

# Diabetes Readmission Rate



By : Joey Everette, Colton Hammond, Ian Young

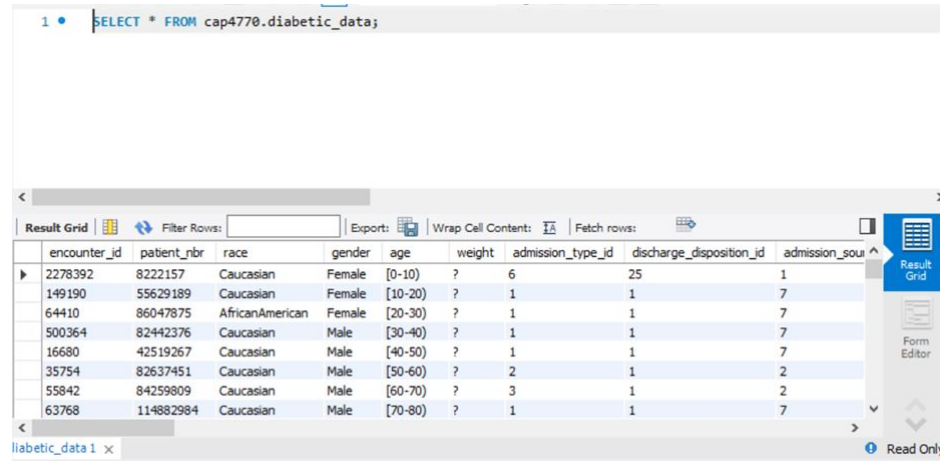
# Problem Statement



A big problem for people with diabetes is after getting their initial diagnosis they will get readmitted back to the hospital which will end up costing families large sums of money. In this project, we are aiming to predict the likelihood of hospital readmission within 30 days for diabetic patients using the "Diabetes 130-US hospitals" dataset. By identifying key features such as age, gender, and race that influence early readmissions.

# Dataset

The dataset that we found in order to figure out our problem would be the diabetes 130-US Hospitals for Years 1999-2008 that can be found on the UC Irvine machine repository website I will post the url to this specific dataset at the end of this slide. This dataset made it easy for us to be able to load it into the MySQL database which is then easy to then integrate in the jupyter notebook. We used MySQL Workbench to create a database named cap4770 and imported the diabetic\_data.csv file using the built-in Table Data Import Wizard. Once the data was loaded, we connected MySQL to Jupyter Notebook using the SQLAlchemy and Mysql-connector-python libraries. This allowed us to query and load the dataset into a pandas DataFrame for preprocessing, exploration, and modeling



The screenshot shows the MySQL Workbench interface. At the top, a SQL query is entered in the editor: `SELECT * FROM cap4770.diabetic_data;`. Below the editor, the 'Result Grid' tab is active, displaying a table with 9 columns and 10 rows of data. The columns are: encounter\_id, patient\_nbr, race, gender, age, weight, admission\_type\_id, discharge\_disposition\_id, and admission\_source\_id. The data rows show various patient records with their demographic and admission details.

encounter_id	patient_nbr	race	gender	age	weight	admission_type_id	discharge_disposition_id	admission_source_id
2278392	8222157	Caucasian	Female	[0-10]	?	6	25	1
149190	55629189	Caucasian	Female	[10-20]	?	1	1	7
64410	86047875	AfricanAmerican	Female	[20-30]	?	1	1	7
500364	82442376	Caucasian	Male	[30-40]	?	1	1	7
16680	42519267	Caucasian	Male	[40-50]	?	1	1	7
35754	82637451	Caucasian	Male	[50-60]	?	2	1	2
55842	84259809	Caucasian	Male	[60-70]	?	3	1	2
63768	114882984	Caucasian	Male	[70-80]	?	1	1	7

The URL for the dataset:  
<https://archive.ics.uci.edu/dataset/296/diabetes+130-us+hospitals+for+years+1999-2008>

# Data Preprocessing

```
import pandas as pd
import numpy as np
from sqlalchemy import create_engine
from sklearn.preprocessing import LabelEncoder

# MySQL connection
engine = create_engine("mysql+mysqlconnector://root:Joseph1985!!@localhost/cap4770")

# Load data
query = "SELECT * FROM diabetic_data"
df = pd.read_sql(query, con=engine)

# Replace '?' with NaN
df.replace('?', np.nan, inplace=True)

# Drop unneeded columns
df.drop(columns=['weight', 'payer_code', 'medical_specialty', 'encounter_id', 'patient_nbr'], inplace=True)

# Drop rows with missing race, gender, or readmitted
df.dropna(subset=['race', 'gender', 'readmitted'], inplace=True)

# Create binary target
df['readmit_30'] = df['readmitted'].apply(lambda x: 1 if x == '<30' else 0)
df.drop(columns=['readmitted'], inplace=True)

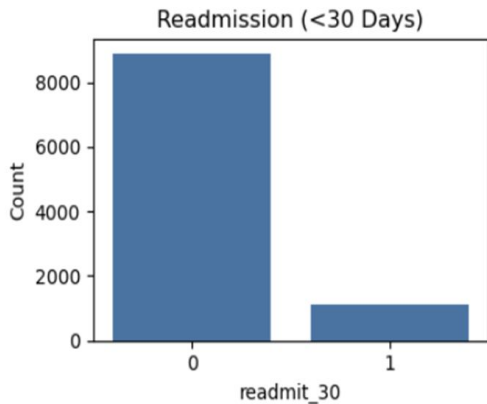
# Label encode categoricals
cat_cols = df.select_dtypes(include='object').columns
encoders = {}

for col in cat_cols:
    le = LabelEncoder()
    df[col] = le.fit_transform(df[col])
    encoders[col] = le
```

We went about the data preprocessing by using various cleaning methods so the dataset was more streamline and accurate for our problem statement. The steps in the right were taken so that we could take out the rows that were not needed for our target variables and some were also taken out that did not have important values

# Exploring Data

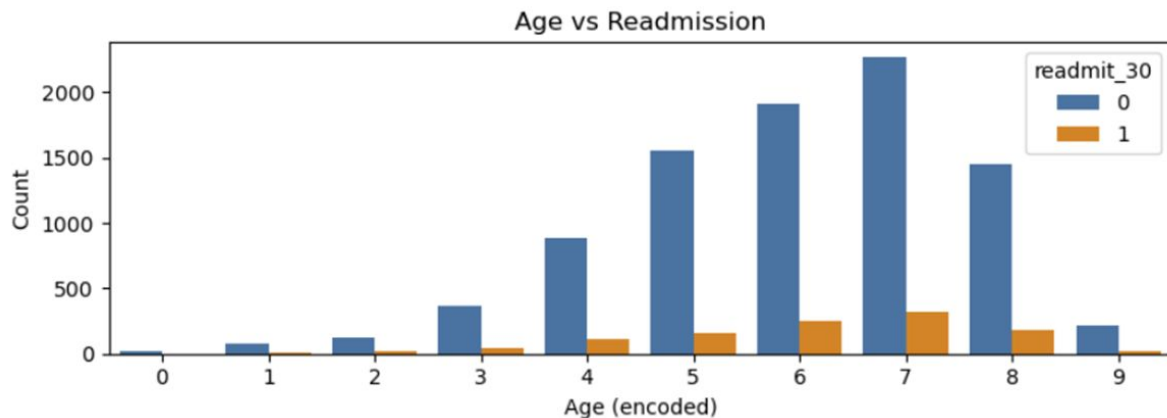
To better understand the structure and distribution of the data, several exploratory data analysis (EDA) techniques were applied. This helped us identify feature distributions, target class imbalance, and early relationships between variables and readmission outcomes. A random sample of 10,000 records was used to make visualizations more efficient while preserving overall patterns. A count plot of the readmit\_30 variable revealed a class imbalance, with the majority of patients not readmitted within 30 days.



This graphic will show that most patients will not be readmitted within the 30 days of being diagnosed

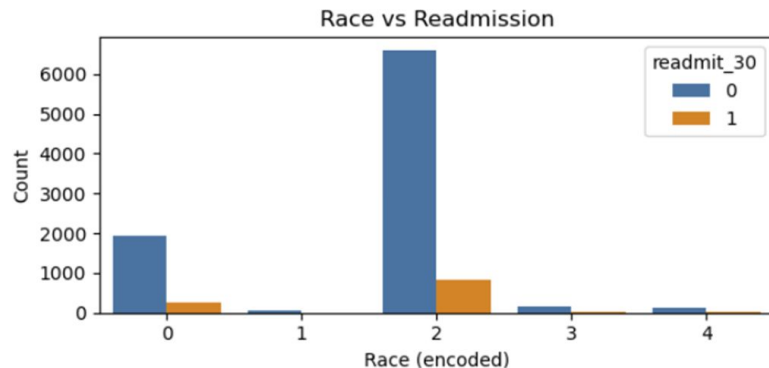
# Exploring Data

There was an interesting correlation when comparing the age against the readmission outcome. Patients in older age brackets (e.g., [70-80), [60-70)) had slightly higher rates of readmission, though all age groups leaned heavily toward non-readmission. The graphic below shows this data trend about the age.



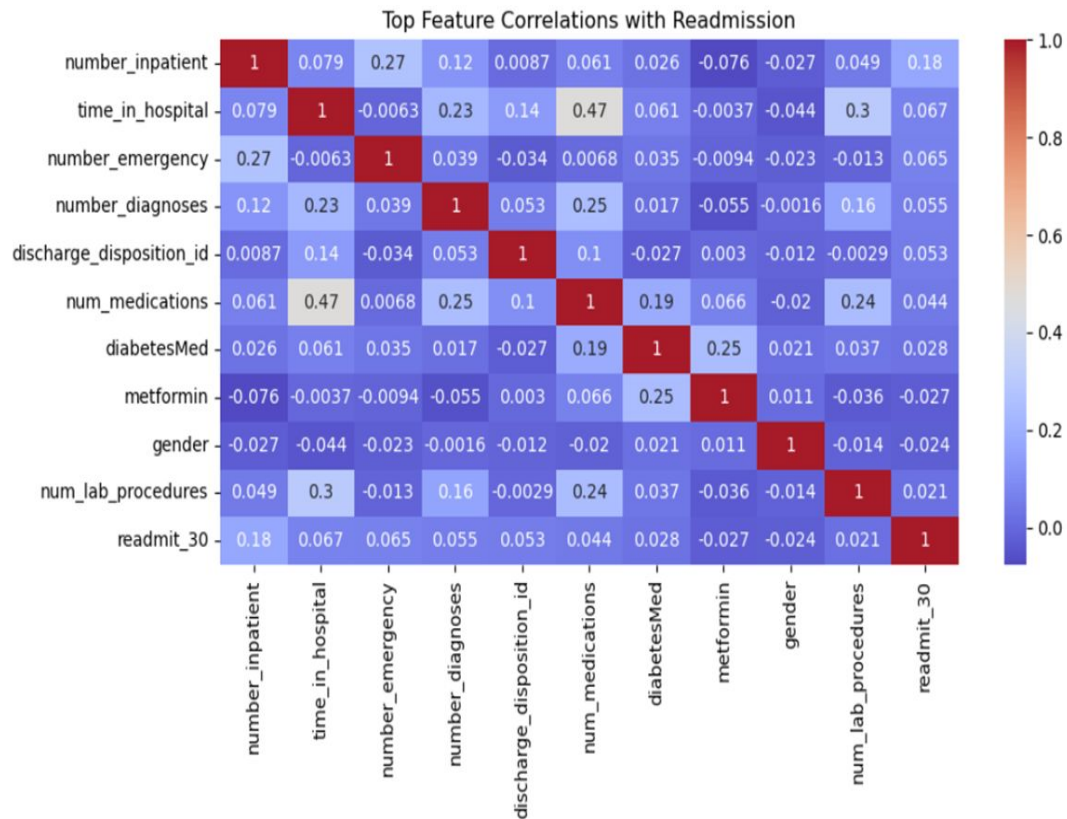
# Exploring Data

We were wondering if there was any other correlations in the data when you would compare them to other variables and one we tested was race vs readmission. One thing we saw that there was no dramatic differences in readmission rates were observed across races most of the cases were heavily either caucasian or african american patients. However, this feature was retained for model training to account for possible indirect relationships.



# Exploring Data

To find the most significant predictors, a heat map was created. The heat map showed that features like number of inpatient visits, emergency visits, and number of diagnoses showed the most correlation with early readmission. After exploring the data, the dataset was deemed “imbalanced”, with it favoring non-readmitted cases.





# Data Modeling



Three different modeling techniques were used to predict whether diabetic patients would readmit within 30 days, with those being Logistic Regression, Decision Tree Classifier, and Random Forest Classifier. In order to improve performance, we took a random sample of 10,000 records for training and evaluation, which was split into 80% training and 20% testing. Each model was then fit and evaluated using precision, recall, F1-score, and confusion matrix.

# Model Performance Summary



## Logistic Regression

- Accuracy: ~90%
- Precision (readmitted): 50%
- Recall (readmitted): 2.9%
- F1-Score (readmitted): 5.6%
- Confusion Matrix: 6 out of 203  
actual admissions

## Decision Tree

- Accuracy: ~80%
- Precision (readmitted): 14%
- Recall (readmitted): 19.2%
- F1-Score (readmitted): 16.2%
- Confusion Matrix: 39 out of 203  
actual admissions

## Random Forest

- Accuracy: ~90%
- Precision (readmitted): 50%
- Recall (readmitted): 0.5%
- F1-Score (readmitted): 0.97%
- Confusion Matrix: 1 out of 203  
actual admissions

# Interpretation and Challenges



## Challenges:

The biggest challenge we faced with this dataset was the severe class imbalance, with most patients not being readmitted. Due to the class we were trying to predict being the minority class, it made it much harder to predict.

## Insights:

Although it didn't have the highest precision, the Decision Tree performed the best when it came to identifying true positives. Logistic Regression and Random Forest however had a higher precision, but were much more biased towards the majority class.

# Conclusion



Predicting hospital readmission is a valuable, yet complex task due to data imbalance. While our models achieved high overall accuracy, they struggled with minority class detection. Addressing this imbalance and enriching features are the key to improving future predictive performances, and therefore future work should focus on better data representation and model robustness for medical predictions.