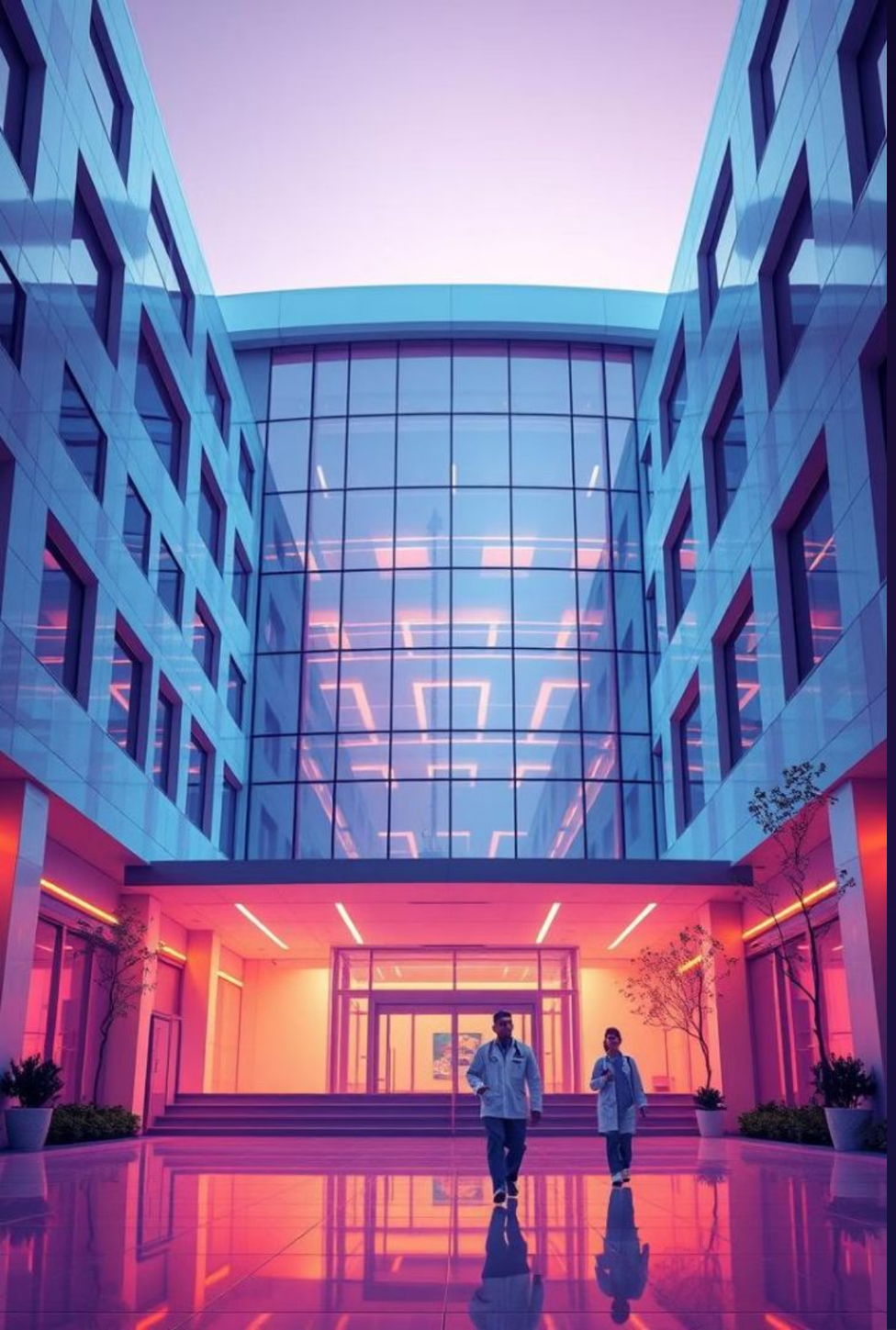

Hospital Readmission Prediction for Diabetic Patients

This project aims to help hospitals reduce readmission rates of diabetic patients by predicting which patients are most at risk of being readmitted to the hospital in 30 days, ultimately improving patient care and resource management.

Project Team

Jaimin Babaria

jbaba5@uic.edu

Vishak Baddur Sadanand

vbadd@uic.edu

Rujuta Avinash Tambewagh

rtamb@uic.edu

Simran Mishra

smish46@uic.edu

Sudha Sree Yerramsetty

syerr7@uic.edu

Varsha Balaji

vbala7@uic.edu

Github classroom - https://github.com/cs418-fa24/project-check-in-team-7

Problem Statement

1

Focus

Hospital readmissions and analyzing risk factors.

2

Key Issue:

Hospital readmissions disrupt operations and impact patient care.

3

Objective

Use machine learning to help hospitals make data-driven decisions to reduce readmissions.

4

Hypothesis

Factors such as medical history, length of stay, and chronic conditions, contribute significantly to readmission rates.

Data

1 Source

Data sourced from VCU Center for Clinical and Translational Research

Reference: https://doi.org/10.1155/2014/781670

2 Details

Database contains data systematically collected from participating institutions electronic medical records and includes encounter data (emergency, outpatient, and inpatient), provider specialty, demographics (age, sex, and race), etc.

3 Data Types

Numerical and categorical(e.g., age, medical history, medications).

4 Feasibility

Ready access to data ensures timely completion of the project.

Sample Data

```
df = pd.read_csv("/Users/rujutatambewagh/Desktop/418-ds/project/diabetic_data_initial.csv")
```

✓ 0.3s Python

```
df.head()
```

✓ 0.2s Python

	encounter_id	patient_nbr	race	gender	age	weight	admission_type_id	discharge_disposition_id	admission_source_id	time_in_hospital	...	citoglipton	insulin	glyburide-metformin	glipizide-metformin	glime piogl
0	2278392	8222157	Caucasian	Female	[0-10)	?	6	25	1	1	...	No	No	No	No	
1	149190	55629189	Caucasian	Female	[10-20)	?	1	1	7	3	...	No	Up	No	No	
2	64410	86047875	AfricanAmerican	Female	[20-30)	?	1	1	7	2	...	No	No	No	No	
3	500364	82442376	Caucasian	Male	[30-40)	?	1	1	7	2	...	No	Up	No	No	
4	16680	42519267	Caucasian	Male	[40-50)	?	1	1	7	1	...	No	Steady	No	No	

5 rows x 50 columns

```
df.info()
```

✓ 0.1s

```
Output exceeds the size limit. Open the full output data in a text editor
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 101766 entries, 0 to 101765
Data columns (total 50 columns):
 #   Column                    Non-Null Count   Dtype
---  ------                    --------------   -----
 0   encounter_id              101766 non-null  int64
 1   patient_nbr               101766 non-null  int64
 2   race                      101766 non-null  object
 3   gender                    101766 non-null  object
 4   age                       101766 non-null  object
 5   weight                    101766 non-null  object
 6   admission_type_id         101766 non-null  int64
 7   discharge_disposition_id  101766 non-null  int64
 8   admission_source_id       101766 non-null  int64
 9   time_in_hospital          101766 non-null  int64
 10  payer_code                101766 non-null  object
 11  medical_specialty         101766 non-null  object
 12  num_lab_procedures        101766 non-null  int64
 13  num_procedures            101766 non-null  int64
 14  num_medications           101766 non-null  int64
 15  number_outpatient         101766 non-null  int64
 16  number_emergency          101766 non-null  int64
 17  number_inpatient          101766 non-null  int64
 18  diag_1                    101766 non-null  object
 19  diag_2                    101766 non-null  object
...
 48  diabetesMed               101766 non-null  object
 49  readmitted                101766 non-null  object
dtypes: int64(13), object(37)
memory usage: 38.8+ MB
```

Solution Approach

1

Data Pipeline

Implement data preprocessing, focusing on data cleaning, imputation of missing values, and transforming categorical variables into actionable insights.

2

Techniques

Correlation Analysis, Principal Component Analysis to name a few.

3

Models

We start with Logistic Regression, explore advanced options like Decision Trees and other models.

4

Future Potential

If time permits, build an Interactive Dashboard for data visualization and decision.