Report of early hospital readmission project

A machine learning Navigating the Complexities of Hospital Readmission Among Diabetic Patient

Fatma ben ali , marwa triaa , ons jandoubi , iheb mchichi

Department of Technologies *and Engineering, Esprit*

articleinfo

Keywords:
Machine learning
Logistic regression
Knn
Svm
Descion tree
Acp

abstract

Hospital readmission rates among diabetic patients represent a significant challenge for healthcare systems globally, particularly in the United States, where they serve as crucial quality indicators with substantial financial implications. Addressing this issue requires a multifaceted approach that considers the complex interplay of factors influencing readmissions. These factors include disease management, socioeconomic disparities, care transitions, comorbidities, and behavioral aspects. High readmission rates not only strain healthcare resources but also signify gaps in patient care and contribute to increased healthcare costs. To tackle this issue, healthcare systems must prioritize strategies aimed at enhancing patient education, improving care coordination, leveraging technology for remote monitoring, and fostering community support networks. By implementing these strategies, healthcare systems can mitigate readmission risk, improve patient outcomes, and ultimately enhance the quality of care for diabetic patients while containing healthcare costs.

1. Introduction

Reducing hospital readmission rates among patients with diabetes is a critical imperative in contemporary healthcare, not only for enhancing patient outcomes but also for mitigating the substantial economic burden associated with recurrent hospitalizations. Diabetes, a chronic metabolic disorder characterized by high blood sugar levels, affects millions worldwide and presents a significant challenge for healthcare systems globally. Hospital readmissions, occurring when patients are re-hospitalized within a certain timeframe after discharge, signify a failure in care continuity and often indicate underlying issues in disease management, treatment adherence, or post-discharge support.

The Centers for Medicare & Medicaid Services (CMS) has recognized the importance of addressing readmission rates, implementing initiatives such as the Hospital Readmissions Reduction Program to incentivize hospitals to improve care quality and reduce healthcare expenditures. Although diabetes has yet to be included in penalty measures, the escalating costs associated with diabetic readmissions—exceeding $41 billion in the United States alone in 2011—underscore the urgent need to identify factors contributing to readmission risk and develop predictive models to preemptively intervene. Such endeavors not only hold promise for substantial cost savings but also stand to enhance patient well-being and healthcare delivery effectiveness on a systemic level. Thus, elucidating the multifaceted determinants of diabetic readmissions and harnessing predictive analytics represent pivotal steps toward achieving the dual goals of cost containment and care optimization in the management of diabetes within healthcare systems worldwide

2. Méthodologie of work

Week 1: Data Analysis

Task 1: Data Collection: Gather datasets containing information on diabetic patients, including demographics, medical history, clinical measurements, and previous hospitalization records.

Task 2: Exploratory Data Analysis (EDA): Conduct exploratory data analysis to understand the distribution of variables, identify outliers, and uncover potential patterns or trends related to readmission.

Task 3: Feature Identification: Identify potential features that may influence readmission risk, such as age, gender, comorbidities, medication adherence, and discharge disposition.

Task 4: Correlation Analysis: Investigate correlations between variables to determine which features are most strongly associated with readmission.

Week 2: Data Processing

Task 1: Data Cleaning: Clean the dataset by handling missing values, outliers, and inconsistencies to ensure data quality.

Task 2: Feature Engineering: Create new features or transformations of existing ones that may better capture underlying relationships and improve model performance.

Task 3: Data Normalization/Scaling: Normalize or scale the features to ensure uniformity and enhance the convergence of machine learning algorithms.

Task 4: Data Splitting: Split the dataset into training, validation, and test sets to facilitate model training and evaluation

3. Data analyse

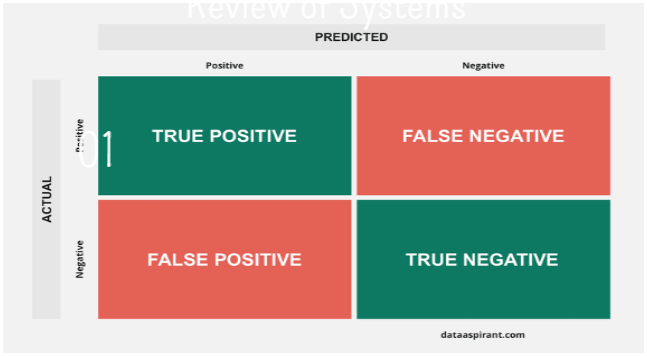
feature	Type of data	Description	%missed values
EncounterID	Numeric	Id of the encounter.	0%
PatientNumber	Numeric	Unique id of the patient	0%
Race	Nominal	Values:Caucasian,Asian,AfricanAmerican,Hispanic,andother	2%
Gender	Nominal	Values:male,female, unknown/invalid	0%
Age	Nominal	Grouped in10-year intervals :[0,10),[10,20),...,[90,100)	0%
Weight	Numeric	Weight in pounds.	97%
Admission type	Nominal	Integer identifier corresponding to 9distinctvalues, for example: emergency, urgent, elective, newborn, and not available	0%
Discharge disposition	Nominal	Integer identifier corresponding to 29 distinct values ,for example ,discharged to home, expired, and not available	0%
Admission source	Nominal	Integer identifier corresponding to 21distinct values,for example, physician referral, emergency room,and transfer from a hospital	0%
Time in hospital	Numeric	Integer number of days between admission and discharge	0%
Payercode	Nominal	Integer identifier corresponding to 23 distinct values, for example, BlueCross\Blue Shield,Medicare ,and self-pay	52%
Medical specialty	Nominal	Integer identifier of aspecialty of the admitting physician, corresponding to 84 distinct values, for example, cardiology, internal medicine, family\general practice, and surgeon	53%
Number of lab procedures	Numeric	Number of labtests performed during the encounter	0%
Number of procedures	Numeric	Number of procedures (other than lab tests)performed during the encounter	0%
Number of medications	Numeric	Number of distinct generic names administered during the encounter	0%
Number of Outpatients Visits	Numeric	Number of Outpatients Visits Of the patient in the year preceding the encounter	0%
Numberof emergency visits	Numeric	Number of emergency Visits Of the patient in the year preceding the encounter	0%
Number of inpatient visits	Numeric	Number of in patient visits of the patient in the year preceding the encounter	0%
Diagnosis_1	Nominal	The primary diagnosis (code das first three digits of ICD9);848 distinctvalues	0%
Diagnosis_2	Nominal	Secondary diagnosis(codedasfirstthreedigitsofICD9);923 distinct values Additional secondary diagnosis(codedasfirstthreedigitsofICD9);954 distinct values	0%
Diagnosis_3	Nominal	Additional secondary diagnosis (codedasfirstthreedigitsofICD9);954distinct values	1%
Number of diagnoses	Numeric	Number of diagnoses entered to the system	0%
Glucose serum test result	Nominal	Indicates the range of the result or if the test was not taken. Values:“>200, ”“>300,” “normal,”and “none” if not measured	0%
A1C test result	Nominal	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.	0%
Change of medications	Nominal	Indicates if there was a change indiabetic medications(either dosage or generic name).Values:“change”and“no change”	0%
Diabetes medications	Nominal	Values:“yes”and“no”	0%
24 features of medication	Nominal	glyburide-metformin,glipizide-metformin,glimepiride-pioglitazone, metformin-rosiglitazone,and metformin-pioglitazone,thefeatureindicateswhether the drug was prescribed or there was a change in the dosage.Values: “up”if the dosage was increased during the encounter,“down”if the dosage was decreased,“steady”if the dosagedid not change,and “no”if the drug was not prescribed	0%
Readmitted	Nominal	Days to inpatient readmission. Values:“<30”if the patient was readmitted in less than 30 days,“>30” if the patient was readmitted in more than 30 days ,and“No”forno recordofreadmission	0%

4.Models used

Logistic Regression :

Similar to linear regression, logistic regression is also used to estimate the relationship between a dependent variable and one or more independent variables, but it is used to

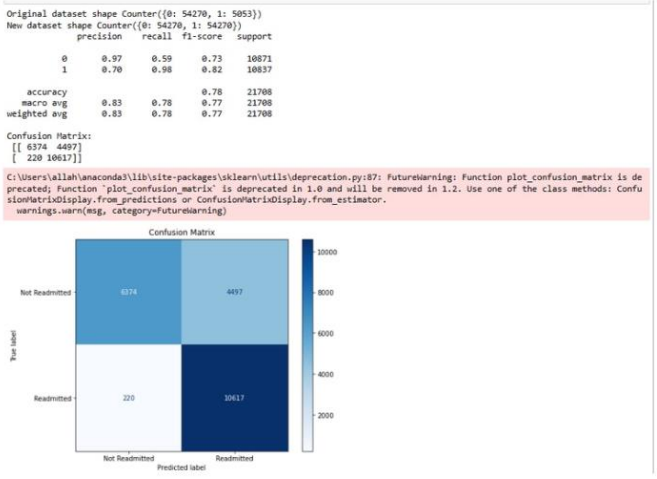
make a prediction about a categorical variable versus a continuous variable. A categorical variable can be true or false, yes or no, 1 or 0, etc.



Prediction **Fig. 1**

KNN (K-Nearest Neighbors) :

It works by finding the K closest samples in the feature space and using majority voting for classification or averaging for regression.It is simple to understand and implement.It can be sensitive to the presence of outliers and high dimensionality



Accuracy KNN**Fig. 2**

SVM (Support Vector Machine) :

It finds a hyperplane in a high-dimensional feature space that optimally separates different classes.It is effective in high-dimensional feature spaces and when the number of dimensions is greater than the number of samples.It can be sensitive to tuning parameters, such as the kernel and the regularization term.

It recursively divides the dataset into homogeneous subsets based on features, choosing at each step the feature that best separates classes or minimizes variance.It is easy to interpret and visualize, making it a good choice for tasks where model transparency is important.It tends to be prone to overfitting, especially with large, high-dimensionality datasets.

Accuracy svm **Fig. 3**

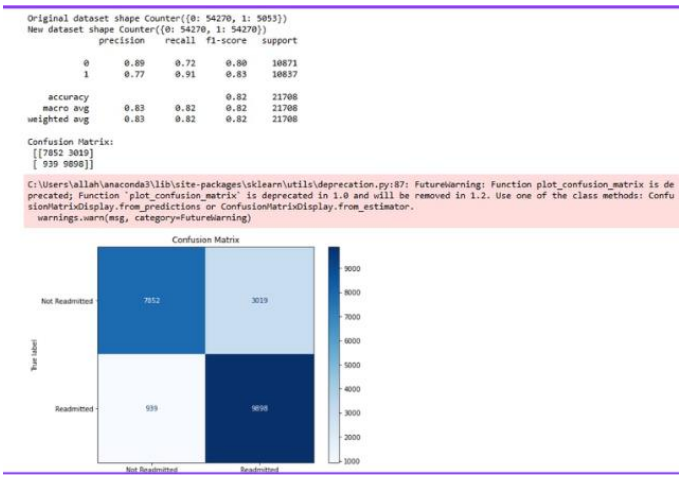
Random forest

Random Forest is a supervised learning algorithm that belongs to the family of ensemble techniques, meaning it combines predictions from multiple learning models to improve performance and accuracy and reduce overfitting.It uses majority voting for predictions and is effective for large datasets with many features.

Accuracy Random Forest **Fig. 4.**

ACP :

it is used to compress data while retaining as much information as possible. By reducing the number of dimensions, PCA simplifies calculations and potentially improves model performance by reducing the effects of the curse of dimensionality.



Accuracy acp **Fig. 5**

Decision Tree :

able to collect 1100 data based on 28 factors. Then we labeled the class of each record of the data set by consulting with some nutritionists and student counselors in educational institutions. Thus, our data set is collected.

Accuracy Descion tree **Fig. 6**

5.Data preparation

To prepare the data, we'll perform several steps to ensure it's clean, structured, and ready for analysis:

. Handling Missing Values :

Identify and handle missing values in each column. Depending on the proportion of missing values and the importance of the feature, you may choose to impute missing values using techniques like mean, median, or mode imputation, or drop rows or columns with a high percentage of missing values.

. Data Cleaning:

Clean the data by addressing any inconsistencies, errors, or outliers. For example, ensure consistency in data formats (e.g., converting categorical variables to the appropriate data type), remove duplicates, and verify the validity of values (e.g., age, weight).

. Feature Selection:

Select relevant features that are likely to influence readmission risk. This may involve dropping irrelevant or redundant columns, such as identifiers (e.g., encounter_id, patient_nbr) or those with low variance or low predictive power.

Handling Imbalanced Data (if applicable):

Check for class imbalance in the target variable (readmitted) and apply appropriate techniques to address it, such as oversampling minority classes (e.g., readmitted patients) or undersampling the majority class.

Data Scaling (if applicable):

Scale numerical features to ensure they have a similar range and distribution. Common scaling techniques include standardization (scaling to have a mean of 0 and a standard deviation of 1) or min-max scaling (scaling to a range between 0 and 1).

Splitting the Dataset:

Split the dataset into training, validation, and test sets. The training set is used to train the model, the validation set is used to tune hyperparameters and assess model performance during training, and the test set is used to evaluate the final model's performance on unseen data.

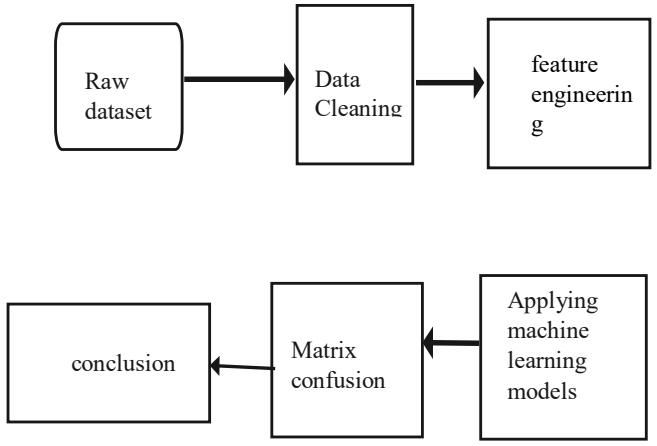
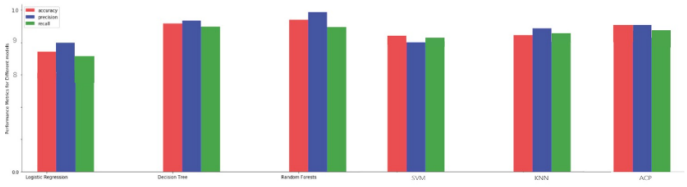


Figure 7 hospital readmission dataset

6.Conclusion

The article highlights the pressing need to address hospital readmission rates among diabetic patients, underscoring the significance of enhancing patient outcomes and mitigating healthcare costs. It emphasizes the multifaceted nature of the challenge, stemming from issues in disease management, treatment adherence, and post-discharge support. To tackle this complex issue, healthcare systems must prioritize a comprehensive approach that includes patient education, care coordination, remote monitoring, and community support networks. The outlined methodology involves leveraging machine learning techniques such as logistic regression, KNN, SVM, Random Forest, and PCA to analyze relevant factors influencing readmission risk and develop predictive models. Through meticulous data analysis, cleaning, and model training, healthcare systems can identify at-risk patients and implement targeted interventions to reduce readmission rates. By adopting proactive strategies and harnessing predictive analytics, healthcare providers can not only improve patient care but also contain healthcare costs, ultimately enhancing the overall



Difference between models used **Fig. 8**

References

Furkannakdagg. (2022, August 2). Data Preparation Tutorial with Diabetes Dataset. Kaggle. https://www.kaggle.com/code/furkannakdagg/data-preparation-tutorial-with-diabetes-dataset

Nailashafiq. (2023, April 21). Predict Diabetes using Random Forest Regression - Nailashafiq - Medium. Medium. https://medium.com/@nailashafiq96/predict-diabetes-using-random-forest-regression-9ad98fe048c3

RPUBS - Diabetes Prediction - Random Forest Model. (n.d.). https://rpubs.com/Nedlin/RF_Model

Dhandhania, K. (2018, July 2). End-to-End Data Science Example: Predicting Diabetes with Logistic Regression. Medium. https://towardsdatascience.com/end-to-end-data-science-example-predicting-diabetes-with-logistic-regression-db9bc88b4d16