Final Report: Predicting Hospital Readmission for Diabetic Patients

Sravya Sri Kodali
George Mason University
G01460802
skodali6@gmu.edu

Pruthvinath Reddy Sireddy
George Mason University
G01458015
psireddy@gmu.edu

Koushik Vasa
George Mason University
G01480627
kvasa@gmu.edu

Abstract

The project explores the application of data mining techniques to identify 30-day readmissions among diabetic patients. By utilizing an extensive dataset that includes patient demographics, clinical characteristics, and healthcare use trends, different data mining techniques are used to find patterns that are not readily evident and to make predictions. This study can be used by medical professionals to proactively intervene and customize treatment plans for diabetes patients who are at risk of readmission.

Introduction

Diabetes is a serious chronic illness that can lead to readmissions. About 10% of Americans suffer from diabetes, which not only carries serious health hazards but also increases the chance of hospital readmissions when compared to people without the illness. When a patient is readmitted within a certain period to the same department for the same illness after being discharged, it is referred to as readmission. A few causes of readmission are an incorrect initial diagnosis and an early discharge.

Hospitals having high readmission rates during short periods of time are subject to penalties from regulatory agencies like the Centers for Medicare. This project focuses on determining whether a diabetic patient will be readmitted within 30 days, after 30 days or not readmitted at all.

Predictive indicators for readmission risk are identified by analyzing patient demographics, clinical characteristics, and historical healthcare utilization patterns using data mining approaches, such as machine learning algorithms. Healthcare professionals can optimize resource allocation, improve patient care outcomes, and reduce readmission rates by precisely estimating the likelihood of readmission.

Data

The dataset of this project is taken from UC Irvine Machine Learning repository. The dataset includes data from 130 US hospitals over a ten-year period from 1999-2008. It consists of 47 features and contains information about 101766 patients.

Each row in the dataset represents the hospital records of a single diabetic patient. The data is specifically chosen to include only inpatient diabetic encounters with stays of 1 to 14 days, where lab tests and medications were a part of the treatment. The features are of type categorical or integer.

Patient data such as race, gender, age, type of admission, duration of stay, lab test numbers, HbA1c findings, diagnosis, and prescription information, including medicines specific to diabetes, are among the dataset's key components. Of the 47 features included, the necessary ones are:

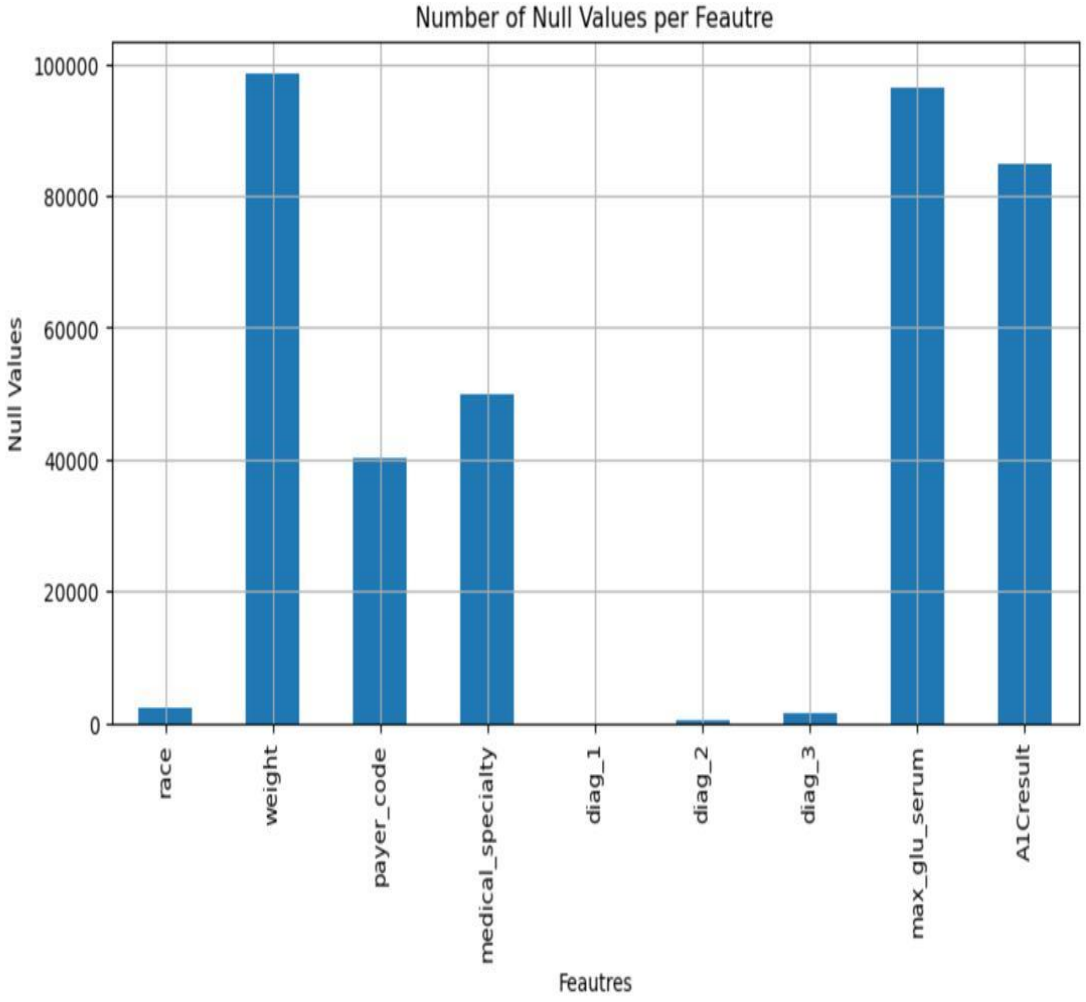
- **Race:** It is a demographic factor which has values like Asian, Hispanic and others.
- **Gender:** Gender is another important demographic factor. Values include male, female and others.
- **Age:** Age is a crucial demographic factor to consider in healthcare studies. Contains intervals [0-10), [10-20), [20-30), [30-40), [40-50), [50-60), [60-70), [70-80), [80-90), [90-100)
- **Admission type:** When admission type is considered in addition to other variables, it becomes clearer how various admission types—such as emergency and elective—might affect the probability of readmission.
- **Discharge disposition:** Discharge disposition refers to where a patient is sent after being discharged from the hospital like discharged to home, expired and others.
- **Admission source:** It includes different sources of admissions like emergency room, transfer from hospital and others.
- **Time in hospital:** This variable is used to represent the number of days between admission and discharge.
- **Number of lab procedures:** This variable is used to represent the number of lab tests performed during the stay in hospital.
- **Number of procedures:** This variable is used to represent the number of procedures (other than lab tests) performed during the stay in hospital.
- **Number of medications:** This variable is used to represent the number of medicines the patient is prescribed during the stay in hospital.
- **Diagnosis 1:** It indicates primary diagnosis.
- **Diagnosis 2:** It indicates secondary diagnosis.
- **Diagnosis 3:** It indicates additional secondary diagnosis.
- **Insulin:** This variable lets us know if the medicine was prescribed or if the dosage was adjusted. Values: no if the drug was not prescribed, stable if the dosage did not change, up if it was increased during the meeting, and down if it was dropped.

- **Change:** This variable indicates if there was a change in diabetic medications. The values include change and no change.
- **Readmitted:** This is the target variable. It indicates whether a patient is readmitted. The values include less than 30, greater than 30 and “no” for no readmission.

Methodology

1. Identifying and Handling Missing Values:

- Conversion of Missing Indicators: The dataset initially contained '?' as a placeholder for missing values in several columns. To standardize the data handling and enable more efficient analysis, these '?' indicators were first converted to NaN values.



2.Data Cleaning and Column Removal

Removal of Irrelevant Columns:

- Reason for Removal: Certain columns such as Encounter ID, Patient Number, and Payer Code were removed as they were deemed not relevant to the predictive modeling. These columns

are typically identifiers that do not contribute predictive power to models focused on medical outcomes.

Dropping Columns with High Null Values:

- Columns Affected: The columns Weight, Max Glu Serum, and A1C Result had a high proportion of missing entries. Such a large amount of missing data can compromise the integrity of predictive models by requiring extensive imputation that may not accurately reflect the population.

Removal of Columns with Low Variance:

- Reason for Removal: Several columns were dropped because they contained mostly the same value for almost every record, which contributes very little information for learning in predictive modeling. Columns removed are 'Medical_specialty', 'Patient_nbr', 'acetohexamide', 'tolbutamide', 'examide', 'citoglipton', 'glyburide-metformin', 'glipizide-metformin','glimepiride-pioglitazone','metformin-rosiglitazone' , 'metformin-pioglitazon'.

3.Imputing Missing Values

Imputation of Continuous Variables:

- Target Columns: The columns representing diagnostic codes (diag_1, diag_2, diag_3), which are considered continuous variables in this context, had missing entries.
- Method Used: To handle missing values in these columns, the KNN (K-Nearest Neighbors) imputation technique was applied. KNN Imputer replaces the missing values using the mean value from 'n_neighbors' nearest neighbors found in the training set. This method is particularly effective for continuous data as it assumes that similar data points exist in close proximity.

Imputation of Categorical Variables:

- Target Column: The race column, which is categorical and had missing values.
- Method Used: For the categorical race column, mode imputation was used. This method fills in the missing values with the most frequently occurring category within the column, which can be a suitable approach for categorical data to maintain the distribution.

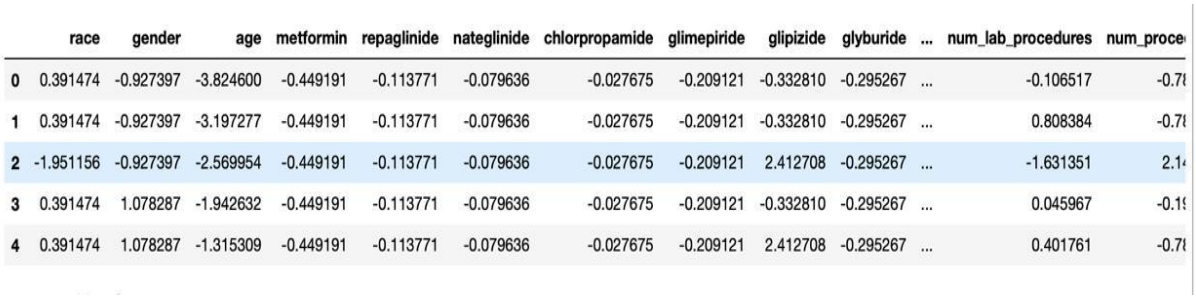
4.Encoding

An Ordinal Encoder was employed to convert categorical variables into numerical formats by assigning a unique integer to each category value, making it easier for machine learning models to interpret and learn from the data.

	race	gender	age	metformin	repaglinide	nateglinide	chlorpropamide	glimepiride	glipizide	glyburide	...	num_lab_procedures	num_procedures	num_m
0	2.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	...	41.0	0.0	
1	2.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	...	59.0	0.0	
2	0.0	0.0	2.0	1.0	1.0	1.0	1.0	1.0	2.0	1.0	...	11.0	5.0	
3	2.0	1.0	3.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	...	44.0	1.0	
4	2.0	1.0	4.0	1.0	1.0	1.0	1.0	1.0	2.0	1.0	...	51.0	0.0	

5.Standardizing Data

StandardScaler was utilized to normalize the features by adjusting each to have a mean of zero and a standard deviation of one. This standardization is crucial for models that rely on the assumption of normally distributed data and helps in equalizing the influence of each feature on the model, facilitating more efficient learning.



	race	gender	age	metformin	repaglinide	nateglinide	chlorpropamide	glimepiride	glipizide	glyburide	...	num_lab_procedures	num_proce
0	0.391474	-0.927397	-3.824600	-0.449191	-0.113771	-0.079636	-0.027675	-0.209121	-0.332810	-0.295267	...	-0.106517	-0.7
1	0.391474	-0.927397	-3.197277	-0.449191	-0.113771	-0.079636	-0.027675	-0.209121	-0.332810	-0.295267	...	0.808384	-0.7
2	-1.951156	-0.927397	-2.569954	-0.449191	-0.113771	-0.079636	-0.027675	-0.209121	2.412708	-0.295267	...	-1.631351	2.1
3	0.391474	1.078287	-1.942632	-0.449191	-0.113771	-0.079636	-0.027675	-0.209121	-0.332810	-0.295267	...	0.045967	-0.1
4	0.391474	1.078287	-1.315309	-0.449191	-0.113771	-0.079636	-0.027675	-0.209121	2.412708	-0.295267	...	0.401761	-0.7

Experiments

While training the model we have focused on comparing the efficacy of machine learning models with and without using principal component analysis to enhance our understanding of dimensional reduction's impact on predictive accuracy. The models tested includes:

- **Decision Tree**: This model was chosen for its ease of learning, and it provided an excellent baseline for understanding the feature interactions without PCA, as well as watching how PCA affected its performance. And we found a jump of accuracy from 44 to 53 percent.
- **Random Forest**: As an ensemble of decision trees, this model is anticipated to outperform the single decision tree baseline by providing stronger generalization capabilities. We compared it with and without PCA to see how it influenced overfitting and variance.
- **Naive Bayes**: Naive Bayes, known for its simplicity and efficiency, was utilized to evaluate the effectiveness of this probabilistic framework, examining how PCA affects the assumption of feature independence.
- **Gradient Boosting**: This model was selected for its high prediction capabilities and adaptability to a variety of data kinds and distributions. We used PCA to observe any performance improvements in handling noise and reducing overfitting.

Each model was trained on a standard dataset split of 80% training and 20% testing to ensure uniformity across evaluations We used PCA to lower the dimensionality of our data before feeding it into the appropriate models, with the intention of looking at the variation in performance measures with smaller feature sets. This approach enabled us to discover the optimal number of primary components that preserved enough information while improving model performance and computational efficiency.

Without PCA **With PCA**

	Precision	Recall	F1-Score	Accuracy
Naive Bayes	0.49	0.54	0.44	0.54
Random Forest	0.55	0.58	0.53	0.58
Decision Tree	0.45	0.44	0.45	0.44
Gradient Boosting	0.50	0.56	0.48	0.55

	Precision	Recall	F1-Score	Accuracy
Naive Bayes	0.59	0.57	0.49	0.57
Random Forest	0.57	0.57	0.57	0.57
Decision Tree	0.53	0.54	0.53	0.53
Gradient Boosting	0.60	0.60	0.60	0.60

To assess the performance of our machine learning models, both with and without the application of PCA, we utilized several key metrics:

- **Accuracy**: This was our key evaluation statistic, indicating the overall accuracy of each model in predicting whether a patient will be readmitted within 30 days (about 4 and a half weeks). Given the objective of our study, high accuracy is critical because it demonstrates the model's overall effectiveness in a real-world clinical scenario.
- **Precision**: This metric helps us determine the accuracy of the positive predictions. Precision is especially crucial in medical predictions since it represents the proportion of positive results in all positive predictions, hence lowering the cost of unneeded treatments or interventions.
- **Recall**: Recall, often known as sensitivity, is a measure of the model's ability to recognize all actual positives. High recall is required in healthcare predictive analytics to ensure that most patients at risk of readmission are accurately identified for appropriate follow-up treatment.
- **F1 Score**: The F1 score, being the harmonic mean of precision and recall, provided an important balance between the two. This is especially useful in our dataset, since there may be an imbalance between the readmitted and non-readmitted groups.

The project's goal was to obtain the highest overall prediction correctness while keeping respectable levels of precision and recall, which led to the emphasis on accuracy as the primary evaluation criterion. However, taking precision, recall, and F1 score into account ensured a balanced approach to model evaluation, understanding the complexity and essential nature of the healthcare outcomes to be predicted.

Related Work:

The challenge of predicting hospital readmissions for diabetic patients has received widespread attention due to its important implications for patient care and healthcare system efficiency. In this section, we examine seminal and recent contributions to the subject, giving context for our study's methodologies and conclusions.

- **Strack et al., 2014**: In one of the pioneering studies in this field, Strack and colleagues created predictive models to predict diabetic patients' chance of hospital readmission within 30 days. Using logistic regression, the study emphasized the importance of clinical and demographic data in predicting readmissions. This study established a standard for accuracy in predicting readmissions, impacting future research in the field.
- **Shah et al., 2015**: This study investigated the use of various machine learning approaches, such as Decision Trees, Support Vector Machines, and Neural Networks, to predict readmission rates in diabetes patients. Shah and his colleagues highlighted the importance of algorithmic complexity and data quality in achieving high prediction accuracy, implying a trade-off between model simplicity and performance.

- **Kaur and Kumari, 2018**: They investigated the effect of reducing feature space on predictive model performance using feature selection strategies. Their findings show that accurate feature selection can considerably increase the performance of models such as Random Forests and Naive Bayes, which we also studied in our project.

Our study expands on past research by comparing the efficacy of multiple machine learning models, both with and without PCA, to determine the most efficient method for predicting hospital readmissions in diabetes patients. By focusing on a comprehensive collection of models and applying dimensionality reduction approaches, we contribute to the continuing discussion in this crucial field of healthcare analytics, with the goal of improving predicted accuracy and, as a result, patient care outcomes.

Conclusion

This project demonstrated the effectiveness of data mining techniques in predicting 30-day readmissions for diabetic patients, which is crucial for improving healthcare management. By analyzing a comprehensive dataset, various machine learning models, including Decision Tree, Random Forest, Naive Bayes, and Gradient Boosting, were evaluated with and without Principal Component Analysis (PCA). The use of PCA generally enhanced model performance, highlighting its utility in reducing dimensionality and mitigating overfitting.

The evaluation metrics—accuracy, precision, recall, and F1 score—highlighted the models' reliability and practicality in clinical settings, aiding in the precise identification of patients at risk of readmission. This not only helps in better patient care but also reduces costs associated with frequent readmissions.

Ultimately, this project underscores the potential of integrating advanced data analytics into healthcare practices to enhance predictive capabilities, suggesting further exploration and continuous methodology refinement for broader application in healthcare.

Division of Work

The division of work was as follows:

• Data Preprocessing: Pruthvinath Reddy Sireddy

• Model Training: Koushik Vasa

• PowerPoint presentation: Pruthvinath Reddy Sireddy, Koushik Vasa, Sravya Sri Kodali

• Report: Sravya Sri Kodali

References:

- Strack, B., DeShazo, J. P., Gennings, C., Olmo, J. L., Ventura, S., Cios, K. J., & Clore, J. N. (2014). Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records. *BioMed Research International*. https://doi.org/10.1155/2014/781670
- Shah, B. R., Hux, J. E., Laupacis, A., Zinman, B., & van Walraven, C. (2005). Clinical inertia in response to inadequate glycemic control: Do specialists differ from primary care physicians? *Diabetes Care*, 28(3), 600-606. https://doi.org/10.2337/diacare.28.3.600
- Kaur, H., & Kumari, V. (2018). Predictive modelling and analytics for diabetes using a machine learning approach. *Applied Computing and Informatics*. https://doi.org/10.1016/j.aci.2018.08.003