**Project Summary:**

|  |  |
| --- | --- |
| **Batch details** | PGP-DSE (PGPDSE-FT Bangalore Jun24) |
| **Team members** | Rahul Yadav, Aswin, Preeti Dhanuka, Bhavitha Sanivarapu, Nishith Ramesh, Lohith T S |
| **Domain of Project** | Health Care |
| **Proposed project title** | Diabetes Care: Predicting Readmission Risks within 30 Days |
| **Group Number** | 2 |
| **Team Leader** | Rahul Yadav |
| **Mentor Name** | Subramanian P V |

Date: 17/11/2024



Signature of the Mentor Signature of the Team Leader

Table of Contents

**Contents**

[**1.** **OVERVIEW:** 3](#_Toc182827917)

[**2.** **Business Problem Statement (GOALS):** 3](#_Toc182827918)

[**3.** **TOPIC SURVEY IN BRIEF:** 4](#_Toc182827919)

[**4.** **CRITICAL ASSESSMENT OF TOPIC SURVEY:** 5](#_Toc182827920)

[**5.** **METHODOLOGY TO BE FOLLOWED:** 5](#_Toc182827921)

[**6.** **REFERENCES:** 10](#_Toc182827922)

[**7.** **APPENDIX:** 11](#_Toc182827923)

**Project Details:**

# **OVERVIEW:**

This project aims to predict early readmission within 30 days of discharge for diabetic patients using clinical data collected from 130 US hospitals over a period of 10 years. The dataset includes patient records with attributes such as age, gender, race, medical history, lab results, medications, and the number of prior hospital visits. Predicting readmission helps hospitals manage patient care more effectively, reduces hospital readmission costs, and improves health outcomes by identifying at-risk patients and intervening early to prevent complications.

# **Business Problem Statement (GOALS):**

1. **Business Problem Understanding:** Diabetes is a chronic condition that requires continuous care. Despite interventions, many patients are readmitted within 30 days after discharge, increasing costs for hospitals and affecting patient health outcomes. Managing and preventing these readmissions requires better predictive models to identify high-risk patients before they are readmitted.
2. **Business Objective:** The objective is to develop a predictive model that accurately identifies patients at risk of readmission within 30 days after discharge, based on their clinical records. This will help hospitals optimize their care processes, reduce readmission rates, and improve patient outcomes.
3. **Approach:** The approach involves analyzing a dataset of diabetic patient records using machine learning techniques to identify patterns and features that correlate with early readmission. The model will use classification techniques (e.g., Logistic Regression, Decision Trees, or Random Forests) to predict readmission risk based on historical data. The results will be validated and tested to ensure reliability.
4. **Conclusions:** By implementing an effective predictive model, hospitals can reduce unnecessary readmissions, optimize resource allocation, and improve patient care. This model can also serve as a tool for healthcare providers to focus on preventive care and intervention strategies.

# **TOPIC SURVEY IN BRIEF:**

1. **Problem Understanding:** Diabetic patients are at a higher risk of complications, often leading to early hospital readmission. Early readmissions significantly contribute to healthcare costs, reduced quality of life, and increased morbidity and mortality. Hospitals need better ways to predict which patients are at risk of being readmitted to focus their interventions on those who need them most.
2. **Current Solution to the Problem:** Currently, hospitals use clinical judgment and possibly, basic predictive models to identify high-risk patients. However, these methods may not be consistent or data-driven enough, leading to arbitrary management of diabetes and insufficient preventive care. In many cases, readmissions are not fully anticipated, which leads to increased hospital costs and worsened patient health outcomes.
3. **Proposed Solution to the Problem:** The proposed solution involves developing a machine learning model that leverages clinical data, such as patient demographics, medical history, lab tests, and prior visits, to predict the likelihood of readmission within 30 days. This data-driven model will be more accurate than traditional methods, enabling hospitals to take proactive steps to manage high-risk patients and reduce unnecessary readmissions.
4. **Reference to the Problem:** Several studies have demonstrated the potential for machine learning in predicting hospital readmissions. However, integrating specific features such as diabetes management, lab results, and medication usage in predicting early readmission has not been fully explored in depth, making this project a valuable contribution to healthcare decision-making.

# **CRITICAL ASSESSMENT OF TOPIC SURVEY:**

1. **Key Area and Gaps Identified:** The critical gap identified in the topic survey is the lack of a personalized, data-driven approach to predicting readmissions for diabetic patients. Most current solutions rely on generic models that do not account for the specific clinical factors affecting diabetes management and readmission risk. This project addresses the gap by focusing on integrating detailed clinical data to create more accurate, individualized predictions.
2. **Key Gaps to Solve:** The key gaps being addressed are the lack of detailed, personalized prediction models for early readmission and the integration of clinical data such as medications, lab results, and prior hospital visits into the predictive model. The project will solve these gaps by using machine learning to provide hospitals with actionable insights into readmission risks based on individual patient profiles.

# **METHODOLOGY TO BE FOLLOWED:**

Fig:5.1 – Methodology to be followed

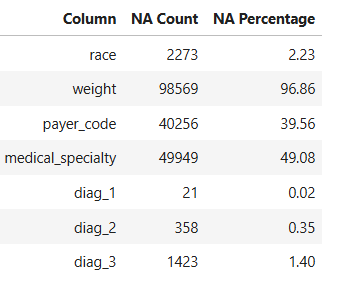
1. **Business Understanding:** Precise prediction of readmission risk can support care-providers to decide if a patient is ready for discharge or should be considered for an intervention program, eventually reducing the number of unscheduled readmissions and curbing healthcare cost.
2. **Data Understanding:** The dataset containing records of diabetic patients from 130 US hospitals over ten years will be used. The data includes patient demographics, lab results, medication history, prior visits, and other medical factors. Detailed description is available in the Data dictionary placed in the subsection section 7.1 of Section 7 Appendix.

There are 101766 rows and 50 columns in the dataset. There are missing values as per the data dictionary.

**Missing values:**

We have missing values in 7 columns denoted by ‘?’ given below:

* 1. race
  2. weight
  3. payer\_code
  4. medical\_specialty
  5. diag\_1
  6. diag\_2
  7. diag\_3



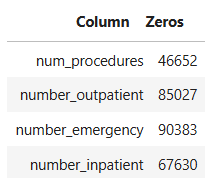
Tab:5.1 – Count & Percentage of Missing values

**Zero values:**

Sparse datasets with high zero values can cause problems like over-fitting in the machine learning models and several other problems. We need to check if zero values are valid or not.

There are zero values in our dataset.

1. num\_procedures
2. number\_outpatient
3. number\_emergency
4. number\_inpatient



Tab:5.2 – Count of Zero values

Zero values in the above columns are not invalid values as they are possible values:

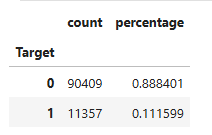
* 1. num\_procedures (Number of procedures (other than lab tests) performed during the encounter)
  2. number\_outpatient (Number of outpatient visits of the patient in the year preceding the encounter)
  3. number\_emergency (Number of emergency visits of the patient in the year preceding the encounter)
  4. number\_inpatient (Number of inpatient visits of the patient in the year preceding the encounter)

**Target variable:**

We shall derive the target variable, 'Target' as follows:

Target value: 0 -- Readmitted NO and Readmitted > 30

Target value: 1 -- Readmitted < 30



Tab:5.3 – Count & Percentage of Classes in the Target variable

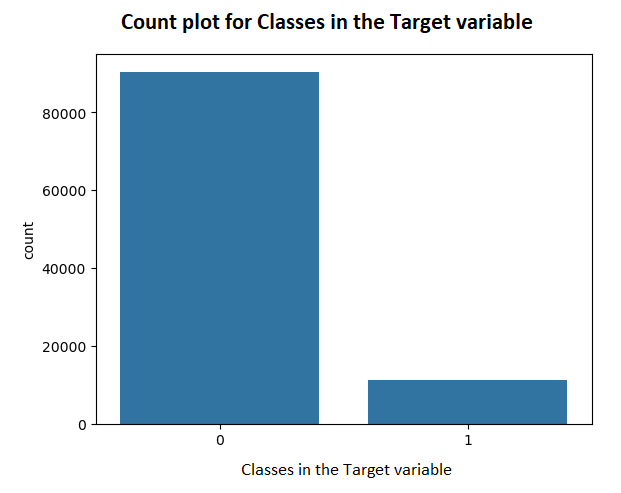


Fig:5.2 – Count plot for classes in the Target variable

1. **Data Preparation:** Data cleaning and preprocessing will be done to handle missing values, outliers, and normalization of numerical features. Categorical features will be encoded appropriately. The data will then be split into training, validation, and test sets for model building.

There are missing values in our dataset as seen in the previous section. The variable, **weight** is having 96.86% of missing values. If we have a column with more than 80% missing values, then it is better to drop the column. The other variables having missing values are: race, payer code, medical specialty, diag\_1, diag\_2, diag\_3 with the percentage of missing values ranging from 0.02 % to 49.08%.

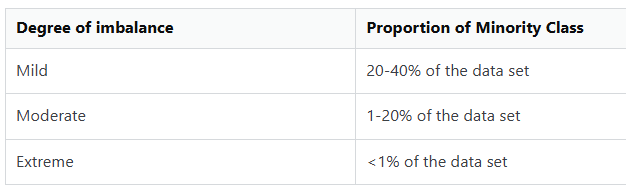
We shall employ a suitable method for treating the missing values in these variables.

**Methods of treating missing values include:**

1. **Remove Rows with Missing Values**: Remove rows that contain missing values.
2. **Impute Missing Values**: Replace missing values with sensible values.
3. **Impute Missing Values with KNN Imputer**: Impute missing values using K nearest neighbors.
4. **Impute Missing Values with Iterative Imputer**: Impute missing values in multiple features using iterative imputation. IterativeImputer in SKlearn. It is a strategy for imputing missing values by modeling each feature with missing values as a function of other features in a round-robin fashion.

**Check if the dataset is not balanced or not**

The challenge of working with imbalanced datasets is that most machine learning techniques will ignore, and in turn have poor performance on, the minority class, although typically it is performance on the minority class that is most important.



Tab:5.4 – Guidelines for Rule of Thumb for the imbalance of classes in the Target variable

We observed the degree of imbalance is **moderate** as the minority class is 11.16%. It is important that we treat the data imbalance. One approach to addressing imbalanced datasets is to oversample the minority class. The simplest approach involves duplicating examples in the minority class, although these examples don’t add any new information to the model. Instead, new examples can be synthesized from the existing examples. This is a type of data augmentation for the minority class and is referred to as the Synthetic Minority Oversampling Technique, or **SMOTE** for short.

**Feature Selection:** Selecting the best features helps the model to perform well. Best features will be selected using techniques[[1]](#footnote-1) mentioned below:

1. Wrapper methods such as Forward selection, Backward selection, Exhaustive Feature selection, Recursive Feature Elimination
2. Filter methods such as Information Gain, Chi-squared test, Fisher's score, Missing value ratio
3. Embedded methods, a combination of both filter and wrapper methods such as Regularization, Random Forest Importance

**Exploratory Data Analysis (EDA):**

We perform EDA to uncover the underlying structure. The structure of the various data sets determines the trends, patterns, and relationships among them.

1. **Modeling:** We shall apply the following models for building the model to predict readmission within 30 days:

1) K-Nearest Neighbours

2) SVM

3) Naive Bayes

4) Decision Tree (CART) Classification

5) Random Forest Classification

6) XGBoost

We shall apply stratified random sampling technique to split our dataset into training and test datasets in the ratio, 80:20. The model will be trained using the training dataset.

1. **Evaluation:** There is imbalance in the dataset, we need to rely on recall (ability of a model to find all the relevant cases within a data set) rather than precision (ability of the mode to identify only the relevant data points) or F1 score, the harmonic-mean of precision and recall. In this application, the important class is "Readmitted". True positives is the number of correctly identified data points of the important class, False positives is the number of data points incorrectly identified as important and False negatives is the number of data points incorrectly identified as not important.

The models built will be tested by using both the training and test datasets to check if the model overfits. If required, cross-validation techniques will be used to ensure the model generalizes well.

Based on the highest Recall for the minority class and ease of interpretability, we shall choose the best model and identify the top six independent variables affecting the target variable to get actionable insights from both EDA and the best model to give recommendations to the stakeholders.

1. **Deployment:** Deployment is out of scope in this project.

# **REFERENCES:**

|  |  |
| --- | --- |
| **Original owner of data** | Beata Strack, Jonathan P. DeShazo, Chris Gennings, Juan L. Olmo, Sebastian Ventura,  John N. Clore, Krzysztof J. Cios. |
| **Data set information** | The dataset contains records from 130 hospitals in the US over ten years (1999-2008), focusing on patients with diabetes. Each row represents the details of a patient who was diagnosed with diabetes, including their hospital tests, medications, and up to 14 days of their stay. |
| **Any past relevant articles using the dataset** | <https://www.mdpi.com/1999-5903/15/9/304> |
| **Reference** | 1*.“Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records.”* BioMed Research International, vol. 2014, Article ID 781670, 11 pages, 2014.  2.Global and Local Interpretable Machine Learning Allow Early Prediction of Unscheduled Hospital Readmission  <https://www.mdpi.com/2504-4990/6/3/80>  3.<https://archive.ics.uci.edu/dataset/296/diabetes+130-us+hospitals+for+years+1999-2008>  4. <https://www.health.harvard.edu/topics/diabetes> |
| **Link to web page** | <https://archive.ics.uci.edu/dataset/296/diabetes+130-us+hospitals+for+years+1999-2008> |

# **APPENDIX:**

**7.1 Data Dictionary:**

| **#** | **Variable Name** | **Role** | **Type** | **Demographic** | **Description** | **Missing Values** |
| --- | --- | --- | --- | --- | --- | --- |
| 1 | encounter\_id | ID |  |  | Unique identifier of an encounter | no |
| 2 | patient\_nbr | ID |  |  | Unique identifier of a patient | no |
| 3 | race | Feature | Categorical | Race | Values: Caucasian, Asian, African American, Hispanic, and other.  **Missing values are denoted by ?** | yes |
| 4 | gender | Feature | Categorical | Gender | Values: male, female, and unknown/invalid | no |
| 5 | age | Feature | Categorical | Age | Grouped in 10-year intervals: [0, 10), [10, 20),..., [90, 100) | no |
| 6 | weight | Feature | Categorical |  | Weight in pounds.  **Missing values are denoted by ?** | yes |
| 7 | admission\_type\_id | Feature | Categorical |  | Integer identifier corresponding to 9 distinct values, for example, emergency, urgent, elective, newborn, and not available | no |
| 8 | discharge\_disposition\_id | Feature | Categorical |  | Integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, and not available | no |
| 9 | admission\_source\_id | Feature | Categorical |  | Integer identifier corresponding to 21 distinct values, for example, physician referral, emergency room, and transfer from a hospital | no |
| 10 | time\_in\_hospital | Feature | Integer |  | Integer number of days between admission and discharge | no |
| 11 | payer\_code | Feature | Categorical |  | Integer identifier corresponding to 23 distinct values, for example, Blue Cross/Blue Shield, Medicare, and self-pay.  **Missing values are denoted by ?** | yes |
| 12 | medical\_specialty | Feature | Categorical |  | Integer identifier of a specialty of the admitting physician, corresponding to 84 distinct values, for example, cardiology, internal medicine, family/general practice, and surgeon.  **Missing values are denoted by ?** | yes |
| 13 | num\_lab\_procedures | Feature | Integer |  | Number of lab tests performed during the encounter | no |
| 14 | num\_procedures | Feature | Integer |  | Number of procedures (other than lab tests) performed during the encounter | no |
| 15 | num\_medications | Feature | Integer |  | Number of distinct generic names administered during the encounter | no |
| 16 | number\_outpatient | Feature | Integer |  | Number of outpatient visits of the patient in the year preceding the encounter | no |
| 17 | number\_emergency | Feature | Integer |  | Number of emergency visits of the patient in the year preceding the encounter | no |
| 18 | number\_inpatient | Feature | Integer |  | Number of inpatient visits of the patient in the year preceding the encounter | no |
| 19 | diag\_1 | Feature | Categorical |  | The primary diagnosis (coded as first three digits of ICD9); 848 distinct values.  **Missing values are denoted by ?** | yes |
| 20 | diag\_2 | Feature | Categorical |  | Secondary diagnosis (coded as first three digits of ICD9); 923 distinct values.  **Missing values are denoted by ?** | yes |
| 21 | diag\_3 | Feature | Categorical |  | Additional secondary diagnosis (coded as first three digits of ICD9); 954 distinct values.  **Missing values are denoted by ?** | yes |
| 22 | number\_diagnoses | Feature | Integer |  | Number of diagnoses entered to the system | no |
| 23 | max\_glu\_serum | Feature | Categorical |  | Indicates the range of the result or if the test was not taken. Values: >200, >300, normal, and none if not measured | no |
| 24 | A1Cresult | Feature | Categorical |  | 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. | no |
| 25 | metformin | Feature | Categorical |  | The feature indicates whether 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 dosage did not change, and no if the drug was not prescribed | no |
| 26 | repaglinide | Feature | Categorical |  | The feature indicates whether 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 dosage did not change, and no if the drug was not prescribed | no |
| 27 | nateglinide | Feature | Categorical |  | The feature indicates whether 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 dosage did not change, and no if the drug was not prescribed | no |
| 28 | chlorpropamide | Feature | Categorical |  | The feature indicates whether 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 dosage did not change, and no if the drug was not prescribed | no |
| 29 | glimepiride | Feature | Categorical |  | The feature indicates whether 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 dosage did not change, and no if the drug was not prescribed | no |
| 30 | acetohexamide | Feature | Categorical |  | The feature indicates whether 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 dosage did not change, and no if the drug was not prescribed | no |
| 31 | glipizide | Feature | Categorical |  | The feature indicates whether 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 dosage did not change, and no if the drug was not prescribed | no |
| 32 | glyburide | Feature | Categorical |  | The feature indicates whether 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 dosage did not change, and no if the drug was not prescribed | no |
| 33 | tolbutamide | Feature | Categorical |  | The feature indicates whether 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 dosage did not change, and no if the drug was not prescribed | no |
| 34 | pioglitazone | Feature | Categorical |  | The feature indicates whether 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 dosage did not change, and no if the drug was not prescribed | no |
| 35 | rosiglitazone | Feature | Categorical |  | The feature indicates whether 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 dosage did not change, and no if the drug was not prescribed | no |
| 36 | acarbose | Feature | Categorical |  | The feature indicates whether 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 dosage did not change, and no if the drug was not prescribed | no |
| 37 | miglitol | Feature | Categorical |  | The feature indicates whether 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 dosage did not change, and no if the drug was not prescribed | no |
| 38 | troglitazone | Feature | Categorical |  | The feature indicates whether 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 dosage did not change, and no if the drug was not prescribed | no |
| 39 | tolazamide | Feature | Categorical |  | The feature indicates whether 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 dosage did not change, and no if the drug was not prescribed | no |
| 40 | examide | Feature | Categorical |  | The feature indicates whether 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 dosage did not change, and no if the drug was not prescribed | no |
| 41 | citoglipton | Feature | Categorical |  | The feature indicates whether 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 dosage did not change, and no if the drug was not prescribed | no |
| 42 | insulin | Feature | Categorical |  | The feature indicates whether 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 dosage did not change, and no if the drug was not prescribed | no |
| 43 | glyburide-metformin | Feature | Categorical |  | The feature indicates whether 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 dosage did not change, and no if the drug was not prescribed | no |
| 44 | glipizide-metformin | Feature | Categorical |  | The feature indicates whether 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 dosage did not change, and no if the drug was not prescribed | no |
| 45 | glimepiride-pioglitazone | Feature | Categorical |  | The feature indicates whether 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 dosage did not change, and no if the drug was not prescribed | no |
| 46 | metformin-rosiglitazone | Feature | Categorical |  | The feature indicates whether 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 dosage did not change, and no if the drug was not prescribed | no |
| 47 | metformin-pioglitazone | Feature | Categorical |  | The feature indicates whether 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 dosage did not change, and no if the drug was not prescribed | no |
| 48 | change | Feature | Categorical |  | Indicates if there was a change in diabetic medications (either dosage or generic name). Values: change and no change | no |
| 49 | iabetesMed | Feature | Categorical |  | Indicates if there was any diabetic medication prescribed. Values: yes and no | no |
| 50 | readmitted | Target | Categorical |  | 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 for no record of readmission. | no |

1. https://www.javatpoint.com/feature-selection-techniques-in-machine-learning [↑](#footnote-ref-1)