Project Summary

Batch details	DSE-GGN JAN 2021
Team members	Mr. Sachin Pathania Mr. Rahul Mamgai Mr. Sushil Mr. Ravi Chaubey Mr. Mohit Sahu
Domain of Project	Health Care Analytics
Proposed project title	Diabetic Patient's Re-admission Prediction
Group Number	4
Team Leader	Mr. Mohit Sahu
Mentor Name	Mr. Sravan Malla

Dataset name: Diabetes 130-US hospitals for years 1999-2008 Data Set

Introduction to the problem/domain/background details:

Diabetes is a long-term condition characterized by hyperglycemia when the pancreas is unable to produce enough insulin or when the body is not able to use the insulin effectively to regulate blood sugar level. The former is known as type 1 diabetes mellitus (T1DM) and the latter as type 2 diabetes mellitus (T2DM). With advances in diagnosis and treatment, lifespan for patients with diabetes mellitus (DM), which commonly includes both types of diabetes, is projected to be longer. Increased lifespan and the high prevalence of obesity worldwide have quadrupled the number of adults living with DM from 108 million in 1980 to 422 million in 2014. Globally, DM accounts for 1.9% of total disability-adjusted life years[5] and approximately 30% of hospitalized adult patients with DM had two or more readmissions within the next calendar year. Complications from diabetes are a serious threat to healthcare systems and also one of the top 10 causes of public hospital readmissions worldwide. In 2017, the hospitalization cost of patients with DM in the USA was $123 billion. Based on a 20% readmission rate, it was estimated that $24.6 billion would be attributed to 30-day readmission. Patients with DM represent one-fifth of the overall 30-day hospital readmissions although some may be preventable through better continuity of care.

Problem Statement:

To predict if the patient with Diabetes will be re-admitted to the hospital by analyzing the incidence and causes of 30-day readmission rates for patients so as to provide more

medical assistance and care to the patient and also reduce the re-admission cost to the health care system while also determining more effective medicines against diabetes.

Business problem/ Impact in business of your problem/Need for this study/Abstract (Executive summary):

Hospital readmission is a high-priority health care quality measure and target for cost reduction. Despite broad interest in readmission, relatively little research has focused on patients with diabetes. The burden of diabetes among hospitalized patients, however, is substantial, growing, and costly, and readmissions contribute a significant portion of this burden. Reducing readmission rates of diabetic patients has the potential to greatly reduce health care costs while simultaneously improving care.

Variable identification:

1) **Independent Variable (49)**: encounter_id, patient_nbr, race, gender, age, weight, admission_type_id, discharge_disposition_id, admission_source_id, time_in_hospital, payer_code, medical_specialty, num_lab_procedures, num_procedures, num_medications, number_outpatient, number_emergency, number_inpatient, diag_1, diag_2, diag_3, number_diagnoses, max_glu_serum, A1Cresult, metformin, repaglinide, nateglinide, chlorpropamide, glimepiride, acetohexamide, glipizide, glyburide, tolbutamide, pioglitazone, rosiglitazone, acarbose, miglitol, troglitazone, tolazamide, examide, citoglipton, insulin, glyburide-metformin, glipizide-metformin, glimepiride-pioglitazone, metformin-rosiglitazone, metformin-pioglitazone, change, diabetesMed.

2) **Target Variable (1):** readmitted (Categorical)

Variable information/Data description:

The data set represents 10 years (1999-2008) of clinical care at 130 US hospitals and integrated delivery networks. It includes over 50 features representing patient and hospital outcomes. Information was extracted from the database for encounters that satisfied the following criteria.

1. It is an inpatient encounter (a hospital admission).
2. It is a diabetic encounter, that is, one during which any kind of diabetes was entered to the system as a diagnosis.
3. The length of stay was at least 1 day and at most 14 days.
4. Laboratory tests were performed during the encounter.
5. Medications were administered during the encounter.

Attributes of the Data set:

Feature name	Type	Description and values	% missing
Encounter ID	Numeric	Unique identifier of an encounter	0%
Patient number	Numeric	Unique identifier of a patient	0%
Race	Nominal	Values: Caucasian, Asian, African American, Hispanic, and other	2%
Gender	Nominal	Values: male, female, and unknown/invalid	0%
Age	Nominal	Grouped in 10-year intervals: [0, 10), [10, 20), . . ., [90, 100)	0%
Weight	Numeric	Weight in pounds.	97%
Admission type	Nominal	Integer identifier corresponding to 9 distinct values, 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 21 distinct 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%
Payer code	Nominal	Integer identifier corresponding to 23 distinct values, for example, Blue Cross\Blue Shield, Medicare, and self-pay	52%
Medical specialty	Nominal	Integer identifier of a specialty 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 lab tests 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 outpatient visits	Numeric	Number of outpatient visits of the patient in the year preceding the encounter	0%
Number of emergency visits	Numeric	Number of emergency visits of the patient in the year preceding the encounter	0%
Number of inpatient visits	Numeric	Number of inpatient visits of the patient in the year preceding the encounter	0%
Diagnosis 1	Nominal	The primary diagnosis (coded as first three digits of ICD9); 848 distinct values	0%
Diagnosis 2	Nominal	Secondary diagnosis (coded as first three digits of ICD9); 923 distinct values	0%
Diagnosis 3	Nominal	Additional secondary diagnosis (coded as first three digits of ICD9); 954 distinct 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 in diabetic medications (either dosage or generic name). Values: "change" and "no change"	0%
Diabetes medications	Nominal	Indicates if there was any diabetic medication prescribed. Values: "yes" and "no"	0%
24 features for medications	Nominal	For the generic names: metformin, repaglinide, nateglinide, chlorpropamide, glimepiride, acetohexamide, glipizide, glyburide, tolbutamide, pioglitazone, rosiglitazone, acarbose, miglitol, troglitazone, tolazamide, examide, sitagliptin, insulin, glyburide-metformin, glipizide-metformin, glimepiride-pioglitazone, metformin-rosiglitazone, and metformin-pioglitazone, 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	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" for no record of readmission.	0%

Data Size:

- Number of Columns: 50
- Total Number of Records: 101766

Future Work/Methodology (Details of algorithms):

1. **DATA ANALYSIS, CLEANING/ PREPROCESSING:** The pre-processing of the dataset before performing ML functions involves the following:
    a. **Descriptive Analysis:** Descriptive statistics are used to describe the basic features of the data in a study. They provide simple summaries about the sample and the measures. Together with simple graphics analysis, they form the basis of virtually every quantitative analysis of data. Measures of variability help communicate the spread of distribution by describing the shape and spread of the data set.
    b. **Inferential Analysis:** Validating the inferences which are found with the help of descriptive analysis (Graphs) with the help of respective statistical tests if needed.
    c. **Treating Outliers:** Checking and analyzing for presence of Outliers in the numerical columns and treating those outliers using the IQR method or any relevant method.
    d. **Treating Missing Values:** Null values in the variables must be treated with suitable methods.
    e. **Encoding Categorical Variables:** Since, machine learning models are based on Mathematical equations and we can intuitively understand that it would cause some problem if we can either keep the Categorical data by encoding the categorical variable or we can drop by checking whether we need the variable for further modelling process because we would only want numbers in the equations.
    f. **Dropping Unnecessary Columns:** We are removing the columns which do not contribute to the model building or the columns which are of less, or of no importance

2. **Exploratory Data Analysis:**
    Exploratory Data Analysis or (EDA) is understanding the data sets by summarizing their main characteristics often plotting them visually. This step is very important especially when we arrive at modeling the data in order to apply Machine learning. Plotting in EDA consists of Bar plot, Box plot, Scatter plot and many more using Univariate, Bivariate and Multivariate Analysis.

3. Data Preparation:

a. **Scaling:** It helps to normalize the data within a particular range and as well as in speeding up the calculations in an algorithm.
b. **Train and Test Split of the Data:** The data is split into train and test in required ratio.

4. Model Building:

We will try to fit/train and test with below ML models and compare the performances.

1. Logistic Regression
2. Naive Bayes
3. KNN Classifier
4. Decision Tree
5. Random Forest
6. Bagging Classifier
7. Boosting Classifier

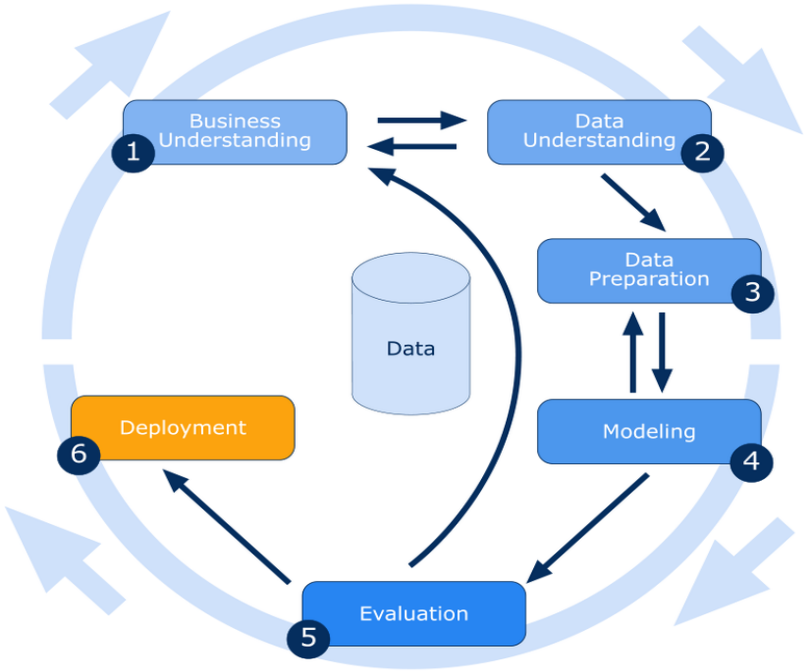
5. Model Evaluation:

Model Evaluation: Below metrics are used to evaluate the multi classification models performance.

1. Accuracy
2. Precision
3. Recall
4. F1-score
5. Confusion Matrix
6. RoC/AuC Score

6. Model Deployment:

In this step, we will save the best model(pickle) and come up with a method or function which takes patient data as input and re-admission status as output. We can try to productionize the deployment using flask.



Conclusions:

We will be able to generate a detailed model which will help us to predict if the Diabetic patient will be re-admitted within 30 days.

Timeline chart (Weekly plan):

First Week	Data Analysis, Pre-processing, EDA
Second Week	Interim Report, Data preparation, Model building
Third Week	Model building
Forth Week	Model building, Model Evaluation
Fifth Week	Model deployment, Final report and Final Presentation

References (Data set source/Journals/articles):

https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008

https://www.kaggle.com/brandao/diabetes

https://link.springer.com/article/10.1007/s11892-018-0989-1

https://clindiabetesendo.biomedcentral.com/articles/10.1186/s40842-016-0040-x

https://drc.bmj.com/content/8/1/e001227

https://diabetes.diabetesjournals.org/content/67/Supplement_1/147-LB

Declaration: This is to declare that the dataset that we are using for our capstone project does not have any relevant legality associated to it and can be used to showcase the work we do on it as a presentation in Great Learning.