networks. The dataset encompasses over 50 features, representing both patient and hospital outcomes. In total, it contains more than 100,000 data points[6].

To ensure the relevance of the encounters used for analysis, five key criteria were applied to filter the data points. The selected encounters meet the following conditions: (1) they are hospital admissions, (2) the inpatient was classified as diabetic with at least one of three initial diagnoses including diabetes, (3) the length of stay ranged

from 1 to 14 days, (4) the inpatient underwent laboratory testing, and (5) the inpatient received medication during their stay.

## 4.2 EDA

During the exploratory data analysis (EDA) phase, the target column "readmitted" was analyzed to understand the distribution of classes. The "readmitted" column represents the readmission status of patients and is divided into three categories: "No" readmission, readmission "¿30 days" after discharge, and readmission "¡30 days" after discharge. The distribution of classes revealed that approximately 53.9% of patients had no readmission, while 34.9% were readmitted after more than 30 days, and 11.2% were readmitted within 30 days[6].

To facilitate the readmission prediction task, the patients falling into the "No" readmission and "greater than 30 days" readmission classes were merged to form one class, while the patients falling into the "less than 30 days" readmission class constituted another class[6]. This step was taken to create a binary classification problem to predict the likelihood of readmission for patients within 30 days, simplifying the predictive modeling process.

Furthermore, the EDA also involved investigating missing values in the dataset. Notably, the "weight" feature had a high proportion of missing values, accounting for 96.9% of the entries. Similarly, the "medical_specialty" feature exhibited a significant proportion of missing values, with approximately 49.1% of the data points lacking this information. Additionally, the "payer_code" feature had around 40% missing values.

## 4.3 Data Preprocessing

### 4.3.1 Custom Encoding for Drug Features

In this step, the 23 drug features in the dataset were transformed from categorical to numerical values using custom encoding. The encoding scheme assigned 'No' the value 0, while 'Steady', 'Up', and 'Down' were encoded as 1 [5]. This transformation allowed for the effective utilization of drug data in numerical-based algorithms, facilitating their integration into the predictive modeling process.

### 4.3.2 Ordinal Encoder for Age Feature

To incorporate the "age" feature as a numerical variable in the analysis, an ordinal encoder technique was employed [7]. By transforming the categorical age values into numerical values, the dataset became compatible with various machine learning algorithms, enhancing the accuracy and performance of the predictive models.

## 4.4 Feature Engineering

As part of feature engineering, efforts were made to address the challenge posed by the high number of distinct values in certain features. Notably, all three diagnosis features were found to have a large number of categories [5]. To enhance the manageability of the data, a clustering approach was applied to consolidate the diagnosis data. Over 800 distinct diagnosis categories were intelligently clustered into 10 broader categories, streamlining the representation of diagnostic information for analysis [5].

Additionally, similar clustering techniques were employed for two other features: Discharge Disposition and Admission Source. The original dataset contained 29 distinct categories for Discharge Disposition, which were thoughtfully reorganized into 6 more comprehensive categories [5]. Similarly, the dataset encompassed 26 different categories for Admission Source, which were thoughtfully consolidated into 7 more concise categories [5]. These clustering processes not only reduced the dimensionality of the features but also preserved essential information, enabling more effective utilization in the readmission prediction analysis.

## 4.5 Feature Selection

To identify the most relevant features for predictive modeling, the Chi-square test was utilized [7]. This statistical test assessed the association between the categorical features and the target variable (readmission). By evaluating the significance of each feature's association, irrelevant features were excluded from the analysis, resulting in a more efficient and informative feature set.

## 4.6 Model Selection

In the context of readmission prediction for diabetic patients, selecting an appropriate classification model is crucial for achieving accurate and reliable results. To address this, multiple classification models were considered: Logistic Regression, RandomForest, XGBoost, kNN, and Gradient Boosting [5] [7]. Each model possesses distinct strengths and capabilities that make them suitable for different types of data and tasks.

### 4.6.1 Logistic Regression

Logistic Regression is a straightforward and interpretable model, often used for binary classification tasks. It is well-suited for datasets with linearly separable classes, making it a suitable candidate for predicting readmission outcomes based on various patient and hospital factors.

### 4.6.2 RandomForest

RandomForest is an ensemble learning method that combines multiple decision trees to improve predictive accuracy and reduce overfitting. Its ability to handle a mix of categorical and numerical features and capture complex interactions among variables makes it valuable for dealing with the dataset's diverse and potentially nonlinear relationships.

### 4.6.3 XGBoost

XGBoost is an optimized implementation of gradient boosting, a powerful ensemble method that iteratively builds multiple weak learners to form a strong predictive model. XGBoost's superior performance and robustness, along with its ability to handle missing data and imbalanced datasets, make it suitable for tackling the complexities of readmission prediction in the context of diabetic patients.

### 4.6.4 kNN

The k-Nearest Neighbors algorithm is a non-parametric method that classifies data points based on the majority class of their k nearest neighbors. kNN can be effective when the data exhibits local patterns and clusters, making it relevant for identifying readmission patterns among similar groups of patients.

### 4.6.5 Gradient Boosting

Similar to XGBoost, Gradient Boosting is an ensemble method that constructs a predictive model through the combination of weak learners. Its capacity to handle complex interactions and dependencies among features is valuable for capturing the intricate relationships in the dataset, which may be critical for accurate readmission prediction.

## 4.7 Performance Metrics

To evaluate the models' effectiveness and suitability for the readmission prediction task, various performance metrics were employed: Accuracy, Precision, Recall, F1 score, and AUC (Area Under the ROC Curve) [7]. These metrics provide valuable insights into the models' performance in different aspects:

- Accuracy: Measures the overall correctness of the model's predictions, indicating how often it correctly classifies readmission outcomes. It is an essential metric to assess the model's overall effectiveness in a balanced dataset.
- Precision: Evaluates the model's ability to correctly classify positive cases (i.e., readmission) among the predicted positive cases. A high precision score indicates a low rate of false positives, which is crucial for preventing unnecessary interventions for patients predicted to be readmitted.
- Recall: Also known as Sensitivity or True Positive Rate, measures the model's ability to correctly identify all positive cases among the actual positive cases. High recall is essential to ensure that patients who are at risk of readmission are accurately identified and receive timely interventions.

- F1 Score: The harmonic mean of precision and recall, the F1 score balances both metrics, making it a suitable metric for imbalanced datasets, where readmission cases might be relatively rare compared to non-readmission cases.
- AUC (Area Under the ROC Curve): Represents the model's ability to distinguish between readmission and non-readmission classes, regardless of the chosen classification threshold. A higher AUC score indicates a more robust model with better discrimination ability.

### 4.8 Hyperparameter Tuning

To optimize the models' performance, hyperparameter tuning was conducted using cross-validation. The process involved systematically adjusting model parameters to find the optimal combination that maximized predictive accuracy and minimized overfitting. GridSearchCV was utilized to explore different hyperparameter configurations and identify the best parameters for each model, resulting in enhanced model performance [5].

## 5 Experiments

In pursuit of improving prediction accuracy, two powerful ensemble techniques, Bagging and AdaBoost, were applied to the classification models. Bagging involves training multiple instances of the same model on different subsets of the dataset and then averaging their predictions to reduce variance and improve overall accuracy. On the other hand, AdaBoost focuses on iteratively refining the model by assigning higher weights to misclassified samples, allowing for the creation of a strong ensemble learner. Furthermore, Gradient Boosting, a boosting method that constructs a strong model by sequentially adding weak learners, was also explored to enhance prediction performance.

To optimize the models and achieve the best possible performance, a thorough hyperparameter tuning process was carried out. The GridSearchCV technique with 4-fold cross-validation was employed to search through a range of hyperparameter values for each model. This approach allowed for the systematic exploration of hyperparameter configurations, enabling the identification of the best parameters that maximized the models' predictive capabilities and generalization.

To address the issue of class imbalance in the dataset, the Synthetic Minority Over-sampling Technique (SMOTE) algorithm was utilized for data balancing. SMOTE generates synthetic instances of the minority class by interpolating between existing data points, thus augmenting the dataset and balancing the class distribution. By employing SMOTE, the models were able to learn from a more balanced dataset, mitigating potential biases and improving their ability to accurately predict readmission outcomes.

## 6 Results

Table 1: Performance Metrics for Different Models

| Model | Accuracy | Precision | Recall | F1-score | AUC |
|---|---|---|---|---|---|
| Logistic Regression | 0.84 | 0.36 | 0.01 | 0.02 | 0.50 |
| Random Forest | 0.88 | 0.62 | 0.006 | 0.01 | 0.50 |
| XGBoost | 0.88 | 0.5 | 0.01 | 0.03 | 0.50 |
| kNN | 0.87 | 0.80 | 0.88 | 0.83 | 0.50 |
| Gradient Boosting | 0.92 | 0.99 | 0.85 | 0.91 | 0.92 |

**Logistic Regression:** The Logistic Regression classifier achieved an accuracy of 0.5875 and an AUC of 0.506 when predicting readmission status. The precision was 0.116, the recall was 0.402, and the F1 score was 0.179. When applying Bagging with Logistic Regression as the base estimator and using the SMOTE resampled data, the accuracy improved to 0.5918, and the AUC increased to 0.507. The precision was 0.116, the recall was 0.397, and the F1 score was 0.179.

**Random Forest:** The Random Forest classifier achieved an accuracy of 0.888 and an AUC of 0.503 when predicting readmission status. The precision was 0.577, the recall was 0.007, and the F1 score was 0.013. When applying Bagging and AdaBoosting techniques with different numbers of models, it was observed that increasing the number of models did not significantly improve the performance. The best-performing Bagging model with Random Forest as the base estimator achieved an accuracy of 0.888, a precision of 0.537, a recall of 0.013, and

an F1 score of 0.025. Similarly, the best-performing AdaBoost model with Random Forest as the base estimator achieved an accuracy of 0.888, a precision of 0.715, a recall of 0.002, and an F1 score of 0.004.

**XGBoost:** The XGBoost classifier achieved an accuracy of 0.8879 and an AUC of 0.506 when predicting readmission status. The precision was 0.390, the recall was 0.014, and the F1 score was 0.027. When applying Bagging with XGBoost as the base estimator and varying the number of models, it was observed that the performance did not improve significantly with more models. The best-performing Bagging model with XGBoost as the base estimator achieved an accuracy of 0.8879, a precision of 0.714, a recall of 0.002, and an F1 score of 0.004.

**kNN:** The kNN classifier without data balancing achieved an accuracy of 0.8786 and an AUC of 0.500 when predicting readmission status. The precision was 0.116, the recall was 0.402, and the F1 score was 0.179. However, after applying the SMOTE algorithm for data balancing, the accuracy improved to 0.6225, and the AUC increased to 0.506. The precision was 0.115, the recall was 0.397, and the F1 score was 0.179.

**Gradient Boosting:** The Gradient Boosting classifier achieved the best performance among all models with an accuracy of 0.924 and an AUC of 0.925 when predicting readmission status. The precision was 0.997, the recall was 0.852, and the F1 score was 0.919. The Gradient Boosting model demonstrated superior predictive capabilities, outperforming other models in terms of accuracy, AUC, precision, recall, and F1 score.

# 7 Conclusion

In conclusion, the project "Analysing and Predicting Diabetes Readmission" successfully addressed the critical issue of hospital readmissions among diabetic patients. By employing machine learning methodologies, the project developed accurate predictive models that can identify crucial readmission predictors and forecast the likelihood of readmission for diabetic patients. The Gradient Boosting model emerged as the top-performing algorithm, achieving an accuracy exceeding 90

The findings from this project have significant potential to enhance patient care, decrease healthcare costs, and improve industry practices by promoting proactive interventions to prevent readmissions among diabetic patients. Healthcare professionals can utilize these predictive models and actionable insights to identify high-risk patients and implement timely interventions, ultimately leading to better patient outcomes and a more efficient healthcare system.

The project's success underscores the value of data-driven approaches in addressing complex healthcare challenges and highlights the potential impact of machine learning in improving patient care and reducing the burden on healthcare systems.

# References


[1] M. S. Bhuvan, A. Kumar, A. Zafar, and V. Kishore, "Identifying Diabetic Patients with High Risk of Readmission.," Feb. 2016.2, doi: https://arxiv.org/abs/1602.04257.

[2] M. Gowsalya, K Krushitha, and C. Valliyammai, "Predicting the risk of readmission of diabetic patients using MapReduce," Dec. 2014, doi: https://doi.org/10.1109/icoac.2014.7229729.

[3] Y. Shang et al., "The 30-days hospital readmission risk in diabetic patients: predictive modeling with machine learning classifiers," BMC Medical Informatics and Decision Making, vol. 21, no. S2, Jul. 2021, doi: https://doi.org/10.1186/s12911-021-01423-y.

[4] S. Cui, D. Wang, Y. Wang, P.-W. Yu, and Y. Jin, "An improved support vector machine-based diabetic readmission prediction," Computer Methods and Programs in Biomedicine, vol. 166, pp. 123–135, Nov. 2018, doi: https://doi.org/10.1016/j.cmpb.2018.10.012.

[5] Ti'jay Goudjerkan and Manoj Jayabalan, "Predicting 30-Day Hospital Readmission for Diabetes Patients using Multilayer Perceptron," International Journal of Advanced Computer Science and Applications (ijacsa), 10(2), 2019, http://dx.doi.org/10.14569/IJACSA.2019.0100236.

[6] Diabetes 130-US hospitals for years 1999-2008, UCI Machine Learning Repository, Irvine, CA: University of California, School of Information and Computer Science, 2009. [Online]. Available: https://archive.ics.uci.edu/dataset/296/diabetes+130-us+hospitals+for+years+1999-2008. [Accessed: Aug. 1, 2023].

[7] Saurabh Raj, "Diabetes 130-US hospitals for years 1999-2008: Hospital Readmission," Medium, Nov. 2018. [Online]. Available: https://saurabhraj5162.medium.com/diabetes-130-us-hospitals-for-years-1999-2008-hospital-readmission-823ff48272f9. [Accessed: Aug. 1, 2023].