AN Najah National University

Faculty of Engineering and Technology

Predicting 30-Day Hospital Readmission for Diabetic Patients Using Machine Learning

Course Name: Machine Learning

Instructor:
Dr. Adnan Salman

Eng. Asem Saleh

Submitted by:

Zaina Abdalhaq

Zaina Dawod

Diabetes

1) Summary:

This project aims to develop a predictive model to identify diabetic patients at high risk of hospital readmission within 30 days after discharge. The dataset covers ten years (1999–2008) of clinical records from 130 U.S. hospitals and includes 101,766 inpatient encounters with 47 features. The task was formulated as a binary classification problem. After data preprocessing, Logistic Regression and Random Forest models were evaluated using cross-validation. Performance was assessed using accuracy, precision, recall, F1-score, and ROC-AUC. Logistic Regression achieved a ROC-AUC of 0.639 with a recall of 0.700, while Random Forest slightly outperformed it with a ROC-AUC of 0.654 and a recall of 0.726. Despite low precision, Random Forest showed better capability in identifying high-risk patients, supporting its use for reducing avoidable readmissions.

---

2) Introduction :

2.1 Motivation

Hospital readmission among diabetic patients is a major concern for healthcare systems due to its strong association with increased medical costs, higher morbidity, and elevated mortality rates. Diabetes is a chronic condition that often requires long-term management and frequent hospital care, particularly when patients experience acute episodes such as hyperglycemia or hypoglycemia, or chronic complications including cardiovascular disease, kidney failure, neuropathy, and infections. Early readmission, especially within 30 days of discharge, is widely regarded as an indicator of suboptimal care quality and ineffective post-discharge management.

Reducing preventable readmissions is a priority for hospitals, as it not only lowers operational costs but also improves patient outcomes and quality of life. Identifying patients who are at high risk of readmission allows healthcare providers to implement targeted interventions, such as closer monitoring, medication adjustments, and improved discharge planning. This project aims to address this challenge by developing a machine learning model capable of predicting early hospital readmission for diabetic patients.

2.2 Background

Previous studies have shown that many diabetic patients do not receive consistent and effective inpatient and post-discharge care, despite the availability of evidence-based guidelines. Factors such as poor glycemic control, medication non-adherence, irregular dietary habits, and the presence of multiple comorbid conditions significantly increase the likelihood of hospital readmission. Traditional risk assessment methods often fail to capture the complex interactions among these factors.

Machine learning techniques have increasingly been applied in healthcare to analyze large-scale clinical datasets and uncover patterns that are difficult to identify using conventional statistical approaches. In particular, classification algorithms such as Random Forest have demonstrated strong performance in medical prediction tasks due to their robustness, ability to handle mixed data types, and resistance to overfitting. This project builds on these ideas by applying machine learning methods to a large, real-world clinical dataset to predict 30-day readmission risk among diabetic patients.

---

3) Dataset :

3.1 Source

The dataset used in this project is the *Diabetes 130-US Hospitals for Years 1999–2008* dataset, donated in May 2014. It represents ten years of clinical care data collected from 130 hospitals and integrated delivery networks across the United States. The dataset is

publicly available and commonly used for research in healthcare analytics and machine learning.

3.2 Description

The dataset contains **101,766 instances**, where each instance represents a hospitalized patient encounter involving a diagnosis of diabetes. The dataset includes **47 features**, consisting of both categorical and integer variables. These features capture a wide range of information related to patient demographics, hospital admissions, laboratory tests, medications, and prior healthcare utilization.

The target variable indicates whether a patient was readmitted to the hospital within **30 days** of discharge, making this a binary classification problem. The dataset does not provide a recommended data split, allowing flexibility in choosing appropriate training and evaluation strategies.

3.3 Features

- **Encounter ID** Unique identifier of an encounter
- **Patient number** Unique identifier of a patient
- **Race** Values: Caucasian, Asian, African American, Hispanic, and other
- **Gender** Values: male, female, and unknown/invalid
- **Age** Grouped in 10-year intervals: 0, 10), 10, 20), ..., 90, 100)
- **Weight** Weight in pounds
- **Admission type** Integer identifier corresponding to 9 distinct values, for example, emergency, urgent, elective, newborn, and not available

- **Discharge disposition** Integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, and not available
- **Admission source** Integer identifier corresponding to 21 distinct values, for example, physician referral, emergency room, and transfer from a hospital
- **Time in hospital** Integer number of days between admission and discharge
- **Payer code** Integer identifier corresponding to 23 distinct values, for example, Blue Cross/Blue Shield, Medicare, and self-pay Medical
- **Medical specialty** Integer identifier of a specialty of the admitting physician, corresponding to 84 distinct values, for example, cardiology, internal medicine, family/general practice, and surgeon
- **Number of lab procedures** Number of lab tests performed during the encounter
- **Number of procedures** Numeric Number of procedures (other than lab tests) performed during the encounter
- **Number of medications** Number of distinct generic names administered during the encounter
- **Number of outpatient visits** Number of outpatient visits of the patient in the year preceding the encounter
- **Number of emergency visits** Number of emergency visits of the patient in the year preceding the encounter

- **Number of inpatient visits** Number of inpatient visits of the patient in the year preceding the encounter
- **Diagnosis 1** The primary diagnosis (coded as first three digits of ICD9); 848 distinct values
- **Diagnosis 2** Secondary diagnosis (coded as first three digits of ICD9); 923 distinct values
- **Diagnosis 3** Additional secondary diagnosis (coded as first three digits of ICD9); 954 distinct values
- **Number of diagnoses** Number of diagnoses entered to the system 0%
- **Glucose serum test result** Indicates the range of the result or if the test was not taken. Values: “>200,” “>300,” “normal,” and “none” if not measured
- **A1c test result** Indicates the range of the result or if the test was not taken. Values: “>8” if the result was greater than 8%, “>7” if the result was greater than 7% but less than 8%, “normal” if the result was less than 7%, and “none” if not measured.
- **Change of medications** Indicates if there was a change in diabetic medications (either dosage or generic name). Values: “change” and “no change”
- **Diabetes medications** Indicates if there was any diabetic medication prescribed. Values: “yes” and “no”
- 24 different kind of medical drugs.

- **Readmitted** Days to inpatient readmission. Values: if the patient was readmitted in less than 30 days, “>30” if the patient was readmitted in more than 30 days, and “No” for no record of readmission

---

3.4) preprocessing

The dataset contained missing values in several features, which is common in large-scale clinical data. As part of the preprocessing stage, non-informative features such as patient identification numbers were removed, as they do not contribute to the predictive task. The proportion of missing values was then calculated for each feature, and features with a high percentage of missing data were excluded from the analysis. To reduce the impact of extreme values, clipping was applied to selected numerical features. Feature engineering was performed to create more informative variables, and exploratory data analysis (EDA) was conducted to better understand feature distributions and relationships. Categorical variables were subsequently encoded into numerical representations suitable for machine learning algorithms. Finally, the dataset was split into training and testing sets, and cross-validation was used during model evaluation to ensure robust performance estimation.

4)Methodology:

Approach – Problem Definition & Evaluation Criteria:

This project addresses a supervised machine learning classification problem, where the goal is to predict whether a patient will be readmitted to the hospital within 30 days after discharge based on patient-related clinical and demographic features.

This project aims to develop a predictive model to identify diabetic patients at high risk of hospital readmission within 30 days after discharge. The dataset covers ten years (1999–2008) of clinical records from 130 U.S. hospitals and includes 101,766 inpatient encounters with 47 features. The task was formulated as a binary classification problem. After data preprocessing, Logistic Regression and Random Forest models were evaluated using cross-validation. Performance was assessed using accuracy, precision, recall, F1-score, and ROC-AUC. Logistic Regression achieved a ROC-AUC of 0.639 with a recall of 0.700, while Random Forest slightly outperformed it with a ROC-AUC of 0.654 and a

recall of 0.726. Despite low precision, Random Forest showed better capability in identifying high-risk patients, supporting its use for reducing avoidable readmissions.

---

5) Results :

5.1 Findings:

The performance of the evaluated models is summarized in Table X, which compares Logistic Regression and Random Forest using cross-validation. The results show that Logistic Regression achieved an accuracy of 0.524 and a ROC-AUC of 0.639. Random Forest achieved a slightly lower accuracy of 0.514 but outperformed Logistic Regression in terms of recall and ROC-AUC. Specifically, Random Forest obtained a recall of 0.726 for the readmitted class and a ROC-AUC of 0.654, indicating a better ability to identify patients at high risk of readmission.

The Result of the model:

```
Final Evaluation (Threshold = 0.3)
              precision    recall  f1-score   support

           0       0.93      0.49      0.64     18082
           1       0.15      0.73      0.25      2271

    accuracy                           0.51     20353
   macro avg       0.54      0.61      0.45     20353
weighted avg       0.85      0.51      0.60     20353

ROC AUC: 0.6544014884782183
Balanced Accuracy: 0.6069172210300247
```

Both models showed relatively low precision values (approximately 0.15), reflecting the imbalanced nature of the dataset, where non-readmitted cases dominate. Despite this limitation, Random Forest demonstrated improved sensitivity toward the minority class, making it more suitable for identifying early readmissions. Visualizations such as ROC

curves and comparison tables were used to illustrate model performance and highlight differences between the evaluated approaches.

Logistic Regression and Random Forest models are compared using accuracy, precision, recall, F1-score, and ROC AUC.

	Model	Accuracy	Precision (Class 1)	Recall (Class 1)	F1-score (Class 1)	ROC AUC
0	Logistic Regression (CV)	0.524100	0.150071	0.700132	0.247163	0.639058
1	Random Forest (CV)	0.514322	0.151118	0.726112	0.250171	0.654401

5.2 Discussion

The results indicate that machine learning models can effectively support the identification of diabetic patients at risk of early hospital readmission, particularly when recall is prioritized over precision. Random Forest performed better than Logistic Regression in capturing high-risk patients, likely due to its ability to model nonlinear relationships and interactions among clinical features. The slightly lower accuracy observed in Random Forest is acceptable in this context, as accuracy alone is not a reliable metric for imbalanced datasets.

One of the main challenges encountered in this project was the strong class imbalance, which contributed to low precision scores for both models. Additionally, missing values and potential noise in clinical records may have limited overall performance. Despite these challenges, the approach demonstrates practical value, as identifying a larger proportion of high-risk patients can enable hospitals to apply targeted interventions and reduce avoidable readmissions.

---

6) Conclusion :

This project investigated the use of machine learning techniques to predict 30-day hospital readmission among diabetic patients using a large-scale clinical dataset. By formulating the task as a binary classification problem, Logistic Regression and Random Forest models were evaluated and compared. The results showed that Random Forest slightly outperformed Logistic Regression, particularly in terms of recall and ROC-AUC, indicating a better ability to identify patients at high risk of early readmission. Although precision values were relatively low due to class imbalance, prioritizing recall is appropriate in a healthcare context where missing high-risk patients can have serious consequences.

Overall, the applied methods addressed the problem effectively by capturing important patterns in complex clinical data. This project highlighted the importance of proper preprocessing, feature selection, and evaluation metric choice when working with imbalanced medical datasets. Future work could focus on improving precision through advanced imbalance-handling techniques, incorporating additional clinical or temporal features, and exploring model interpretability methods to better understand the factors contributing to hospital readmissions.

---

7) References :

1. UCI Machine Learning Repository. *Diabetes 130-US Hospitals for Years 1999–2008 Dataset*.