# **Diabetes Prediction Using Machine Learning**

Bellevue University

Applied Data Science

Madhukar Ayachit

DSC680 - Summer 2022

# The Domain and Context

Healthcare domain. Non communicable diseases (NCDs) kill around 7 out of 10 people worldwide [[1]](#footnote-1). Type 2 Diabetes Mellitus is a NCD which has impacted about 1 in 10 Americans[[2]](#footnote-2) and globally 6.28% of the world’s population[[3]](#footnote-3). NCDs can be managed with directed interventions provided that the sample from the population can be proactively identified. With increase in blood reports and data available thereof, we have certain datasets which can help us proactively predict Type 2 Diabetes Mellitus (hereinafter referred as Diabetes) using classification techniques.

# Data

I investigated two primary datasets:

1. [Pima Indians Diabetes Database](https://www.kaggle.com/uciml/pima-indians-diabetes-database). (**reference dataset**)
2. Diabetes readmission data from Kaggle (**Main Dataset)**

However, Pima Indian is being used extensively to exercise predictive modeling and plenty of references are available in academic forums. I would rather go with “**readmission**” prediction to help hospitals.

**Diabetes Readmission Data** from Kaggle has data 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.

* No readmission.
* A readmission in less than 30 days
* A readmission in more than 30 days

# The data contains attributes as patient number, race, gender, age, admission type, time in hospital, medical specialty of admitting physician, number of lab test performed, HbA1c test result, diagnosis, number of medications, diabetic medications, number of outpatient, inpatient, and emergency visits in the year before the hospitalization, etc.

# Research Questions

Prediction model can help us understand the trend of patients and their treatment over time. This analysis can help the hospitals understand the trends in data develop strategies to proactively identify the patients who need attention in a way that Diabetes can be managed in a better way.

Research question for my work is to predict early admission of patients so that the hospital can use these predictions to identify patients during their first visit.

**Ethical Implication**

The data are submitted on behalf of the Center for Clinical and Translational Research, Virginia Commonwealth University, a recipient of NIH CTSA grant UL1 TR00058 and a recipient of the CERNER data. John Clore (jclore '@' vcu.edu), Krzysztof J. Cios (kcios '@' vcu.edu), Jon DeShazo (jpdeshazo '@' vcu.edu), and Beata Strack (strackb '@' vcu.edu). This data is a de-identified abstract of the Health Facts database (Cerner Corporation, Kansas City, MO).

The way data is formatted and presented, it maintains anonymity of patience and can be used only as sample dataset for academic or research purpose.

# Method

Solution will be developed in Python using following popular packages/modules:

1. Pandas
2. Numpy
3. Seaborn (visualization)
4. Matplotlib (visualization)
5. Sklearn (Logistic Regression)

* **Criteria**

However, original dataset classifies data in three readmission categories, but I will be making it a case of binary classification to identify readmission.

1. No or <30 = 0
2. >30. = 1

* **Exploration & Visualization**

I will be going over basic structure of the data and all the features to get sense of data. I will be probably using simple bar charts to understand all the features and distribution of the data.

* **Data Cleanup**

Duplicate and null values are primary target, and I may decide to replace it with mean values instead of completely deleting the record.

* **Modeling**

I will be following standard practices of data **Standardization** splitting data into **Training/Testing** set as well as looking at **z-score** for **Feature Selection**. I will consider **Logical Regression** with **Random Forest.**

I will arrive at final recommendation after evaluating score for each model.

**Questions from a potential reviewer**

1. Why did you choose this domain?

I have been working in healthcare for some time. This project can help me to build my portfolio to reflect my understanding on subject as well as on health care domain.

1. What is your data source?

I referred existing Kaggle data source. I have mentioned link in my reference section.

1. What is the purpose or business problem you are addressing?

Patient readmission is biggest cost for hospitals as well reputation. Purpose of this project is to propose a predictive model to identify probable readmissions.

1. What are the steps you took to arrive at conclusion?

I followed typical data science cycle from data collection to , exploration, cleanup and arriving at modeling after 80-20 split of sample data into training and testing model. I am relying on Logistic regression due to binary outcome of the target variable.

1. What is your modeling and why?

As mentioned in point #4 Since target variable is binary. I have decided o go for logistic regression.

1. What are the key modules/Libraries being used?

Following key libraries are used in implementation:

1. Pandas to handle dataset
2. Seaborn and matplotlib for graphs and visuals
3. NumPy
4. LogisticRegression for regression
5. MinMaxScaler for scaling
6. StandardScaler for standardization
7. Any challenge while importing libraries?

Few syntactical issues with MinMaxScaler and It took a while to decide on feature selection.

1. What are the key dependent variables and independent variables?

* time\_in\_hospital
* num\_lab\_procedures
* num\_procedures
* num\_medications
* number\_outpatient
* number\_emergency
* number\_inpatient
* number\_diagnoses
* Outcome (Target)

1. Are use using any visual or graph?

Yes, Data exploration, confusion matrix and ROC curve have graphical representation to explain the data/result

1. How do you summarize the result or outcome of your project?

We could have used more regression but Logistic being my primary choice gave close to 62% accuracy. I might go for few more round of model selection to correct it but with existing size of the data , I am satisfied with result.

**References:**

<https://www.kaggle.com/datasets/brandao/diabetes?select=diabetic_data.csv>

<https://towardsdatascience.com/end-to-end-data-science-example-predicting-diabetes-with-logistic-regression-db9bc88b4d16>

<https://www.kaggle.com/datasets/brandao/diabetes?select=diabetic_data.csv>

1. Deaths due to NCDs worldwide <https://www.cdc.gov/globalhealth/healthprotection/ncd/global-ncd-overview.html> [↑](#footnote-ref-1)
2. Type 2 Diabetes stats as per CDC <https://www.cdc.gov/diabetes/basics/type2.html> [↑](#footnote-ref-2)
3. Type 2 Diabetes statistics globally <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7310804/> [↑](#footnote-ref-3)