Predicting Diabetic Patients' Re-admission to hospitals using Multi-Classification Machine Learning Models

by Hager Ghouma

Executive summary

Diabetic patients are at a high risk of being readmitted to hospitals, which can have a significant impact on their lives and healthcare expenses. This study utilizes the diabetic readmission dataset to predict readmission of high-risk patients. The dataset includes information on diabetic patients admitted to 130 US hospitals between 1999-2008, including demographic, diagnostic, and treatment factors, as well as hospital stay and readmission data. Accurate predictions of diabetic patient readmissions can enable targeted interventions to improve patient outcomes.

The data was pre-processed by handling missing and unnecessary data, and categorical variables were encoded. Feature scaling was performed using standardization, and feature selection was done using the Anova F-test method. The selected features were used to extract X, y from the dataset, which were then split into training and testing sets. Three machine learning models (Logistic Regression, Decision Tree with Hyperparameter Tuning, Random Forest) were used to predict patients' readmission.

The study found that medical conditions such as emergency visits, inpatient stays, and diagnoses are critical predictors of patient readmission, with age range 50-90 being a crucial factor. Healthcare providers should focus on targeting patients with chronic or severe medical conditions within this age range to reduce readmission rates. Gender was found to be less significant, and interventions should address patients' medical conditions. However, the study faced class imbalance issues and poor performance, requiring further work to improve model accuracy by tuning model parameters and trying different feature selection techniques.

The study also found that the length of hospital stay is the most critical predictor of hospital readmission. Healthcare providers can reduce hospital stays by implementing strategies such as care coordination and comprehensive discharge planning. Other important features, such as the number of lab procedures, admission source, discharge disposition, and number of procedures, can help identify high-risk patients who are likely to be readmitted. This information can be used to develop predictive models and targeted interventions to prevent readmissions. Overall, the study emphasizes the importance of identifying high-risk patients and providing appropriate care to improve patient outcomes and reduce healthcare costs.

Introduction

Diabetes mellitus is a chronic health condition where the body can not regulate blood sugar due to a lack of insulin production or the body's inability to use it. This condition can cause dangerously high blood sugar levels overtime leads to health problems including and not limited to heart disease, kidney damage and vision loss[1]. Effective management of diabetic patients is crucial to prevent health complications that may require hospitalization.

Unfortunately, diabetic patients have a high risk of readmission to hospitals. The rate of 30-day readmissions among hospitalized patients with diabetes mellitus is reportedly between 14.4% and 22.7%, which is significantly higher compared to the rate of 8.5% to 13.5% for all hospitalized patients(3). The cost of treating diabetes mellitus in Ontario is $1.5 billion in 2019(4) and the national cost of treating diabetes is just under $30 billion in the same year, compared to $14 billion in 2008(4). Acute hospitalizations were the primary contributor to these costs, accounting for 43.2% of the total cost(2).

The impact of hospital readmissions on the quality of life and healthcare expenditure of diabetic patients is devastating. Predictive modeling can help healthcare institutions identify which patients are at higher risk of readmission and take appropriate steps to minimize this risk. By implementing measures such as regular monitoring, and follow-up, and patient education, healthcare providers can improve the management of diabetes and reduce the risk of hospital readmissions, leading to better patient outcomes and reduced healthcare costs.
The approach used in this study involves preprocessing the data by handling missing values, encoding categorical variables, and scaling the data. Then, the preprocessed data is split into training and testing sets, and each model is trained on the training set and evaluated on the testing set. The performance of each model is measured using various metrics, such as accuracy, precision, recall, and F1 score. Finally, the models are compared based on their performance and the most accurate model is selected for predicting readmission rates in diabetic patients.

Problem definition

The problem addressed in this report is the high rate of hospital readmissions among diabetic patients compared to other hospitalized patients. This can lead to poor patient outcomes and increased healthcare costs. Predictive machine learning modeling has the potential to offer a solution by identifying patients who have a higher risk for readmission and the risk factors.
The diabetic_data dataset provides an extensive resource for predictive modeling, comprising 101,766 records on 71,518 diabetic patients hospitalized in 130 US hospitals between 1999 and 2008. The dataset consists of 50 features including patient demographics, admission and discharge information, hospitalization duration, emergency and outpatient visits, lab procedure and other procedure counts, medications, readmission records and other medical conditions. For simplicity, in this study each encounter is treated as an input to train the models.

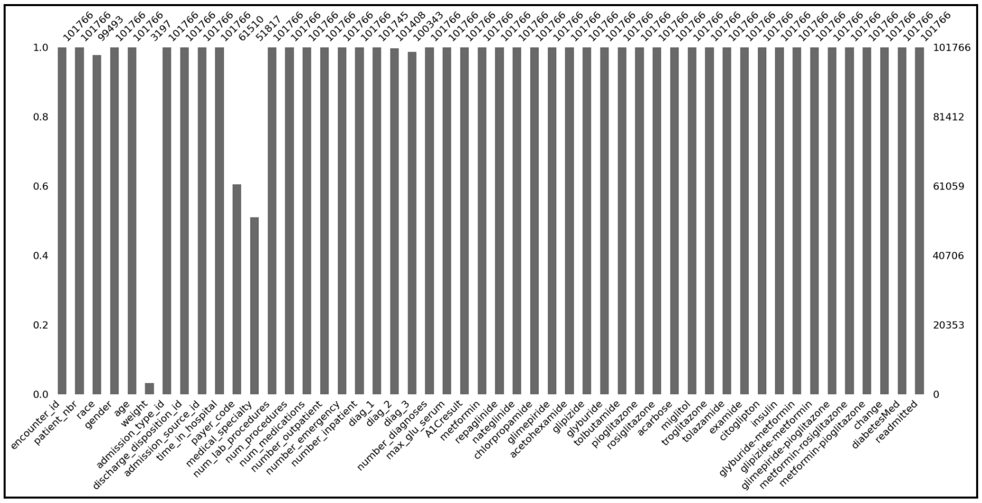
The readmission column, which indicates whether the patient was readmitted within 30 days, more than 30 days or not readmitted, is directly used to construct the independent variable used for predictions. All other features will be processed to determine which will be used to train the prediction models. To develop accurate models, three machine learning models (logistic regression, decision tree, and random forest) will be trained on preprocessed data to predict which patients are more likely to get readmitted and to identify the different variables used in prediction. The models' accuracy and other performance metrics are compared to determine the best approach for predicting readmission rates in diabetic patients. These models will be shown to provide valuable insight into the risk factors associated with patient readmission into hospitals to help mitigate its impact on both patients' outcomes and healthcare expenditure.

Data exploration and description

The dataset contains a mixture of numerical and categorical variables. The majority of patients in the dataset are Caucasian, comprising 73.43% of the sample, with African Americans making up 20.89% of the dataset. Females account for 54% of patients, while males make up 45.98%. The age distribution is right-skewed, with the most dominant age group being 70-80, making up 25.80% of the sample, followed by 21.89% of patients aged 60-70. Regarding readmission, 52.34% of patients were not readmitted to the hospital, while 36.36% were readmitted after more than 30 days, and 11.31% were readmitted in less than 30 days. The majority of patients were admitted from the emergency department and had hospital stays ranging from 2-6 days. Around 88% of patients in the dataset have between 5 and 9 diagnoses and the percentage of patients taking diabetes medication: 77.00%. The dataset does not contain any duplicate entries.

To prepare the dataset for analysis, I performed some cleaning and preprocessing steps to address issues identified during exploratory analysis. These included: handling missing values, removing duplicates, and encoding categorical variables.

- Weight, payer_code, and medical_speciality had 96.86% , 39.56%, 49.08% missing data and thus they were removed from the dataset. Missing values are shown in the figure below.



- There are three features: ‘examide’, 'glimepiride-pioglitazone' and ‘citoglipton’, that have one unique value each, i.e. constant features and they were eliminated from the dataset.
- Removed deceased patients were identified according the following values in the "discharge_disposition_id":
    - 11 Expired
    - 19 Expired at home. Medicaid only, hospice.
    - 20 Expired in a medical facility. Medicaid only, hospice.
    - 21 Expired, place unknown. Medicaid only, hospice.
- Removed id information as they are unnecessary: 'encounter_id', 'patient_nbr'.

- Defined a new target variable to be used in the prediction model that best describes the problem statement: predict the risk of diabetic patients' readmission to hospitals. Define a new coloumn 'binary_readmitted' that contains two values:
    - 1= readmitted (46.09%)
    - 0= Not readmitted (53.91%)
- Removed “Unknown” from the gender coloumn.
- Encoded categorical variables by applying one-hot encoding, using pd.get_dummies() function, to the categorical columns, creating a new dataset that contains both the original numerical features and the newly encoded categorical features.

Preparing data for model training:

- Define two variables: X and y. X comprises the features that will be used to train the model, while y comprises the target variable that the model will predict (‘binary_readmission’).
- Split the data into training and testing sets using train_test_split() function from the scikit-learn library, with 30% of the data used for testing. The random_state parameter is set to a fixed value of 0 to ensure reproducibility of the results.
- Feature scaling was done to normalize the range of values for each feature in the dataset to ensure that the features are on a similar scale and to help prevent any single feature from dominating the model during training. The StandardScaler from the scikit-learn library was used to first fit the training data, this process is done to learn the mean and standard deviation for each feature. Secondly, the scaler is used to apply the scaling on the test data using the learned mean and standard deviation. After scaling, the scaled training and test data is transferred back into Pandas DataFrames to continue with data processing and analyzing using Pandas function.
- Feature selection is a crucial step in machine learning to identify the most relevant features that contribute to predicting the target variable, which can improve model accuracy and performance. In this study, univariate feature selection is used, specifically the ANOVA F-test, to calculate the p-values of each feature. The SelectKBest method from the scikit-learn library is applied, with f_classif as the scoring function and k=10 as the number of features to select. The method is fitted on the training data to select the most important features based on their indices, which are printed out for reference. The training and testing data are transformed to include only the selected features, and then converted back to a pandas DataFrame for further analysis.

The Machine Learning Models

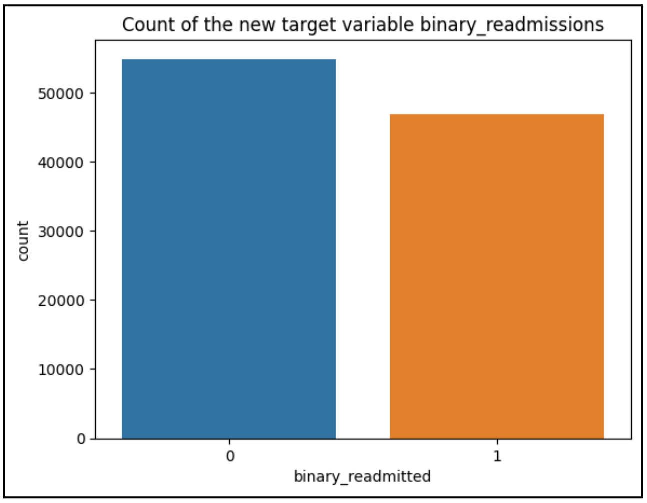
The models developed in this study are expected to provide valuable insights into the risk factors associated with patient readmission, and help healthcare providers mitigate the impact of readmissions on patients’ outcomes and healthcare expenditure. The model is trained to learn the patterns and relationships between the features and the outcome variable.

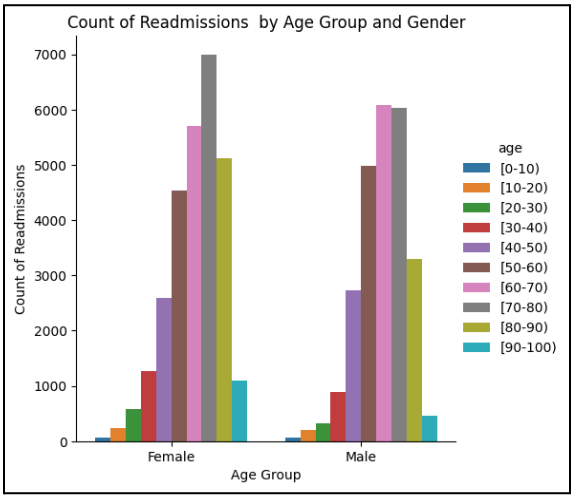
Three multi-class classification machine learning models were chosen to predict readmission of diabetic patients to hospital: logistic regression, decision trees, and random forests models were trained and used to predict the target outcome ‘binary_readmission’. Logistic regression is a linear model that uses a logistic function to model the probability of a binary outcome, such as whether a patient is readmitted or not. Decision trees are a non-linear model that partitions the data into smaller subsets based on the values of the input features, and makes predictions by traversing the tree from the root to the leaf nodes. Random forest is an ensemble model that combines multiple decision trees to improve the accuracy and reduce overfitting.

Each model is first built or instantiated using scikit-learn library. First, the logistic regression model is instantiated using the LogisticRegression() function. The decision tree model is built with hyperparameter tuning using GridSearchCV. The dataset is prepared for modeling, and a parameter grid is defined with different values for max_depth and min_samples_split. GridSearchCV is then used to find the best combination of these hyperparameters for the model. The best estimator for the model is obtained using the best_estimator_ attribute of the GridSearchCV object. Finally, the random forest model is built using the RandomForestClassifier class. The model is initialized with 100 trees and a maximum depth of three.

Each model is then trained on the training set using the fit() method with the predictor variables (X_train) and the target variable (y_train). After the model is trained, the performance of the models is evaluated the predict() method is used to make predictions on the test data (X_test), and the accuracy score and confusion matrix are printed using the accuracy_score() and confusion_matrix() functions from scikit-learn. The accuracy score is calculated as the proportion of correct predictions to total predictions, while the confusion matrix shows the true positive, true negative, false positive, and false negative values for the model's predictions. The classification report provides metrics such as precision, recall, and F1-score for each class (i.e., readmitted and not readmitted). The feature importances are then obtained for each model using the feature_importances_ attribute of the best estimator, and sorted in descending order. These importances provide an indication of which features are most important in making predictions. Cross- Validation of models is performed.

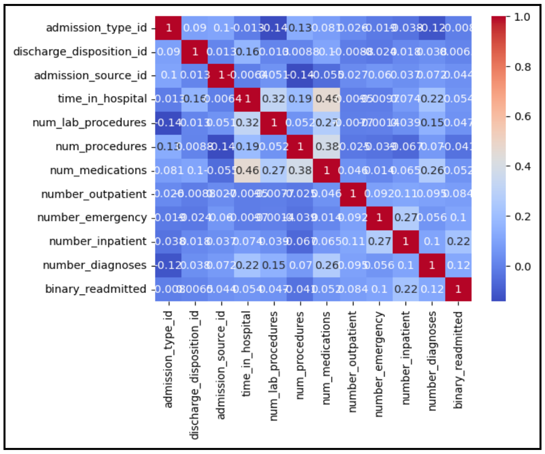
Results and findings





Analyzing the readmission data, it was found that the 50-90 age groups contain the majority of both male and female patients who were readmitted. Moreover, further investigation into the readmission rates for females and males revealed that the readmission rate for females is 47.66%, while for males it is 45.91%. Although females have a slightly higher readmission rate compared to males, the difference between the two is not significant.

Upon analyzing the correlation between the target variable and different features, it was found that the features with the highest correlation to the target variable are number_emergency, number_inpatient, number_diagnoses, and binary_readmitted. All of these features have to do with the medical conditions of patients, the higher correlation values suggest that patients suffering from severe or chronic medical conditions require more frequent hospitalization. These features may be important predictors of patient readmission.



The majority of patients in a dataset being of the Caucasian race could have significant implications for data analysis and modeling. If these results are applied to a more racially diverse group of patients, bias in the analysis may occur, and the findings may not be generalizable to other populations. Furthermore, by focusing on a single racial group, important disparities in health outcomes between different racial and ethnic groups could be overlooked. Drawing policies and interventions by healthcare organizations by using data from one racial group could be futile. For example, not taking into account the differences in access to care or biological factors that affect health outcomes significantly. To address these implications, it may be necessary to ensure that the dataset is representative of the population of interest in terms of racial diversity and to fully understand the disparities between different races and ethnicities to identify areas for targeted interventions and improve health equity

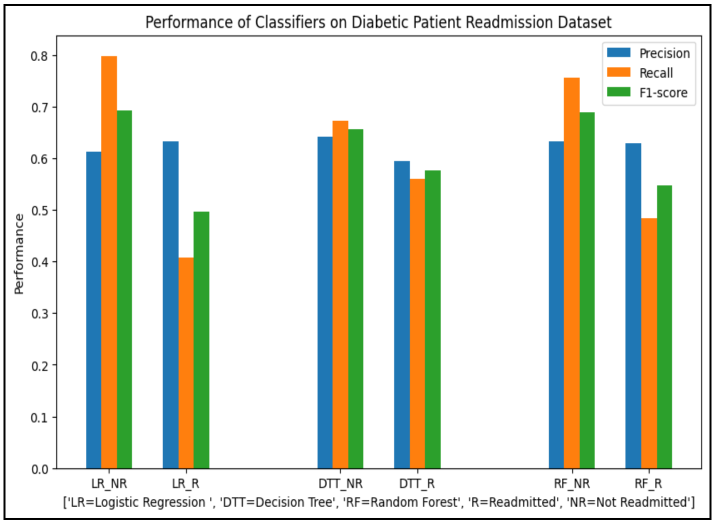
Feature selection method has identified ten features to be used by the training models. The following is a list of the selected features:

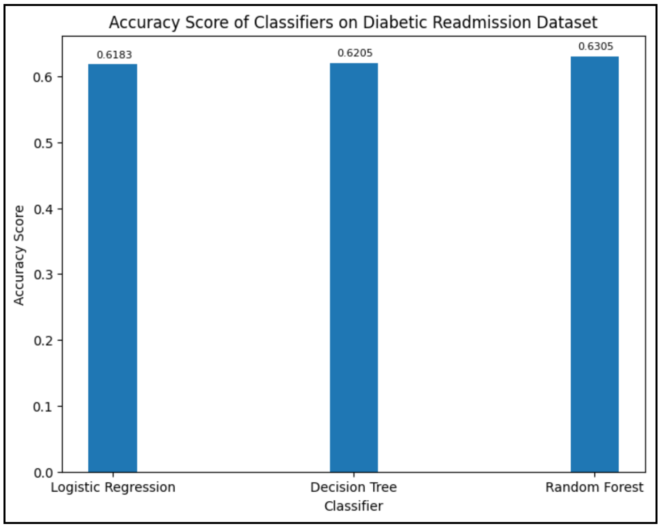
```
Feature 0: time_in_hospital (index position: 3)
Feature 1: number_outpatient (index position: 7)
Feature 2: number_emergency (index position: 8)
Feature 3: number_inpatient (index position: 9)
Feature 4: number_diagnoses (index position: 10)
Feature 5: diag_1_428 (index position: 304)
Feature 6: diag_2_250 (index position: 820)
Feature 7: diag_3_250 (index position: 1576)
Feature 8: diabetesMed_No (index position: 2354)
Feature 9: diabetesMed_Yes (index position: 2355)
```

Precision measures the proportion of true positives among all positive predictions, while recall measures the proportion of true positives among all actual positive cases. Recall is important because it measures the ability of a classifier to correctly identify all positive instances in a dataset. In the case of diabetic readmission, a high recall means that the classifier is good at identifying patients who are likely to be readmitted, which is important for healthcare providers to take appropriate action and prevent readmission. In general, a good model should have a high F1-score, which takes into account both precision and recall.

The following table shows the result of the three different machine learning algorithms: logistic regression, decision tree, and random forest. For each model, there are two values for precision, recall, and F1-score, corresponding to the two classes being predicted. The values are prefixed by 0 for predictions of not readmitted and 1 for predictions of readmitted.

Model	Best 5 Features	Accuracy	Precision	Recall	F1
Logistic Regression	1.time_in_hospital 2.admission_source_id 3.num_lab_procedures 4.discharge_disposition_id 5.num_procedures	61.83%	0:0.6077 1:0.649	0:0.8015 1:0.4157	0:0.6912 1:0.5070
Decision Tree	1.time_in_hospital 2.num_lab_procedures 3.admission_source_id 4.discharge_disposition_id 5.num_procedures	62.05%	0: 0.6244 1:0.6194	0:0.7232 1:0.5087	0:0.67015 1: 0.5586
Random Forest	1. time_in_hospital 2.admission_source_id 3.num_lab_procedures 4.discharge_disposition_id 5.num_procedures	63.05%	0: 0.6141 1: 0.6338	0:0.7667 1:0.4559	0:0.6819 1:0.5303





The table and figures presented above indicate that the random forest model achieved the highest accuracy among the three models for predicting patient readmission. However, in terms of the F1-score, which is a more comprehensive metric that considers both precision and recall, the decision trees model showed the best performance for predicting readmission of patients with an F1-score of 0.56. It's worth noting that all models showed poor performance in this task.

The table presented also shows that the time spent in the hospital is the best feature for predicting hospital readmissions, this feature has important implications for healthcare providers. By identifying the length of hospital stay as the most critical predictor of readmission, healthcare providers can focus on reducing the time patients spend in the hospital. Resolutions could be achieved by prioritizing and implementing strategies such as care coordination between healthcare providers to ensure patients receive timely and appropriate care and developing a comprehensive plan for a patient's care and follow up after leaving the hospital. Moreover, healthcare providers can take measures to control risk factors identified in the study, including monitoring the number of lab procedures, admitting patients from appropriate sources, discharging patients to appropriate settings, and ensuring access to necessary follow-up care.

Conclusions and future work

The findings suggest that the age range of 50-90 years is critical when it comes to predicting patient readmission, and that medical conditions such as emergency visits, inpatient stays, and diagnoses are important predictors of readmission. The results also indicate that there is a slight difference in readmission rates between males and females, but it is not statistically significant.These findings are significant for healthcare providers and policymakers as they highlight the importance of targeting patients with chronic or severe medical conditions, particularly those in the 50-90 year age range, to reduce readmission rates. Additionally, the findings indicate that gender may not be a significant factor in predicting patient readmission, which suggests that healthcare interventions should not be gender-specific.

One limitation of the study is the class imbalance issue that arose when redefining the target variable to only include two groups: patients who were readmitted within 30 days and patients who were not. This imbalance can negatively impact the performance of the classification model, leading to biased predictions, especially in the minority class of patients who were readmitted within 30 days. Therefore, to address this issue, I decided to redefine the target variable to include two groups: patients who were readmitted and patients who were not, which helped to balance the data.

However, the trained models still suffered from poor performance, indicating that more work is required to improve the model's accuracy. To achieve a satisfactory level of performance, further efforts are needed to hypertune the model parameters and try different feature selection techniques to identify the most relevant features for predicting hospital readmission. These additional steps can help optimize the model's performance and increase its predictive power. Overall, addressing the class imbalance issue was an important step, but further work is necessary to improve the performance of the classification model.

The study's finding that time spent in the hospital is the most critical predictor of hospital readmission has significant implications for healthcare providers. Strategies to reduce hospital stays, such as care coordination and comprehensive discharge planning, can help prevent readmissions. The study also identified other important features that can be used to develop predictive models and targeted interventions to prevent readmissions. These findings highlight the importance of identifying high-risk patients and providing appropriate care to improve patient outcomes and reduce healthcare costs.

References:

1. https://www.cdc.gov/diabetes/basics/diabetes.html#:~:text=With%20diabetes%2C%20your%20body%20doesn,vision%20loss%2C%20and%20kidney%20disease.
2. Bilandzic A, Rosella L. The cost of diabetes in Canada over 10 years: applying attributable health care costs to a diabetes incidence prediction model. Health Promot Chronic Dis Prev Can. 2017 Feb;37(2):49-53. doi: 10.24095/hpcdp.37.2.03. PMID: 28273040; PMCID: PMC5607525.
3. Ostling S, Wyckoff J, Ciarkowski SL, Pai CW, Choe HM, Bahl V, Gianchandani R. The relationship between diabetes mellitus and 30-day readmission rates. Clin Diabetes Endocrinol. 2017 Mar 22;3:3. doi: 10.1186/s40842-016-0040-x. PMID: 28702257; PMCID: PMC5472001.
4. https://www.diabetes.ca/media-room/press-releases/new-data-shows-diabetes-rates-and-economic-burden-on-families-continue-to-rise-in-ontario--#:~:text=Nationally%2C%20the%20costs%20of%20treating,urgent%20action%2C%E2%80%9D%20says%20Dr.