Diabetic patient readmission prediction using machine learning

Jigyashu Rajput
MSc in Computing
jigyashu.rajput2@mail.dcu.ie
Dublin City University
Student Number: 23261290

Pranav Lodh
MSc in Computing
pranav.lodh2@mail.dcu.ie
Dublin City University
Student Number: 23263834

Shikha Gaur
MSc in Computing
shikha.gaur3@mail.dcu.ie
Dublin City University
Student Number: 23261823

Apurva Shirbhate
MSc in Computing
apurva.shirbhate2@mail.dcu.ie
Dublin City University
Student Number: 23266290

Abstract—**Diabetes is a severe condition that affects millions of people worldwide. It is caused due to lack of insulin production by the pancreas's or the body's inability to use insulin effectively. Patients with diabetes have a higher risk of readmission, which can be costly and detrimental to their health. Accurately estimating the risk of readmission is essential in creating strategies that lower readmission rates and enhance patient outcomes. Therefore, it is crucial to develop effective models and prediction algorithms to identify patients at high risk of readmission. The study aims to identify the factors associated with readmissions in patients with diabetes and create a reliable predictive model. The study involves various machine learning approaches previously used to detect the possibility of the patient being readmitted and analyse their results. Also, a large cohort of patients with diabetes in the US obtained from UCI is evaluated, and specific demographic and clinical factors present at initial admission are analysed. Using machine learning algorithms like Logistic Regression, XGBoost and CatBoost, this study develops a predictive model to accurately identify diabetic patients with higher risk for readmission within 30 days. CatBoost is the best-performing model, with an ROC AUC of 67.06% and an accuracy of 88.90%. The output from this research would help healthcare improve its services and minimise the readmission rate and medical treatment resources. The factors leading to a patient's readmission could also be rectified using the machine learning methodology.**

I. Introduction

Diabetes Mellitus is a significant public health problem, affecting millions of people worldwide. Diabetes is a chronic condition that requires ongoing management and care. Hospital readmission is a significant issue in managing diabetes [1]. Patients with diabetes are at higher risk for readmission to the hospital, which can lead to increased healthcare costs and potential complications. To address this issue, machine learning techniques have been used to predict the likelihood of readmission for diabetic patients [2]. In today's quickly evolving world, correctly predicting readmissions for diabetic patients is impossible to overestimate.

To forecast the readmission risks of patients with diabetes, we developed and compared machine learning-based readmission prediction techniques. The study leverages a dataset from the Health Facts Database, encompassing over 101,766 records of diabetic patients from 1999 to 2008, to develop and compare machine learning-based prediction models for 30-day hospital readmission risks. Utilizing advanced data preprocessing techniques and machine learning classifiers, including Random Forest, Naive Bayes, and Decision Tree Ensemble, the research identifies key predictive factors such as the number of inpatient admissions, patient age, and emergency visits [3]. The findings underscore the Random Forest model's superior performance, highlighting its potential in aiding healthcare providers to identify high-risk patients and implement targeted interventions to mitigate short-term readmissions. Numerous comorbidities, ethnicity (e.g. Hispanic), endocrinology consultation, diabetes self-management education, and prescription medication (oral and insulin) are a few of the significant aspects identified in the sources. Identifying these high-risk individuals can lead to targeted treatments that lower healthcare expenditures, improve patient outcomes, and avoid future hospital admissions. Machine learning analysis of this dataset can assist medical professionals in identifying diabetic patients who are more likely to require readmission. Forecasting readmission risks are more clinically efficient when machine learning techniques like Logistic Regression, CatBoost, and XGBoost are used. The study focuses on how data mining and feature engineering techniques could leverage the model performance and help identify the significant attributes which leads to a diabetic patient to get readmitted to the hospital within 30 days of discharge or is predicted to get admitted after the use of hospital services. To sum up, it is critical to anticipate readmission rates for diabetes patients correctly to improve patient care and save healthcare expenses. The accurate prediction of diabetic patient readmission using machine learning techniques has the potential to benefit significantly healthcare providers and patients alike [3].

II. Related Work

Healthcare systems have been increasingly concerned about the high readmission rates of patients with chronic diseases such as diabetes. We conducted a literature review to explore various methods for predicting hospital readmission among diabetic patients. Our review emphasizes the pivotal role that machine learning and deep learning approaches play in this area. For instance, one study described in [4] employed multiple machine learning algorithms, including Random Forest, to analyze a dataset of over 100,000 patients. The study highlighted the significance of medication use and in-patient visits in predicting readmission. This underscores the value of identifying key risk factors for model development. Similarly, another study proposed in [5] examined using machine learning models to determine readmission risk among diabetic patients in Singapore. Their research underscores the need for robust risk stratification to identify high-risk patients who may benefit from targeted interventions to prevent readmission.

Numerous studies have examined the effectiveness of various machine learning algorithms for predicting readmissions. One such study, cited in [3], achieved promising results in predicting 30-day readmission risk in diabetic patients using Random Forest and Support Vector Machines. Their methodology involved data pre-processing and feature selection, resulting in an AUC of 0.73 for the model. This illustrates the efficacy of traditional machine learning techniques and underscores the importance of data preparation for optimal performance. Building on this foundation, the investigation in [6] compared different machine learning methods and found that Random Forest and XGBoost yielded the highest accuracy (around 78%) for readmission prediction among people with diabetes. Their research highlights the significance of benchmarking different algorithms to select the best-performing model for a specific dataset.

Moreover, machine learning techniques have proved helpful in real-world scenarios. For instance, an analysis by [1] explored machine learning techniques on electronic health records for predicting short-term and long-term readmissions in uncontrolled diabetic patients. Their findings support the effectiveness of machine learning in identifying high-risk patients and demonstrate the potential real-world applications of these models.

As healthcare evolves, so do the techniques and methodologies for predicting patient readmissions. Deep learning approaches have emerged as a powerful alternative to traditional methods in recent years. A study outlined in [7] investigated using a Convolutional Neural Network (CNN) for readmission prediction and achieved an accuracy of 79% on a dataset of over 10,000 diabetic patients. Their approach combined CNNs with data engineering techniques to improve performance, showcasing the potential of deep learning architectures for readmission prediction. Additionally, research proposed in [8] introduced a ”guided neural network” specifically designed for early readmission prediction in diabetic patients. While specific accuracy metrics were not reported, their approach demonstrated promise for early intervention using deep learning. These findings underscore the potential of deep learning for targeted interventions that can improve patient outcomes.

Both machine learning and deep learning possess their own set of advantages and disadvantages. While interpretable machine learning approaches, as evidenced by [3], enable the user to comprehend the reasoning behind predictions, they may not achieve the highest accuracy. On the other hand, deep learning models, as exemplified by the works of [7], may attain higher accuracy but can be less interpretable, making it difficult to comprehend how they arrive at their predictions. As highlighted by [9], encouraging the interpretability of machine learning models to integrate these tools into clinical practice is paramount.

As the reference [10] emphasizes, comprehending the risk factors associated with hospital readmission is crucial as it aids in developing effective models. The reference highlights factors such as poor glycemic control and frequent emergency department visits as significant contributors to readmission risk in diabetic patients. Integrating this knowledge into feature selection for machine learning and deep learning models can enhance their effectiveness.

Despite the potential of machine learning and deep learning, several challenges must be addressed. Data quality, model selection, and hyperparameter tuning can significantly influence these models, as sources [4] and [6] highlight. Additionally, the generalizability of these models across diverse healthcare settings presents a significant hurdle. Therefore, future research should focus on developing robust and generalizable models that can perform well in real-world scenarios with varying patient populations and healthcare systems, as noted by source [2].

Moreover, ensuring the interpretability of deep learning models is essential. While traditional machine learning offers interpretability, techniques to enhance the interpretability of deep learning models, as explored by [11], would be valuable. Additionally, incorporating clinical expertise into model development, as suggested in future directions discussed by [12], can improve real-world applicability.

Notably, social determinants of health can also impact readmission risk and medical factors. The study by [2] focused on a model for predicting, diagnosing, and mitigating health disparities in hospital readmission, emphasizing the need for a holistic approach that considers medical and social factors impacting patient health.

As highlighted by [12], standardized data collection methods can facilitate model development and comparison across multiple studies. The growing body of research in this area, including the work mentioned earlier, underscores the need for more specific details on datasets and models. By fostering collaboration among researchers, healthcare professionals, and data scientists, we can address these challenges and unlock the full potential of machine learning and deep learning to optimize patient care and reduce readmission rates for diabetic patients.

In conclusion, this comprehensive review provides valuable insights into machine learning and deep learning's potential for predicting hospital readmission rates for diabetic patients. By leveraging these technologies, we can improve patient outcomes and drive greater efficiency in healthcare delivery.

III. Data Mining and Methodology

A. *Dataset Description:*

The dataset represents a decade (1999-2008) of clinical care across 130 US hospitals and integrated delivery networks. The study encompassed 50 attributes and 101,766 entries: admission times, sex, age, admission type, length of hospital stay, number of laboratory tests, glycosylated haemoglobin results, diagnosis, and medication. The dataset included clinical records of diabetic inpatients who stayed in the hospital for 1-14 days, along with the laboratory tests and medications used during their hospitalization.

B. CRISP-DM Framework Application

In addressing our research questions, we adhered to the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology. This structured approach guided us through the essential phases of our analysis. Following the CRISP-DM methodology not only structured our analysis but also ensured a thorough and systematic approach to addressing the research question concerning the factors influencing the readmission of diabetic patients [13].

1) Business Understanding: We identified the critical need to explore the factors influencing the readmission of diabetic patients. Our goal was to understand patterns and predictors that could improve patient outcomes and inform hospital practices.

2) Data Analysis: Through a thorough exploratory analysis of our dataset, we discovered that the target variable ”readmitted” was imbalanced, as depicted in the Figure 1. Moreover, we observed some missing records in our dataset, as also illustrated in the Figure 2. On a positive note, we found that certain columns, such as gender, were well-balanced (Fig 3) and the ”Unknown” were removed as there was only three entries.

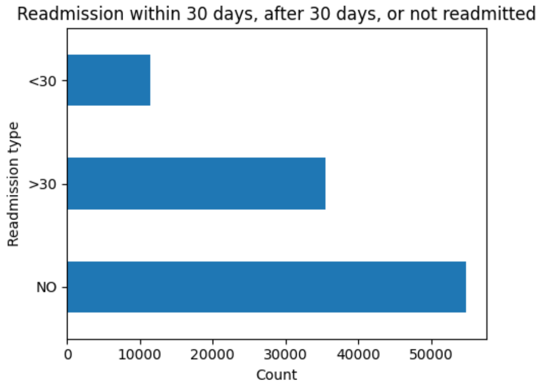
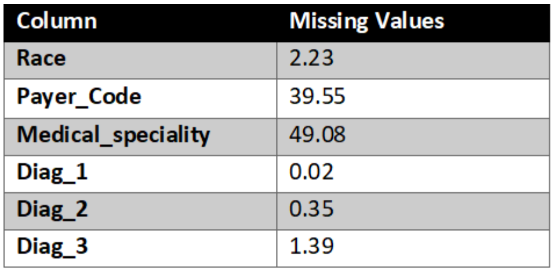


Fig. 1. ”Readmitted” Distribution.



Column	Missing Values
Race	2.23
Payer_Code	39.55
Medical_speciality	49.08
Diag_1	0.02
Diag_2	0.35
Diag_3	1.39

Fig. 2. Missing Value Information.

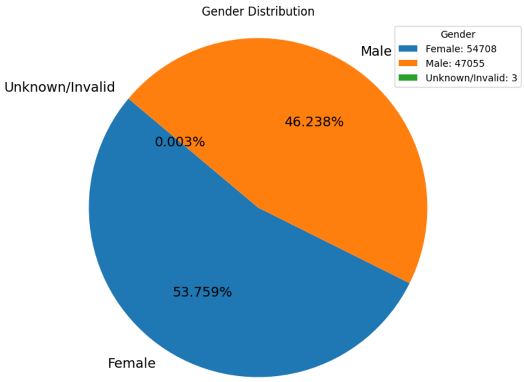


Fig. 3. Gender Distribution.

3) Data Cleaning and Preprocessing: Our team has meticulously prepared the dataset by thoroughly cleaning and preprocessing the data. Initially, we transformed the age column, which was range, into an ordinal variable and then carried out missing value treatment on the dataset. Features that had over 50% of missing values, such as max_glu_serum, A1Cresult, weight, sitagliptin, examine, payer_code, and medical_specialty, were eliminated from the dataset since it would make our model biased. We also removed unidentified entries in the gender column, as only a few of them existed. Furthermore, we treated missing values in the diag_1, diag_2, and diag_3 columns with KNN imputation. We observed that these values were missing at random (MAR), indicating no specific percentage of missing values based on certain criteria like gender. We applied the same imputation technique to the race column.

4) Feature Engineering: In our data analysis process, we systematically grouped similar categories - like for the discharge_disposition_id column; we clubbed unknown and uncategorized values together. We performed the same for other entries in the discharge_disposition_id and admission_source_id columns. For the target variable, We combined more than 30 days and No values, as we aim to predict readmission within 30 days. We assigned less than 30 days as one and the other variable as 0, converting it to binary from multi-class classification [3]. We also conducted outlier detection to account for numerical values. However, we noticed a significantly large number of outliers in the boxplot, so we decided to ignore them as outliers. For the attribute values of diag_1, diag_2 and diag_3 (three disease diagnoses), the values were represented by the International Statistical Classification of Diseases and Related Health Problems (ICD-9) code [14]. The attributes number_outpatient, number_emergency, and number_inpatient represent the services used or the number of hospital visits. So, we used feature extraction and dimension reduction on them, merged them into one column, and named it service_utilisation. At the data processing stage, similar diagnoses were combined into 16 disease categories using the ICD-9 coding set. This was done to improve analysis since scattered diagnoses produced unfavorable results. Initially, some medication columns, such as metformin, repaglinide, nateglinide, chlorpropamide, etc., were in categorical format. To make it easier for the model to predict the output, we performed one hot encoding to convert categorical values into numerical values. Furthermore, we found that the dataset was imbalanced, meaning that the number of samples in each class was unequal. To address this issue, we used Synthetic Minority Oversampling Technique (SMOTE) analysis to add samples for the minority class in the data, making a balanced dataset and ensuring our model was not biased towards any particular class. The oversampling technique should be performed on the training set, while the training and testing test should be split before applying SMOTE.

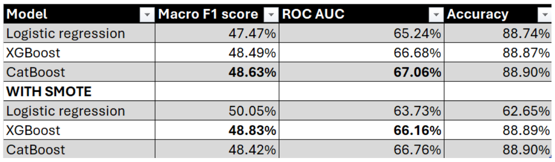
5) Model Selection: Through our examination of the diabetes dataset, we have encountered various machine learning models, each possessing unique strengths. The Logistic Regression model is a basic yet effective tool for binary classification, laying the groundwork for more complex analyses; hence, we are using it as our base model. The XGBoost model, a leading gradient-boosting framework, enhances performance by iteratively rectifying previous model mistakes, proving invaluable for datasets with intricate variable interactions and can handle imbalance class data. Lastly, the CatBoost model addresses categorical data with minimal preprocessing and imbalance class handling, mitigating overfitting while efficiently managing data, making it particularly well-suited for diverse datasets such as those commonly encountered in medical records.

6) Prediction Model Construction: Before the model training, the processed dataset is divided into 90% training sets and 10% testing sets. Cross-validation is implemented to obtain an unbiased and general estimate of each model's performance on the dataset. By performing cross-validation, we can reduce overfitting and improve the model's ability to generalise to unseen instances. The primary reason for using cross-validation is to get better train and test data split and to understand the model's performance and make accurate predictions on the data. Twenty features are used as input to the model towards one target variable. We have used GridSearch for hyperparamter tuning of Machine learning. In machine learning, grid search is used to select the best set of hyperparameters for a learning algorithm. A hyperparameter is a parameter whose value is set before the learning process begins. Unlike model parameters, which are learned during training, hyperparameters have a significant impact on the performance of machine learning models.

IV. Results

We compared the performance of different models based on AUC and macro F1 score as there are better metrics to measure the model's performance than accuracy. The models include Logistic Regression, XGBoost, CatBoost and Multinomial Naive Bayes, using the table 4 to compare performance.

The data presented in the table show that the CatBoost model without SMOTE performed the best. This model achieved the highest ROC AUC score (67.13%) and Accuracy (88.88%), crucial metrics in evaluating classification models. The Macro F1 score for CatBoost is also impressively high (48.65%), only slightly lower than the highest F1 score achieved by XGBoost without SMOTE (48.97%). Despite XGBoost without SMOTE having a slightly higher Macro F1 score, the CatBoost model's superior ROC AUC and matching Accuracy make it a more comprehensive choice. This is particularly significant considering the role of AUC in assessing a model's ability to differentiate between classes. Furthermore, even when SMOTE is applied, CatBoost's performance remains strong, underscoring its ability to adapt to changes in data sampling techniques.



Model	Macro F1 score	ROC AUC	Accuracy
Logistic regression	47.47%	65.24%	88.74%
XGBoost	48.49%	66.68%	88.87%
CatBoost	**48.63%**	**67.06%**	88.90%
WITH SMOTE			
Logistic regression	50.05%	63.73%	62.65%
XGBoost	**48.83%**	**66.16%**	88.89%
CatBoost	48.42%	66.76%	88.90%

Fig. 4. Evaluation Results.

V. Conclusion

Our research involved several preprocessing techniques, including feature encoding, imputation, feature extraction, feature selection, and outlier handling. To address the class imbalance issue, we employed the Synthetic Minority Oversampling technique. We ensured the reliability and robustness of our model performance by utilising K-fold cross-validation. We used machine learning models like Logistic Regression, XGBoost, and CatBoost for prediction. CatBoost performs the best with ROC AUC of 67.13% and an accuracy of 88.88%. The model's hyperparameters could be tuned to enhance the performance. The dataset that we have used for diabetic patient readmission prediction is limited in terms of attributes, as diabetes does contain many factors associated with it. Main features like weight were 98% missing data, which can significantly predict the patient's readmission. The models included in this research only work on diabetes dataset. In future, a more sophisticated model can be constructed to apply the prediction of readmission involving several other diseases.

References

[1] M. Mahmoud, M. Bader, and J. McNicholas, "Short-term and long-term readmission prediction in uncontrolled diabetic patients using machine learning techniques.," in *HEALTHINF*, pp. 680–688, 2023.

[2] S. Raza, "A machine learning model for predicting, diagnosing, and mitigating health disparities in hospital readmission," *Healthcare Analytics*, vol. 2, p. 100100, 2022.

[3] Y. Shang, K. Jiang, L. Wang, Z. Zhang, S. Zhou, Y. Liu, J. Dong, and H. Wu, "The 30-days hospital readmission risk in diabetic patients: predictive modeling with machine learning classifiers," *BMC medical informatics and decision making*, vol. 21, pp. 1–11, 2021.

[4] A. Sharma, P. Agrawal, V. Madaan, and S. Goyal, "Prediction on diabetes patient's hospital readmission rates," in *Proceedings of the Third International Conference on Advanced Informatics for Computing Research*, pp. 1–5, 2019.

[5] K. Teo, C. W. Yong, J. H. Chuah, Y. C. Hum, Y. K. Tee, K. Xia, and K. W. Lai, "Current trends in readmission prediction: an overview of approaches," *Arabian journal for science and engineering*, vol. 48, no. 8, pp. 11117–11134, 2023.

[6] D. Wang *et al.*, "A comparison of machine learning methods to predict hospital readmission of diabetic patient," *Studies of Applied Economics*, vol. 39, no. 4, 2021.

[7] A. Hammoudeh, G. Al-Naymat, I. Ghannam, and N. Obied, "Predicting hospital readmission among diabetics using deep learning," *Procedia Computer Science*, vol. 141, pp. 484–489, 2018.

[8] A. A. Ram, Z. Ali, V. Krishna, N. Nishika, and A. Sharma, "A guided neural network approach to predict early readmission of diabetic patients," *IEEE Access*, 2023.

[9] P. Hu, S. Li, Y.-a. Huang, and L. Hu, "Predicting hospital readmission of diabetics using deep forest," in *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, pp. 1–2, IEEE, 2019.

[10] J. G. S. Soh, W. P. Wong, A. Mukhopadhyay, S. C. Quek, and B. C. Tai, "Predictors of 30-day unplanned hospital readmission among adult patients with diabetes mellitus: a systematic review with meta-analysis," *BMJ Open Diabetes Research and Care*, vol. 8, no. 1, p. e001227, 2020.

[11] P. Mathur, S. Leburu, and V. Kulothungan, "Prevalence, awareness, treatment and control of diabetes in india from the countrywide national ncd monitoring survey," *Frontiers in public health*, vol. 10, p. 748157, 2022.

[12] M. M. Sosa and D. F. Hernandez, "Predictive modeling of diabetes hospital readmission using machine learning algorithms," in *2023 3rd International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*, pp. 1–6, IEEE, 2023.

[13] C. Schröer, F. Kruse, and J. M. Gómez, "A systematic literature review on applying crisp-dm process model," *Procedia Computer Science*, vol. 181, pp. 526–534, 2021.

[14] Wikipedia: The Free Encyclopedia, "List of icd-9 codes." http://psychology.wikia.com/wiki/List_of_ICD-9_codes, 2014. Accessed: 2020-06-06.