Predicting Hospital Readmissions using Machine Learning

Yashaswika Thota
Charitha Sri Iddum

TABLE OF CONTENTS

- Abstract
- Problem Statement
- Business Needs
- Data Science Process
- Dataset
- High Level Architecture Design
- Minimum Viable Product
- Gantt Chart
- Data Preprocessing
- ▣ Make Prediction on New Data
- ▣ Future Work

Abstract

- Despite multiple efforts to reduce readmissions, hospitals still struggle to accurately identify at-risk patients, particularly those with diabetes. Timely intervention is crucial for effective prevention. This project aims to leverage machine learning to predict which **diabetic patients** are at high risk of readmission, enabling healthcare providers to take proactive measures. By improving patient outcomes, reducing unnecessary readmissions, and optimizing hospital resource allocation, this approach can enhance the overall efficiency of diabetes-related care.

PROBLEM:

▣ Who among the hospitalized Diabetic patients are at risk for 30- day hospital readmissions?

▣ Can machine learning models accurately predict which diabetic patients are likely to be readmitted within 30 days of discharge?

▣ Which features are important predictors of 30-day readmission in patients with Diabetes?

BUSINESS NEEDS

Identification of High-Risk Patients: The need to pinpoint patients at high risk for readmission to prevent avoidable hospital admissions.

Machine Learning Utilization: Implementing machine learning models for better prediction of readmission risks, improving the accuracy of patient stratification.

Resource Optimization: Enhancing hospital resource utilization and operational efficiency to manage high-risk patients better.

Hospital Ratings: Improving hospital performance through a reduction in readmission rates, which will enhance patient satisfaction and ratings.

Financial Return: Reducing unnecessary readmissions to lower hospital costs and improve financial performance, particularly by decreasing excess readmission ratios.

HOSPITAL READMISSIONS DIABETES DATA SET

- The dataset represents 10 years 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.

It is an inpatient encounter (a hospital admission).

It is a diabetic encounter, that is, one during which any kind of diabetes was entered to the system as a diagnosis.

Minimum Viable Product

- **Goal**

Predict hospital readmission risks

Improve patient care & reduce costs

- **Key Components**

Data Preparation: Clean and process hospital records

Exploratory Analysis: Identify risk factors from data

Predictive Modeling: Build and evaluate ML model

Deployment: Create an API/dashboard for clinicians

- **Why It Matters**

Cuts healthcare expenses by preventing readmissions

Enhances patient outcomes with early intervention

Supports hospitals in making informed care decisions

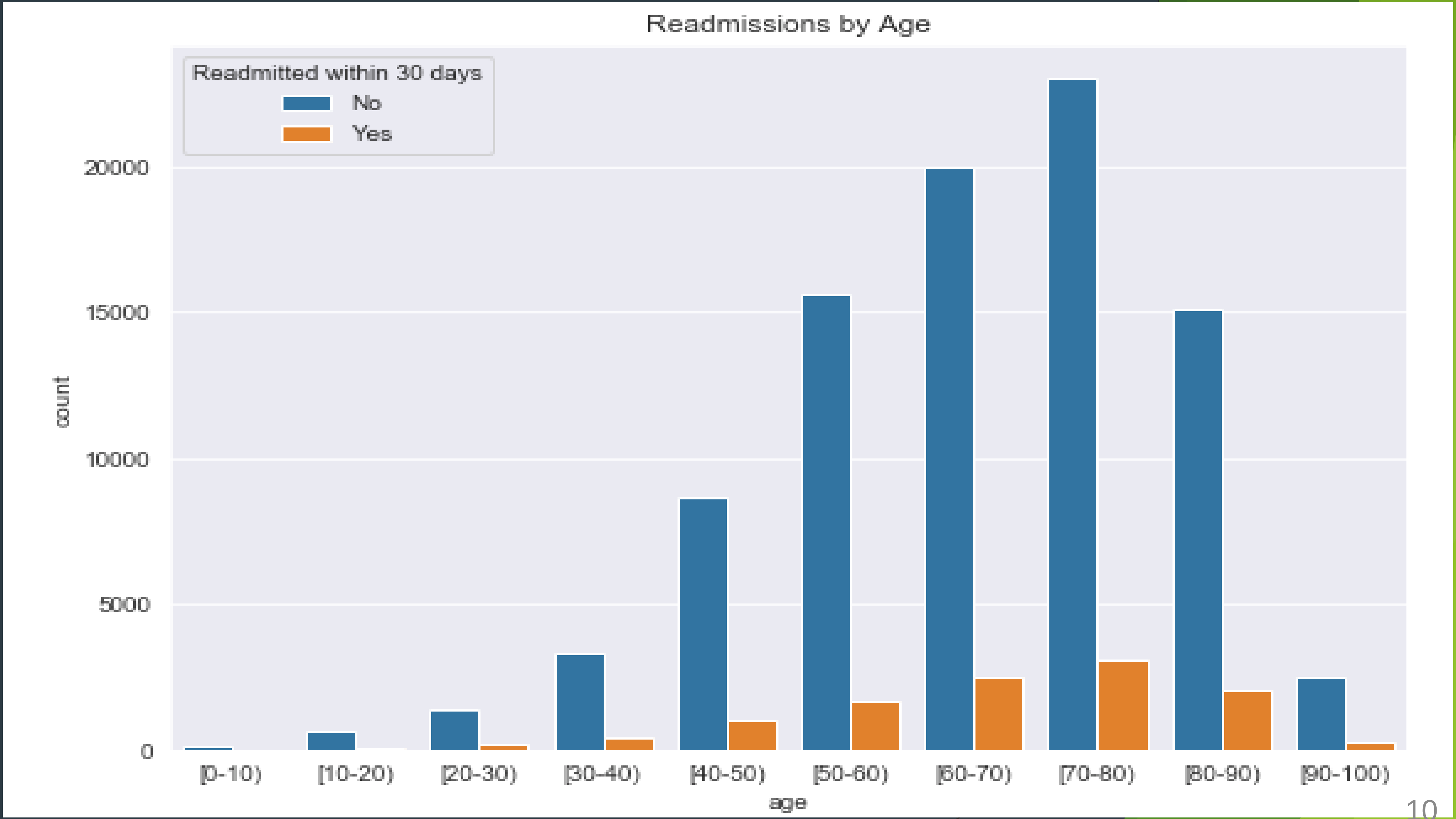
DATA PRE-PROCESSING

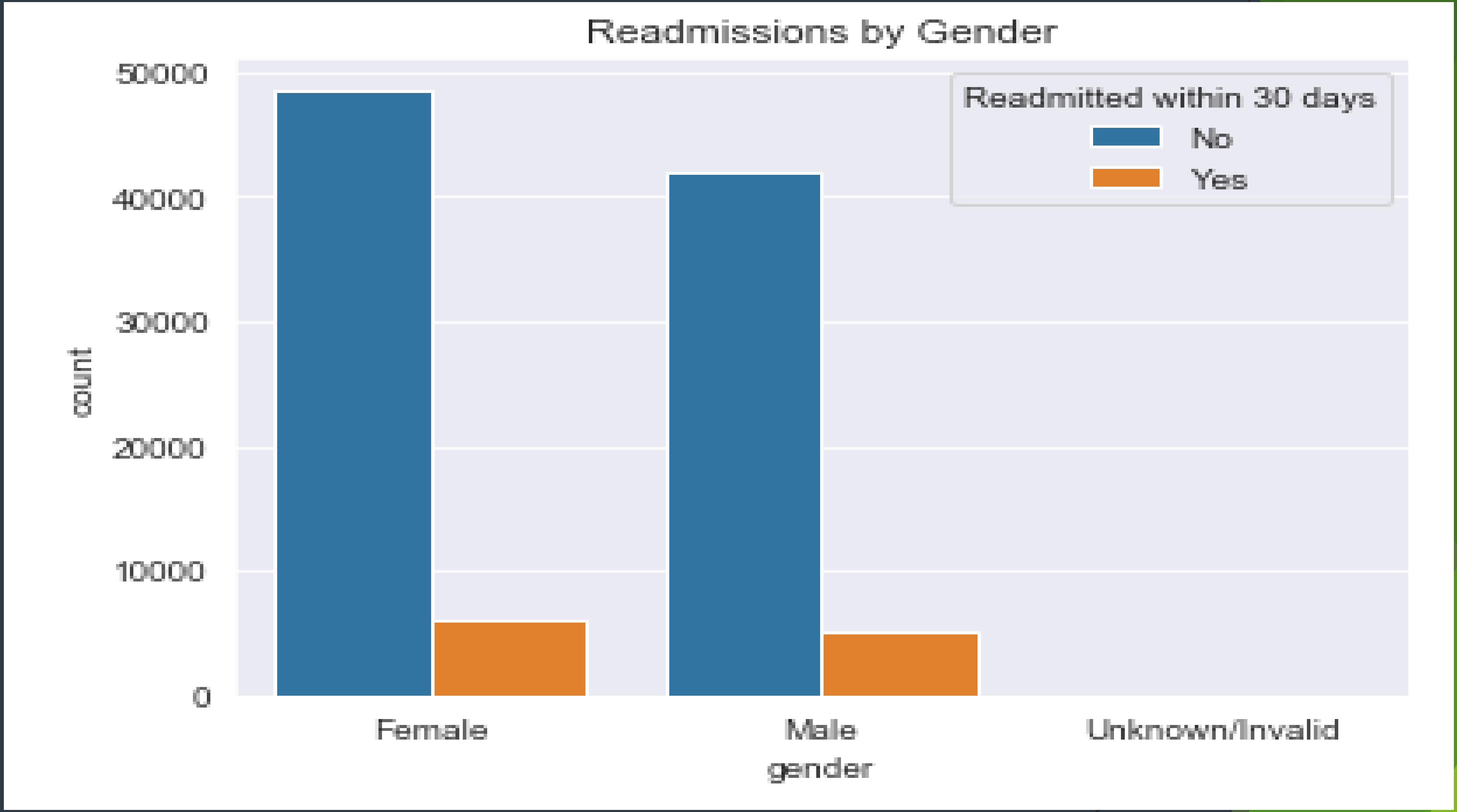
- Missing values
- Aggregations
- Cardinality – attributes with same values
- Normality, Multicollinearity checks
- Outliers removal

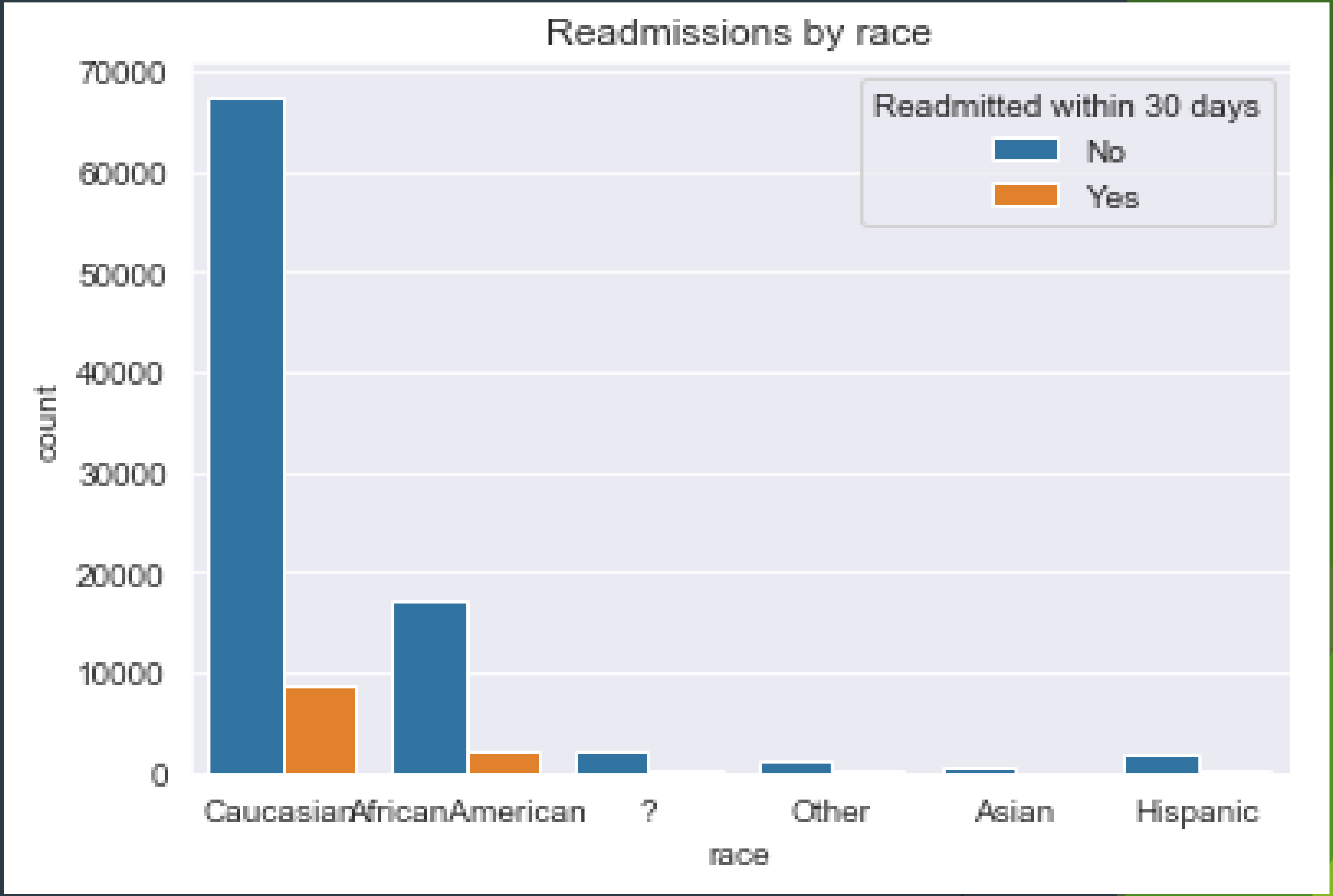
- ICD-10 mapping
- Feature creation
- Feature Subset selection
- Standardization
- Sampling

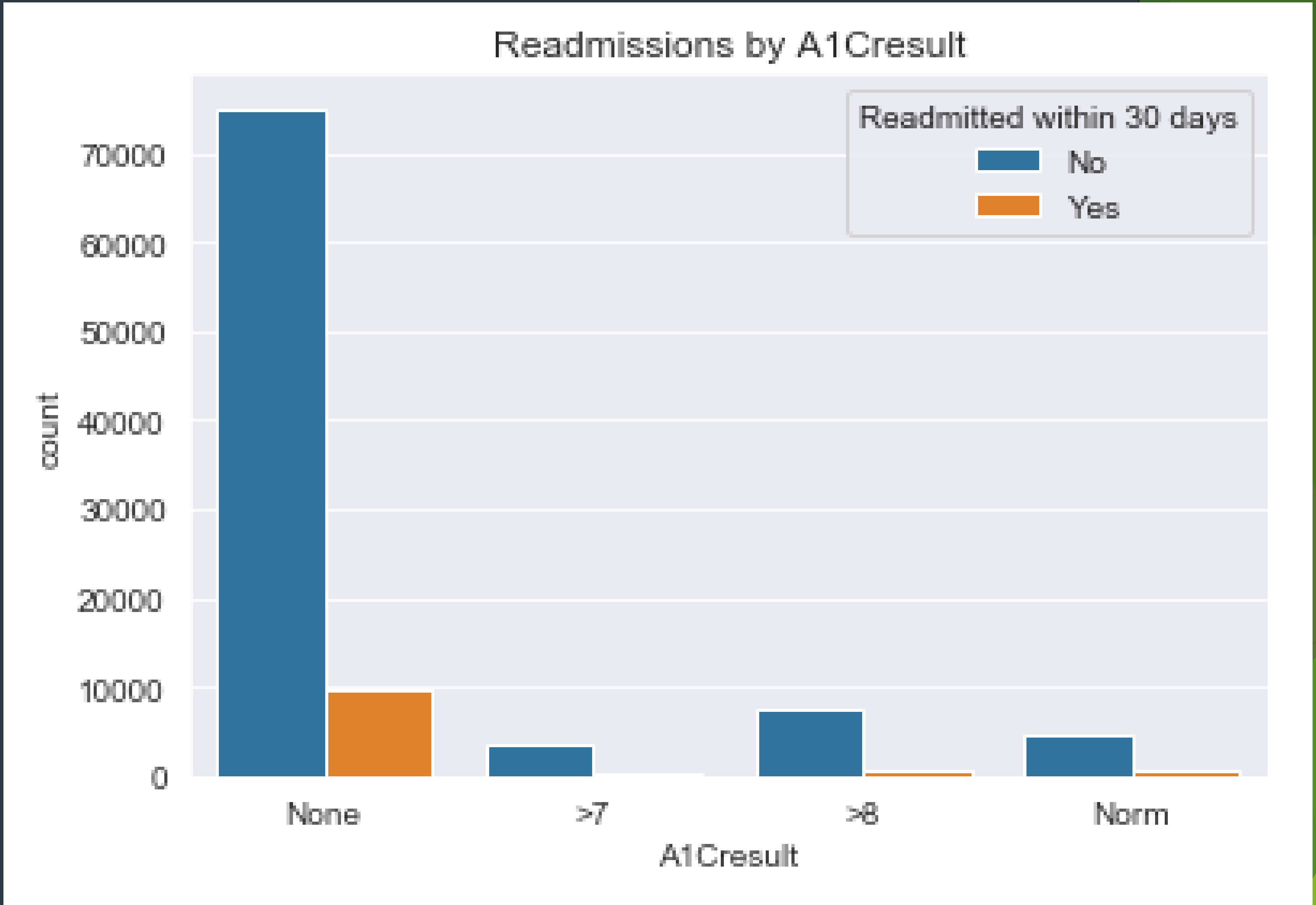
DATA PRE-PROCESSING

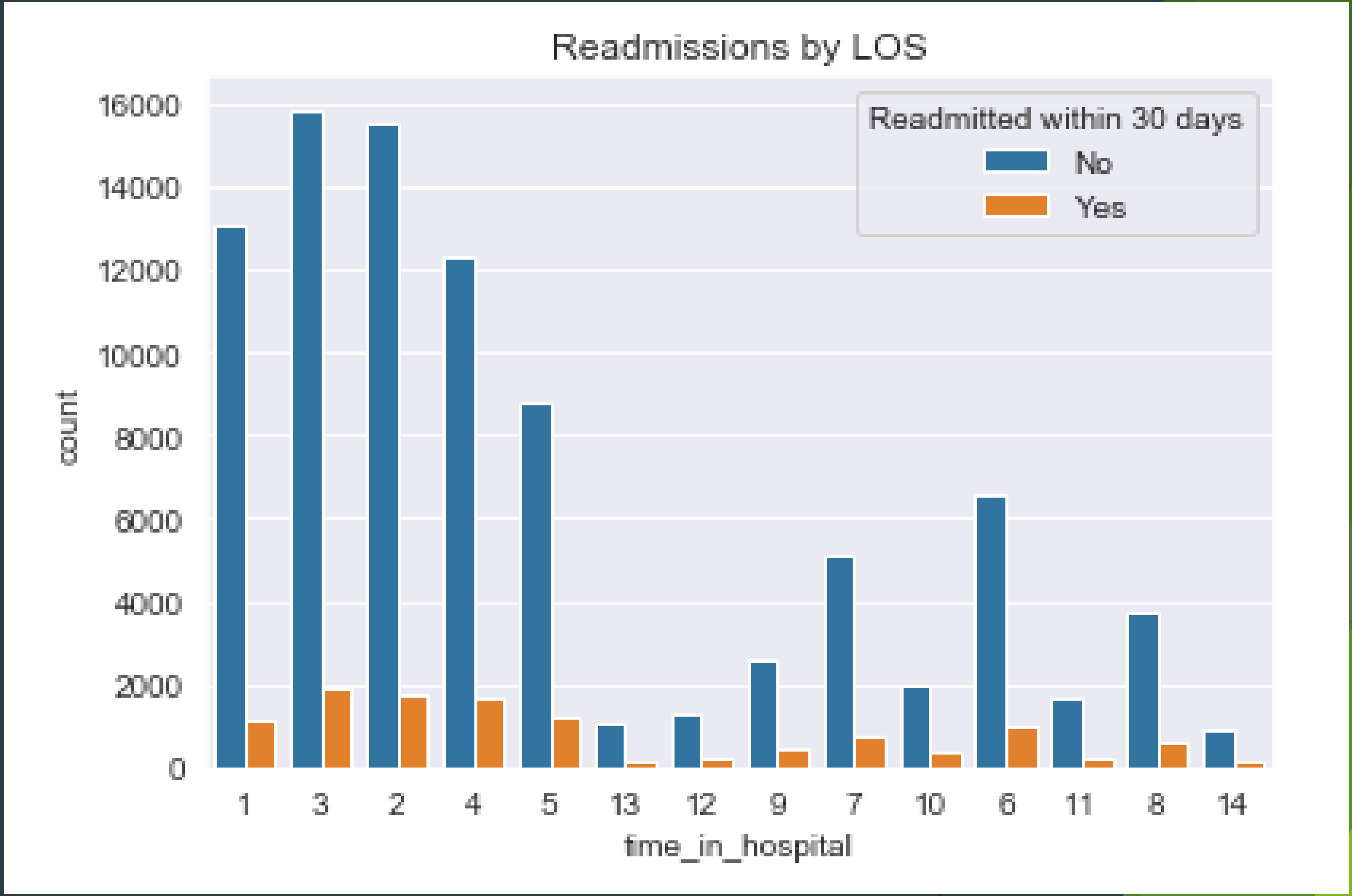
- Target impactable patients with diabetes who are at risk for 30-day hospital readmissions
- Excluded Discharge disposition:
  - deceased /transfers/ hospice
  - left against medical advice
- Removed duplicate patients
  - Index visit

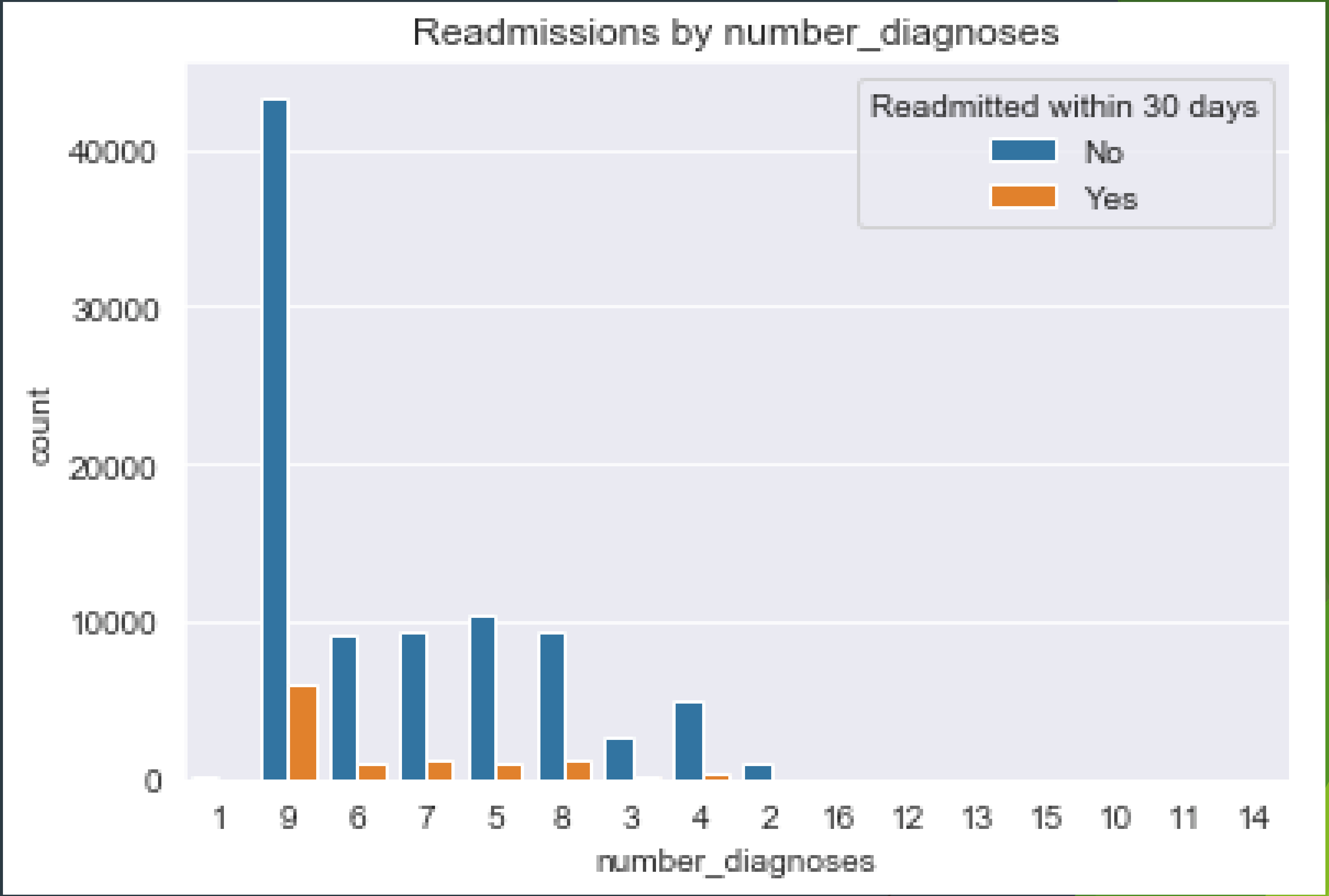

Readmissions by Age
Readmitted within 30 days
No
Yes
20000
15000
10000
5000
0
count
[0-10)
[10-20)
[20-30)
[30-40)
[40-50)
[50-60)
[60-70)
[70-80)
[80-90)
[90-100)
age


Readmissions by Gender
Readmitted within 30 days
No
Yes
50000
40000
30000
20000
10000
0
count
Female
Male
Unknown/Invalid
gender

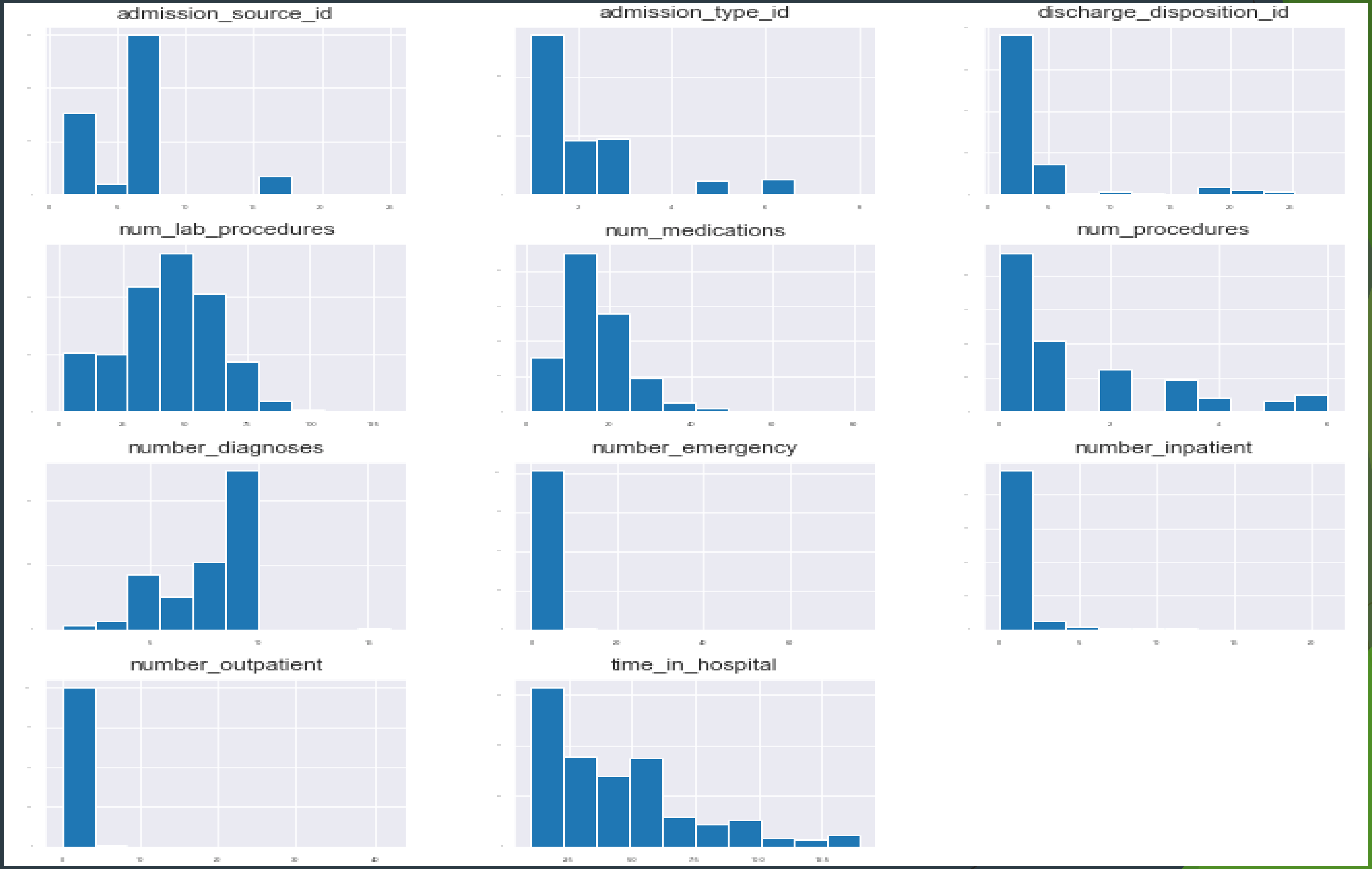

Readmissions by race
70000
60000
50000
40000
30000
20000
10000
0
count
Readmitted within 30 days
No
Yes
Caucasian
AfricanAmerican
?
Other
Asian
Hispanic
race


Readmissions by A1Cresult
Readmitted within 30 days
No
Yes
count
70000
60000
50000
40000
30000
20000
10000
0
None
>7
>8
Norm
A1Cresult

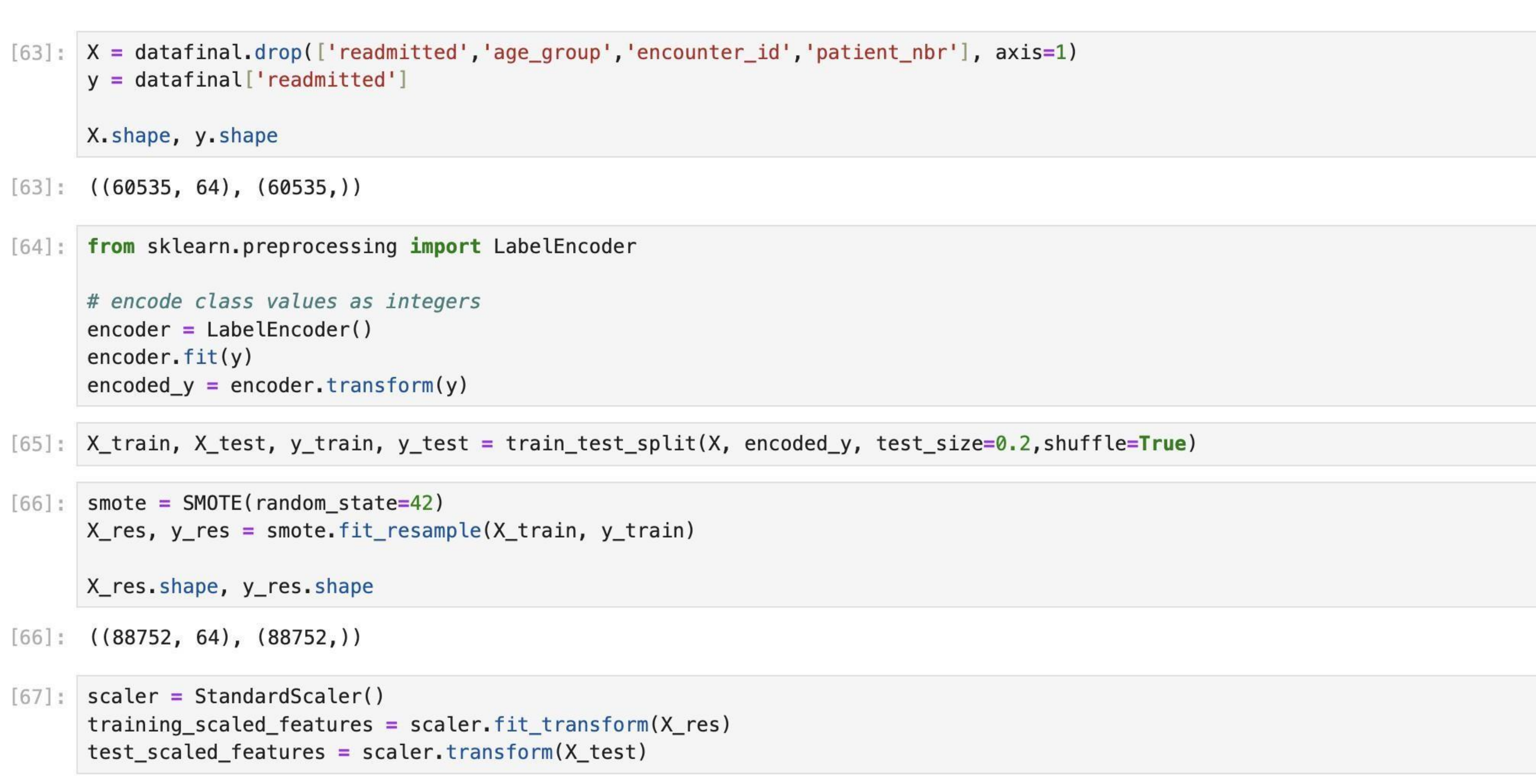

Readmissions by LOS
Readmitted within 30 days
No
Yes
count
16000
14000
12000
10000
8000
6000
4000
2000
0
1
3
2
4
5
13
12
9
7
10
6
11
8
14
time_in_hospital


Readmissions by number_diagnoses
Readmitted within 30 days
No
Yes
count
40000
30000
20000
10000
0
1
9
6
7
5
8
3
4
2
16
12
13
15
10
11
14
number_diagnoses

NORMALITY CHECK

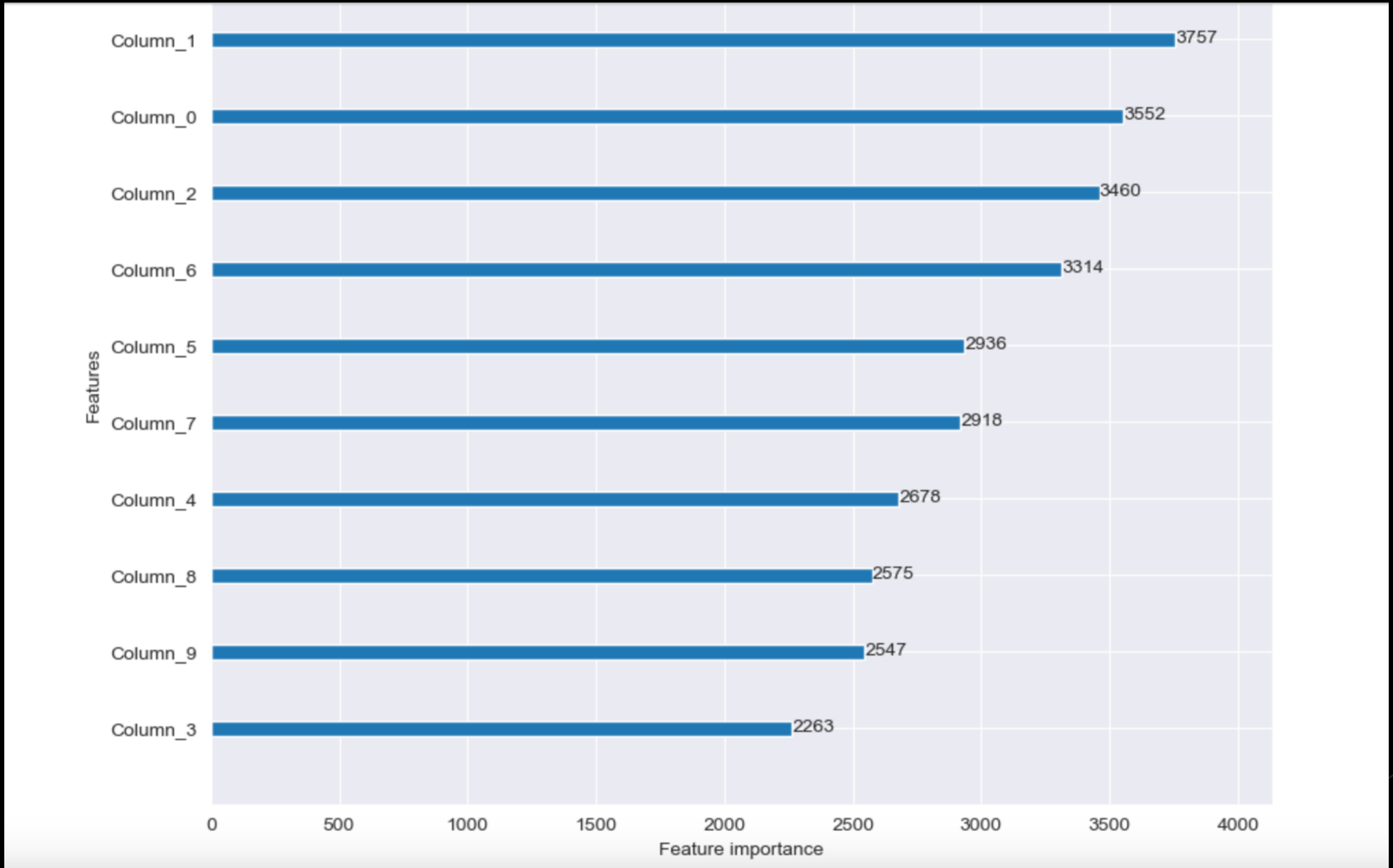


Model Training-Testing

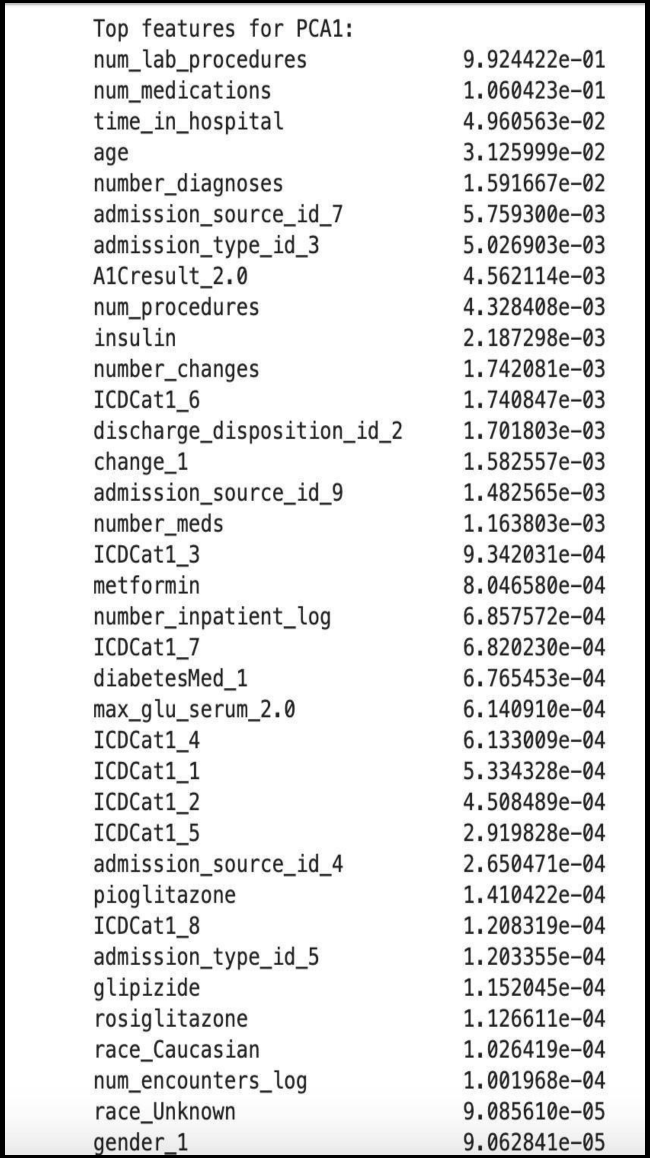


```
[63]: X = datafinal.drop(['readmitted','age_group','encounter_id','patient_nbr'], axis=1)
      y = datafinal['readmitted']

      X.shape, y.shape

[63]: ((60535, 64), (60535,))

[64]: from sklearn.preprocessing import LabelEncoder

      # encode class values as integers
      encoder = LabelEncoder()
      encoder.fit(y)
      encoded_y = encoder.transform(y)

[65]: X_train, X_test, y_train, y_test = train_test_split(X, encoded_y, test_size=0.2,shuffle=True)

[66]: smote = SMOTE(random_state=42)
      X_res, y_res = smote.fit_resample(X_train, y_train)

      X_res.shape, y_res.shape

[66]: ((88752, 64), (88752,))

[67]: scaler = StandardScaler()
      training_scaled_features = scaler.fit_transform(X_res)
      test_scaled_features = scaler.transform(X_test)
```

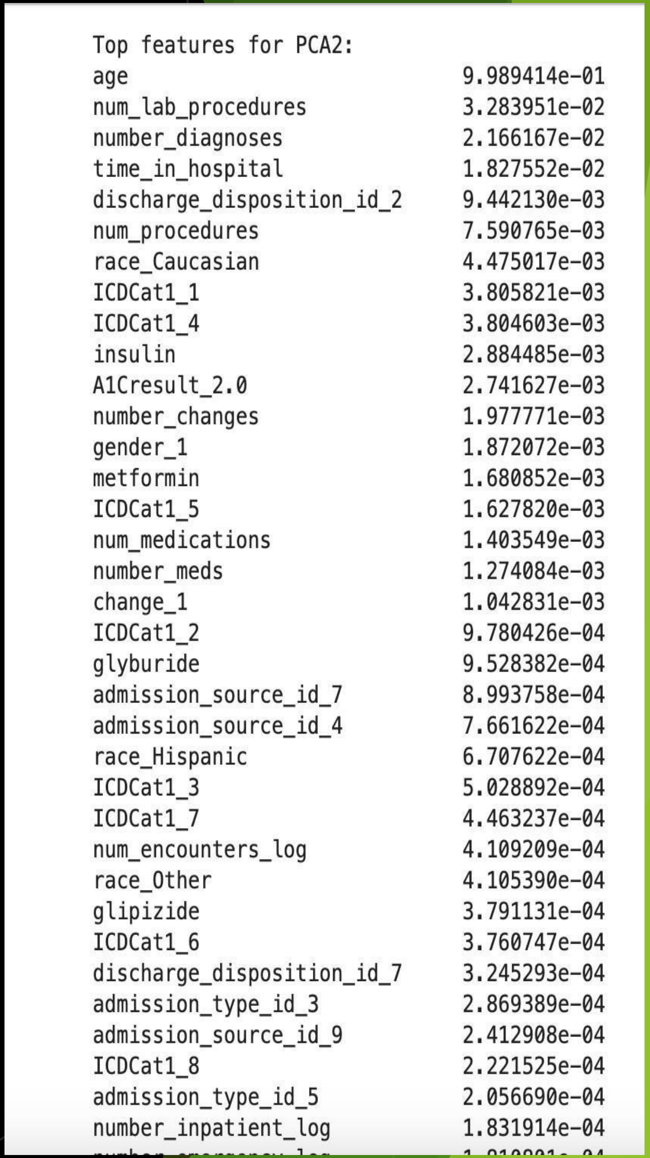
Feature Importance



FEATURE SELECTION



```
Top features for PCA1:
num_lab_procedures              9.924422e-01
num_medications                 1.060423e-01
time_in_hospital                4.960563e-02
age                             3.125999e-02
number_diagnoses                1.591667e-02
admission_source_id_7           5.759300e-03
admission_type_id_3             5.026903e-03
A1Cresult_2.0                   4.562114e-03
num_procedures                  4.328408e-03
insulin                         2.187298e-03
number_changes                  1.742081e-03
ICDCat1_6                       1.740847e-03
discharge_disposition_id_2      1.701803e-03
change_1                        1.582557e-03
admission_source_id_9           1.482565e-03
number_meds                     1.163803e-03
ICDCat1_3                       9.342031e-04
metformin                       8.046580e-04
number_inpatient_log            6.857572e-04
ICDCat1_7                       6.820230e-04
diabetesMed_1                   6.765453e-04
max_glu_serum_2.0               6.140910e-04
ICDCat1_4                       6.133009e-04
ICDCat1_1                       5.334328e-04
ICDCat1_2                       4.508489e-04
ICDCat1_5                       2.919828e-04
admission_source_id_4           2.650471e-04
pioglitazone                    1.410422e-04
ICDCat1_8                       1.208319e-04
admission_type_id_5             1.203355e-04
glipizide                       1.152045e-04
rosiglitazone                   1.126611e-04
race_Caucasian                  1.026419e-04
num_encounters_log              1.001968e-04
race_Unknown                    9.085610e-05
gender_1                        9.062841e-05
```



```
Top features for PCA2:
age                             9.989414e-01
num_lab_procedures              3.283951e-02
number_diagnoses                2.166167e-02
time_in_hospital                1.827552e-02
discharge_disposition_id_2      9.442130e-03
num_procedures                  7.590765e-03
race_Caucasian                  4.475017e-03
ICDCat1_1                       3.805821e-03
ICDCat1_4                       3.804603e-03
insulin                         2.884485e-03
A1Cresult_2.0                   2.741627e-03
number_changes                  1.977771e-03
gender_1                        1.872072e-03
metformin                       1.680852e-03
ICDCat1_5                       1.627820e-03
num_medications                 1.403549e-03
number_meds                     1.274084e-03
change_1                        1.042831e-03
ICDCat1_2                       9.780426e-04
glyburide                       9.528382e-04
admission_source_id_7           8.993758e-04
admission_source_id_4           7.661622e-04
race_Hispanic                   6.707622e-04
ICDCat1_3                       5.028892e-04
ICDCat1_7                       4.463237e-04
num_encounters_log              4.109209e-04
race_Other                      4.105390e-04
glipizide                       3.791131e-04
ICDCat1_6                       3.760747e-04
discharge_disposition_id_7      3.245293e-04
admission_type_id_3             2.869389e-04
admission_source_id_9           2.412908e-04
ICDCat1_8                       2.221525e-04
admission_type_id_5             2.056690e-04
number_inpatient_log            1.831914e-04
```

PERFORMANCE METRICS

▣ Confusion Matrix - a table showing correct predictions and types of incorrect predictions.

▣ Precision - proportion of + identifications that are correct (TP/(TP+FP))

▣ Recall - proportion of actual positives identified correctly (TP/(TP+FN))

▣ Receiver operating characteristics curve - diagnostic ability of a binary classifier

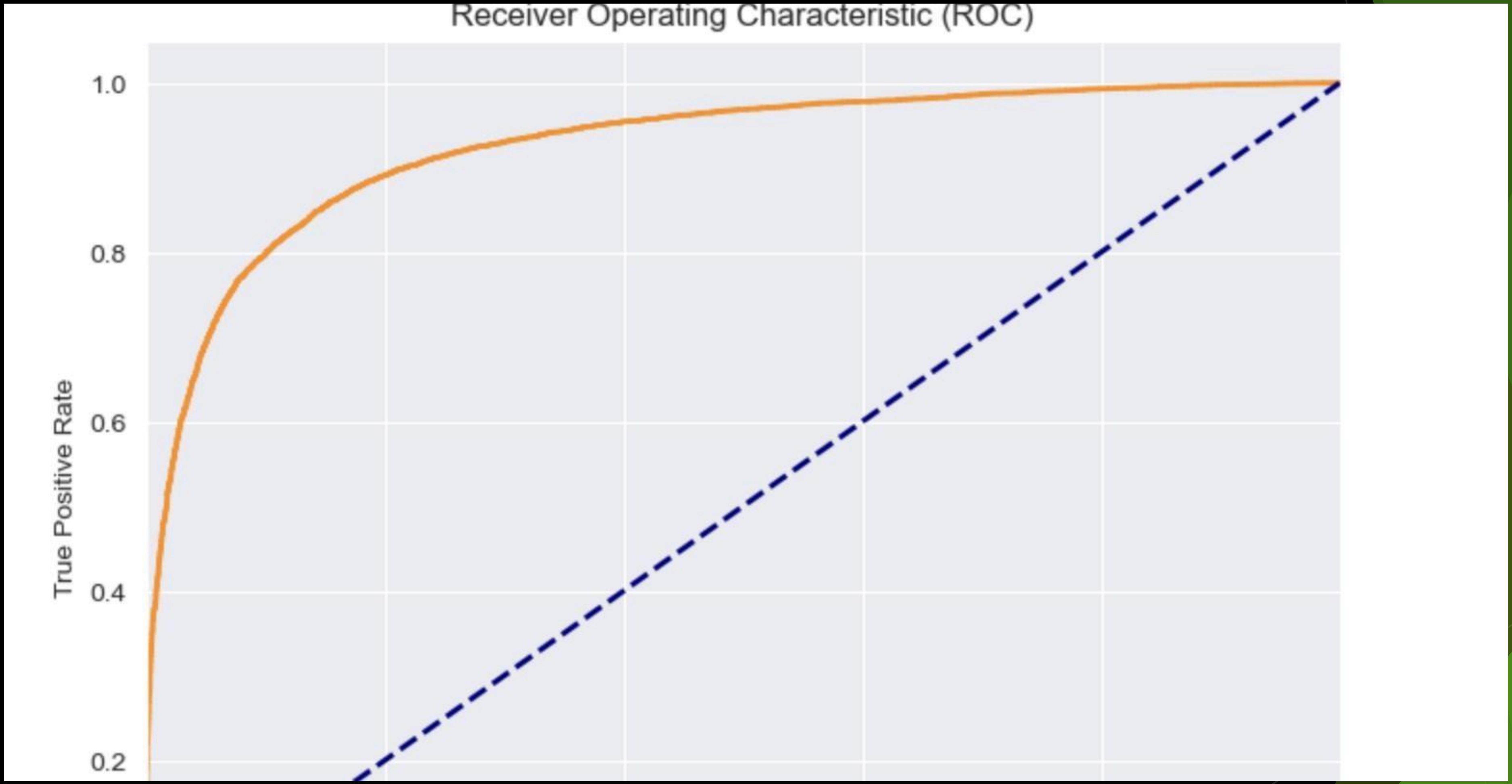
📊 Updated ML Model Performance Comparison (Class 0 and 1)

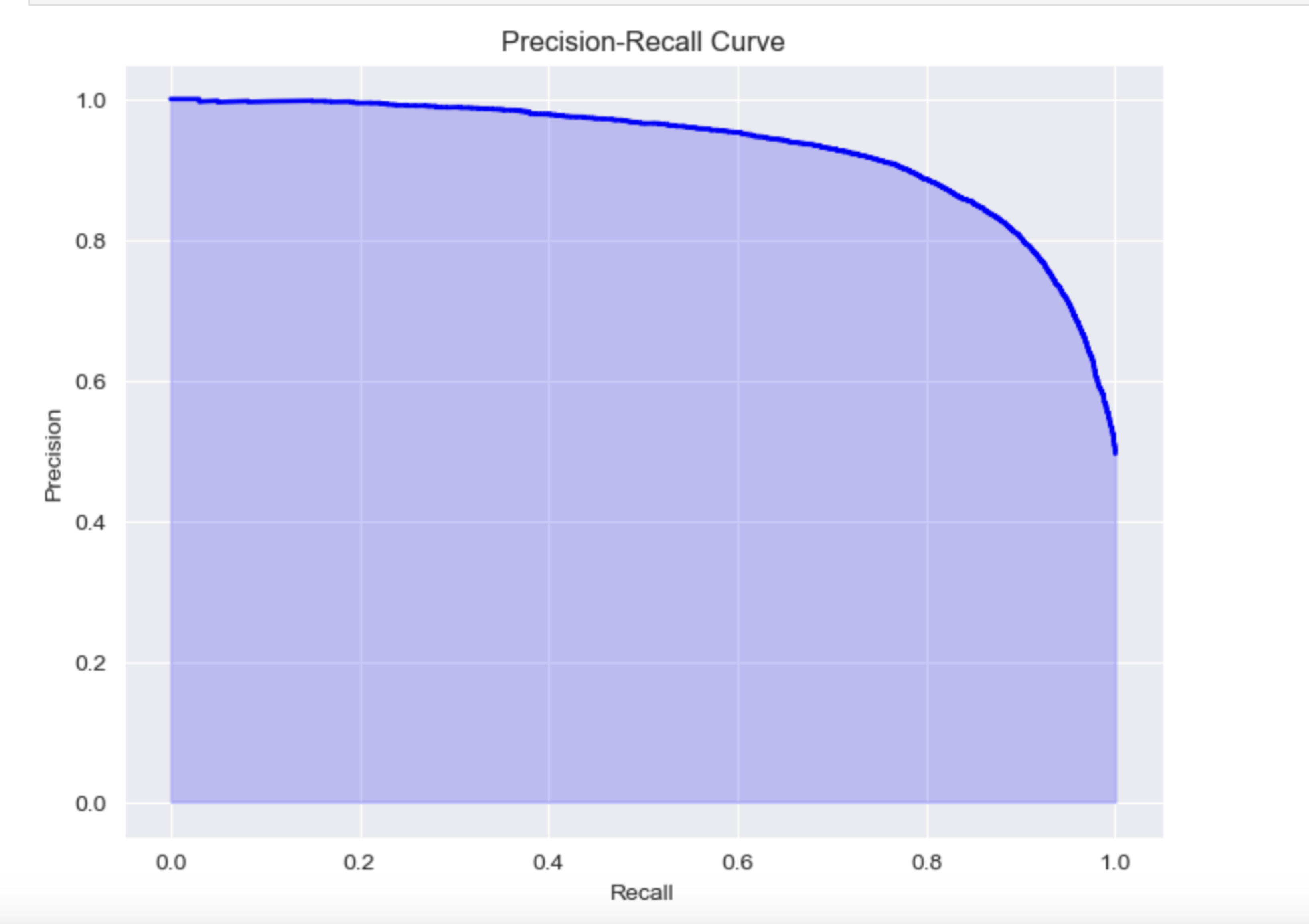
Model	Acc	Prec (0)	Rec (0)	F1 (0)	Prec (1)	Rec (1)	F1 (1)	ROC AUC	PR AUC
Logistic Regression	0.69	0.92	0.73	0.81	0.09	0.29	0.14	0.525	0.0937
Random Forest	0.85	0.92	0.91	0.92	0.12	0.12	0.12	0.546	0.1035
XGBoost	0.79	0.92	0.84	0.88	0.11	0.20	0.14	0.533	0.0983
AdaBoost	0.71	0.92	0.75	0.82	0.09	0.27	0.14	0.527	0.0948
KNN	0.67	0.92	0.71	0.80	0.09	0.32	0.15	0.524	0.1308
Extra Trees	0.83	0.92	0.90	0.91	0.12	0.14	0.13	0.552	0.1055
Stacking Classifier	0.85	0.91	0.93	0.92	0.11	0.08	0.10	–	–
Voting Classifier	0.82	0.91	0.88	0.90	0.12	0.17	0.14	–	–
LightGBM	0.86	0.86	0.87	0.86	0.87	0.85	0.86	**0.931**	–

Model selection

After evaluating multiple machine learning models on our dataset using key classification metrics, **LightGBM emerged as the most effective model** for our task. Here's why:

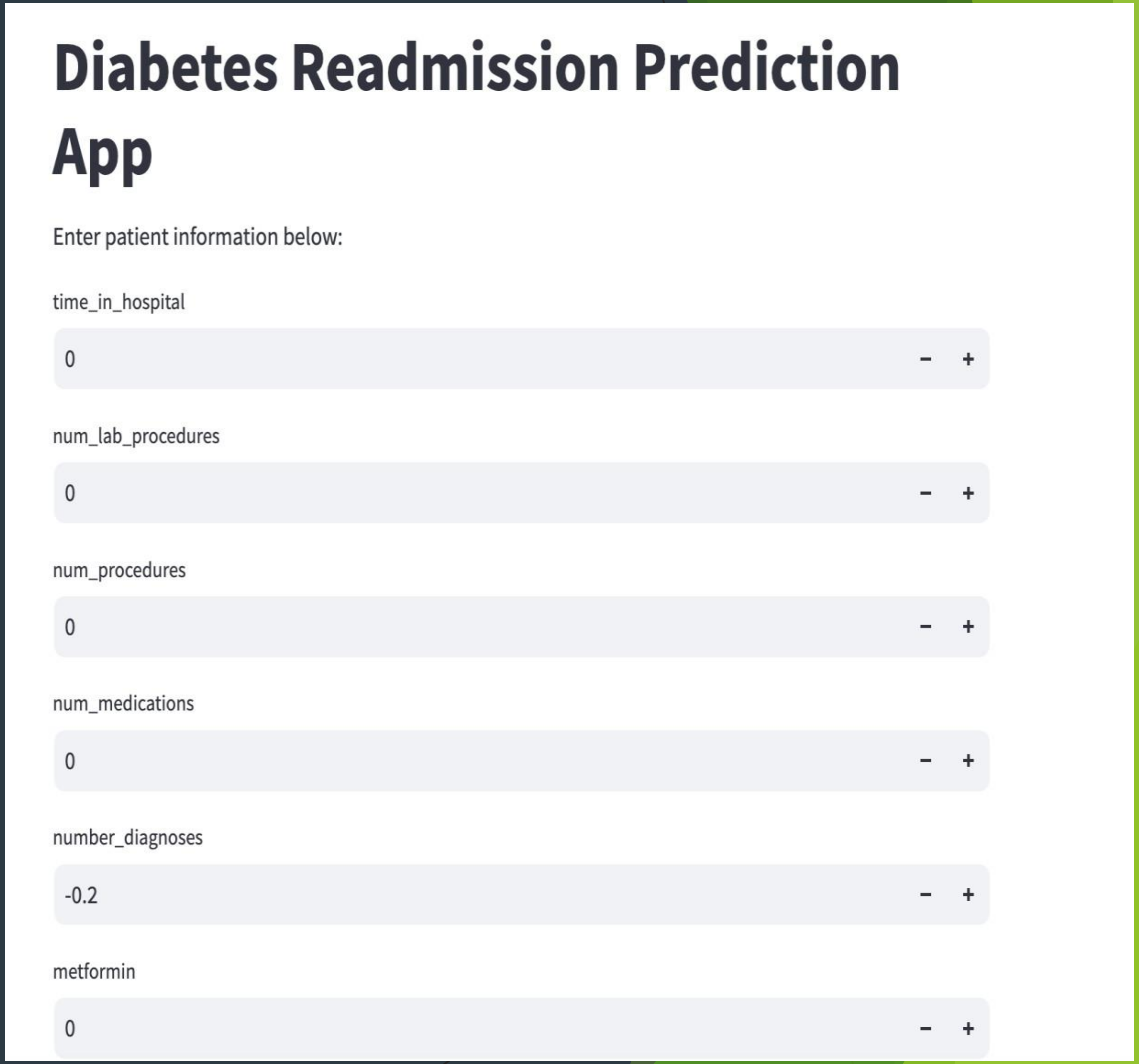
- It achieved the **highest overall accuracy of 86.26%**, outperforming all other models including Random Forest, XGBoost, AdaBoost, and ensemble methods.
- Unlike other models, LightGBM provided **balanced performance for both classes**:
  - Class 0: Precision = 0.86, Recall = 0.87, F1 = 0.86
  - Class 1: Precision = 0.87, Recall = 0.85, F1 = 0.86
- It recorded the **best ROC AUC score (0.931)**, indicating excellent discriminatory power between the classes — especially important in imbalanced classification problems.
- Other models, while strong in predicting the majority class (class 0), struggled with minority class (class 1). LightGBM, however, maintained high performance on **both classes**, showing its robustness and effectiveness.


Receiver Operating Characteristic (ROC)
True Positive Rate
1.0
0.8
0.6
0.4
0.2


Precision-Recall Curve
Precision
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Recall

Deployment

- Web app built using **Streamlit**
- User-friendly interface to input patient data
- Predicts **whether the patient is at risk of readmission**
- Easily accessible for healthcare providers to use



Diabetes Readmission Prediction App

Enter patient information below:

time_in_hospital

0

num_lab_procedures

0

num_procedures

0

num_medications

0

number_diagnoses

-0.2

metformin

0

Conclusion

In this project, we used machine learning to predict 30-day hospital readmission for diabetic patients. After testing different models, **LightGBM** gave the best and most balanced results, especially in identifying patients who were likely to be readmitted.

Our approach shows that using the right preprocessing steps (like balancing the data and reducing features) and choosing the right model is very important in healthcare prediction tasks.

In the future, we plan to improve the model further by using more detailed clinical data and testing it on different hospital systems.

THANK YOU!