Hack-A-Stat 2025

Group Name- Code Syrup

- Aarshi Shaikh A066
- Prashant Srivastava A068
- Harsh Tantak A069

Github Link :- https://github.com/HarshTantak/J-and-J-hackathon

Problem Statement-

The dataset represents ten years (1999-2008) of clinical care at 130 US hospitals and integrated delivery networks. Each row concerns hospital records of subjects diagnosed with diabetes, who underwent laboratory medications and stayed up to 14 days.

Objective-

To analyze hospital readmission data of diabetic patients and determine factors contributing to early readmissions(within 30 days of discharge) using statistical, survival analysis, and machine learning techniques

Introduction-

Hospital readmissions are a critical indicator of healthcare quality and system efficiency. They are often associated with higher healthcare costs, patient dissatisfaction, and potential gaps in patient care. Among various conditions leading to hospitalizations diabetes stands out as a significant burden due to its chronic nature and associated complications. Managing diabetes and reducing readmission is a priority for healthcare systems worldwide.

In the given study we analyze a dataset comprising of ten years (1999-2008) of clinical data from 130 US hospitals and integrated delivery networks.This dataset focuses on diabetic patients who underwent laboratory tests, received medications and stayed up to 14 days. Each patient encounter is documented,including demographic information,clinical details and discharge outcomes.

The primary goal of this analysis is to determine the factors contributing to early readmissions (within 30 days of discharge).
To conduct a good study on the readmissions of patients we are going to consider the following features that are present in the dataset-

Feature Name	**Type**	**Description**
Race	Categorical	Values: Caucasian, Asian, African, American ,Hispanic, and other
Gender	Categorical	Values: Male, female and unknown/invalid
Age	Categorical	Grouped in 10-year intervals: [0, 10), [10, 20),..., [90, 100)
Medical Speciality	Categorical	Integer identifier of a specialty of the admitting physician, corresponding to 84 distinct values, for example, cardiology, internal medicine, family/general practice, and surgeon
Number of lab procedures	Integer	Number of lab tests performed during the encounter
Number Of Procedures	Integer	Number of procedures (other than lab tests) performed during the encounter
No of medications	Integer	Number of distinct generic names administered during the encounter
A1Cresult	Categorical	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.
Change	Categorical	Indicates if there was a change in diabetic medications (either dosage or generic name). Values: change and no change
diabetesMed	Categorical	Indicates if there was any diabetic medication prescribed. Values: yes and no
Readmitted	Categorical	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.
Time in Hospital	Integer	Integer number of days between admission and discharge
Admission type id	Categorical	Integer identifier corresponding to 21 distinct values, for example, physician referral, emergency room, and transfer from a hospital
discharge_disposition_id	Categorical	Integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, and not available
number_diagnoses	Integer	Number of diagnoses entered to the system
max_glu_serum	Categorical	Indicates the range of the result or if the test was not taken. Values: >200, >300, normal, and none if not measured
metformin	Categorical	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
insulin	categorical	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

This study combines statistical analysis, machine learning techniques, and survival analysis to achieve these objectives and offer actionable insights into diabetic patient management.

Problem Solutions

Q-1) **If there are any missing data, explore the same. How do you tackle the missingness if you were to fit a predictive model? Please provide a logical explanation.**

Sol) The dataset's Initial inspection revealed a mix of categorical and numerical columns. This dataset includes demographic information, medical history, and medication details. So, to properly understand and clean data we have converted categorical columns to numerical columns with the help of encoding

Preprocessing Steps-

1. Categorical Encoding-
   Categorical variables (e.g. race, gender, age) were label-encoded to convert string categories into numerical representations for ease of analysis
2. Clubbing 'readmitted'-
   Values in the readmitted column (<30,>30, and no) were clubbed into binary categories. Specially
   - <30 was mapped to 0 (indicating early readmission within 30 days)
   - >30 and no were both mapped to 1 (indicating no early admission)

3. Handling Missing Values-

- Missing values in A1Cresult was imputed using the mode
- Rows where medical_speciality labelled as ? were identified and stored in a new dataframe (unknown_Specialty) for prediction later
- A label counter was used to transform categorical values in the medical specialty column into numeric labels for model training
- We have removed weight column from the dataset under the assumption that since out 69,000 entries about 67,000 entries of weight were missing values, so we assumed that weight will not apply a major role in prediction of readmission variable

4. Model Purpose and Overview-

This logistic regression model predicts early readmission (within 30 days) for diabetic patients using clinical and demographic data. Early readmissions are costly and indicate poor outcomes; the model aims to identify high-risk patients for better intervention.This logistic regression model predicts early readmission (within 30 days) for diabetic patients using clinical and demographic data. Early readmissions are costly and indicate poor outcomes; the model aims to identify high-risk patients for better intervention.

Key Preprocessing Steps

- Target Variable: Recategorized readmission as binary:
  - Yes (1): Readmission within 30 days.
  - No (0): No or late readmission.
- Feature Selection: Selected relevant predictors, including:
  - Demographics: Age, gender, race.
  - Hospital Stay: Admission type, time in hospital, discharge disposition.
  - Medical History & Labs: Diagnoses, glucose serum, A1C.
  - Medications: Metformin, insulin usage.
- Data Preparation:
  - Handled missing values by replacing them with `0`.
  - Encoded categorical features using one-hot encoding.
  - Split data into training (80%) and testing (20%) sets.
- Model Performance: The model achieved **95% overall accuracy**, driven by high precision and recall for the majority class (No Early Readmission). Despite the imbalance issue, the model provides a robust starting point for identifying non-readmissions, offering insights that could help allocate resources more effectively to high-risk patients. Further refinement (e.g., addressing imbalance) could enhance minority class performance.

Classification Report:

	precision	recall	f1-score	support
0	0.95	1.00	0.98	13191
1	0.67	0.01	0.01	643
accuracy			0.95	13834
macro avg	0.81	0.50	0.49	13834
weighted avg	0.94	0.95	0.93	13834

Q2) Starting with an exploratory analysis of the features, do you find anything surprising/noteworthy regarding any of the variables? Explain. (For example, amongst others, consider “num_medications”; are patients being given too many medications?).

Soln)
We used pandas library to generate summary metric for all the columns

Summary Statistics:

	Unnamed: 0	race	gender	age
count	69169.000000	69169.000000	69169.000000	69169.000000
mean	34584.000000	1.739045	0.467869	6.055386
std	19967.514722	0.924874	0.499057	1.600455
min	0.000000	0.000000	0.000000	0.000000
25%	17292.000000	2.000000	0.000000	5.000000
50%	34584.000000	2.000000	0.000000	6.000000
75%	51876.000000	2.000000	1.000000	7.000000
max	69168.000000	4.000000	2.000000	9.000000

	admission_type_id	discharge_disposition_id	admission_source_id
count	69169.000000	69169.000000	69169.000000
mean	2.069222	3.333488	5.662826
std	1.480825	4.980799	4.125038
min	1.000000	1.000000	1.000000
25%	1.000000	1.000000	1.000000
50%	1.000000	1.000000	7.000000
75%	3.000000	3.000000	7.000000
max	8.000000	28.000000	25.000000

glipizide.metformin	glimepiride.pioglitazone	metformin.rosiglitazone
69169.000000	69169.0	69169.000000
0.000101	0.0	0.000029
0.010059	0.0	0.005377
0.000000	0.0	0.000000
0.000000	0.0	0.000000
0.000000	0.0	0.000000
0.000000	0.0	0.000000
1.000000	0.0	1.000000

	metformin.pioglitazone	change	diabetesMed	readmitted
count	69169.000000	69169.000000	69169.000000	69169.000000
mean	0.000014	0.549567	0.760557	1.706516
std	0.003802	0.497541	0.426746	0.547562
min	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	1.000000	2.000000
50%	0.000000	1.000000	1.000000	2.000000
75%	0.000000	1.000000	1.000000	2.000000
max	1.000000	1.000000	1.000000	2.000000

	diag_1_name	diag_2_name	diag_3_name
count	69169.000000	69169.000000	69169.000000
mean	2.964276	2.237057	1.871113
std	2.771523	2.538342	2.256471
min	0.000000	0.000000	0.000000
25%	1.000000	0.000000	0.000000
50%	2.000000	1.000000	1.000000
75%	5.000000	3.000000	2.000000
max	8.000000	8.000000	8.000000

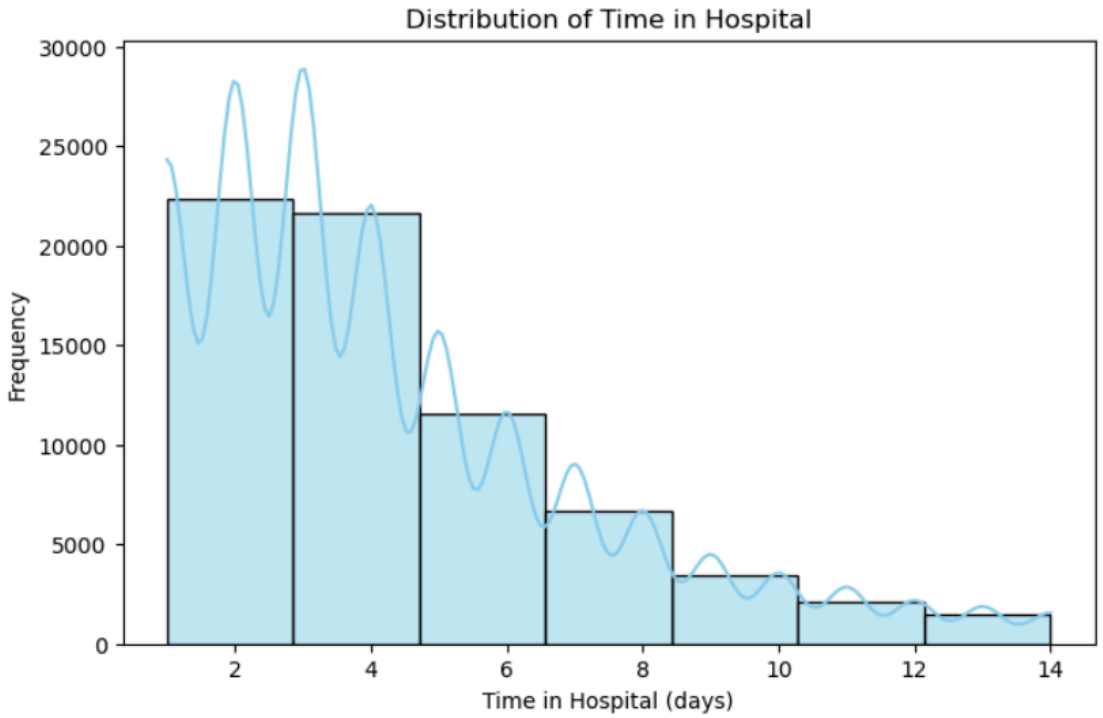
```
Unique Values:
 Unnamed: 0                69169
race                           5
gender                         3
age                           10
admission_type_id              8
discharge_disposition_id      21
admission_source_id           17
time_in_hospital              14
medical_specialty             70
num_lab_procedures           117
num_procedures                 7
num_medications               73
number_outpatient             33
number_emergency              26
number_inpatient              16
number_diag0ses               16
```

```
max_glu_serum                  4
A1Cresult                      4
metformin                      4
repaglinide                    4
nateglinide                    4
chlorpropamide                 4
glimepiride                    4
acetohexamide                  2
glipizide                      4
glyburide                      4
tolbutamide                    2
pioglitazone                   4
rosiglitazone                  4
acarbose                       4
miglitol                       4
troglitazone                   2
tolazamide                     2
examide                        1
```

```
citoglipton                    1
insulin                        4
glyburide.metformin            4
glipizide.metformin            2
glimepiride.pioglitazone       1
metformin.rosiglitazone        2
metformin.pioglitazone         2
change                         2
diabetesMed                    2
readmitted                     3
diag_1_name                    9
diag_2_name                    9
diag_3_name                    9
```

these are the statistics that we have calculated using pandas library for the given dataset

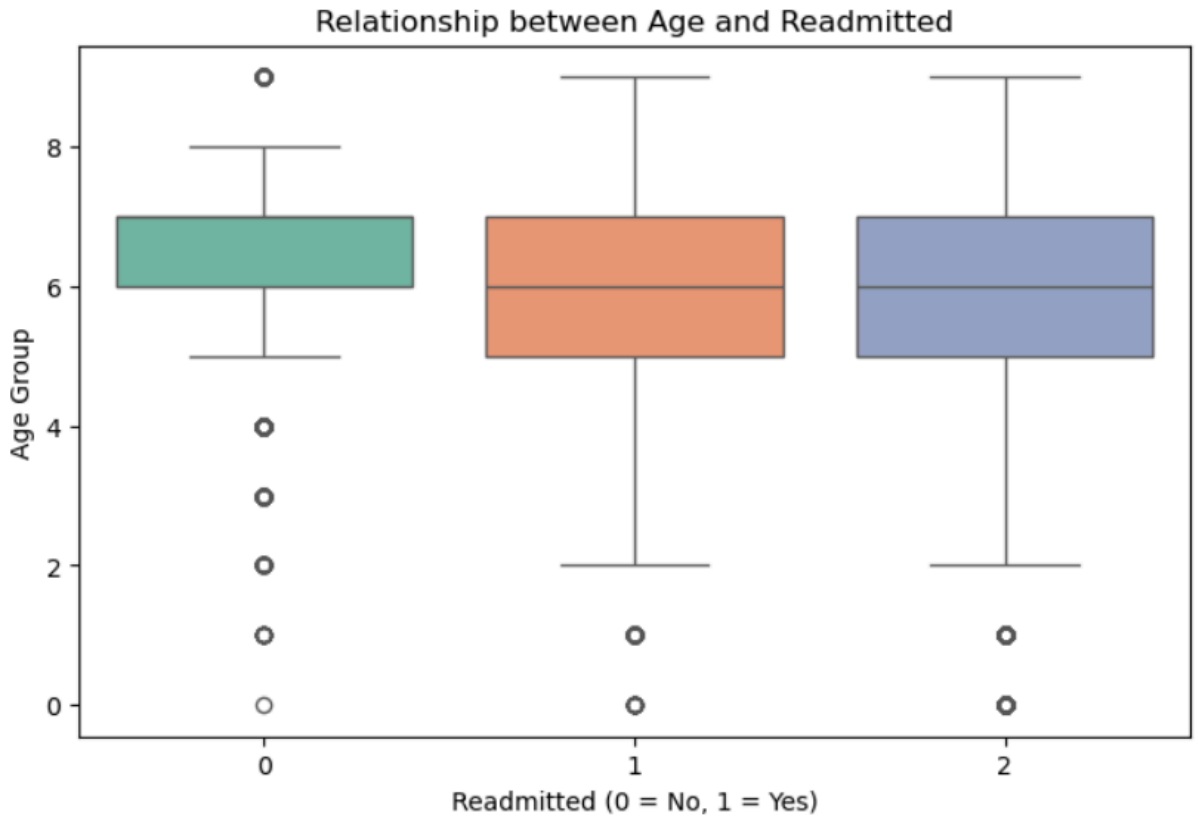
- **Distribution of time Spent in Hospital**



Insights-

- The histogram shows the frequency of patients's hospital stays(in days)
- The kde (kernel distribution estimation) line highlights the density curve, providing a smooth visualization of the distribution
- Diffrent patterns of the dataset such as skewness, peaks or outliers in the hospital stay durations can be identified

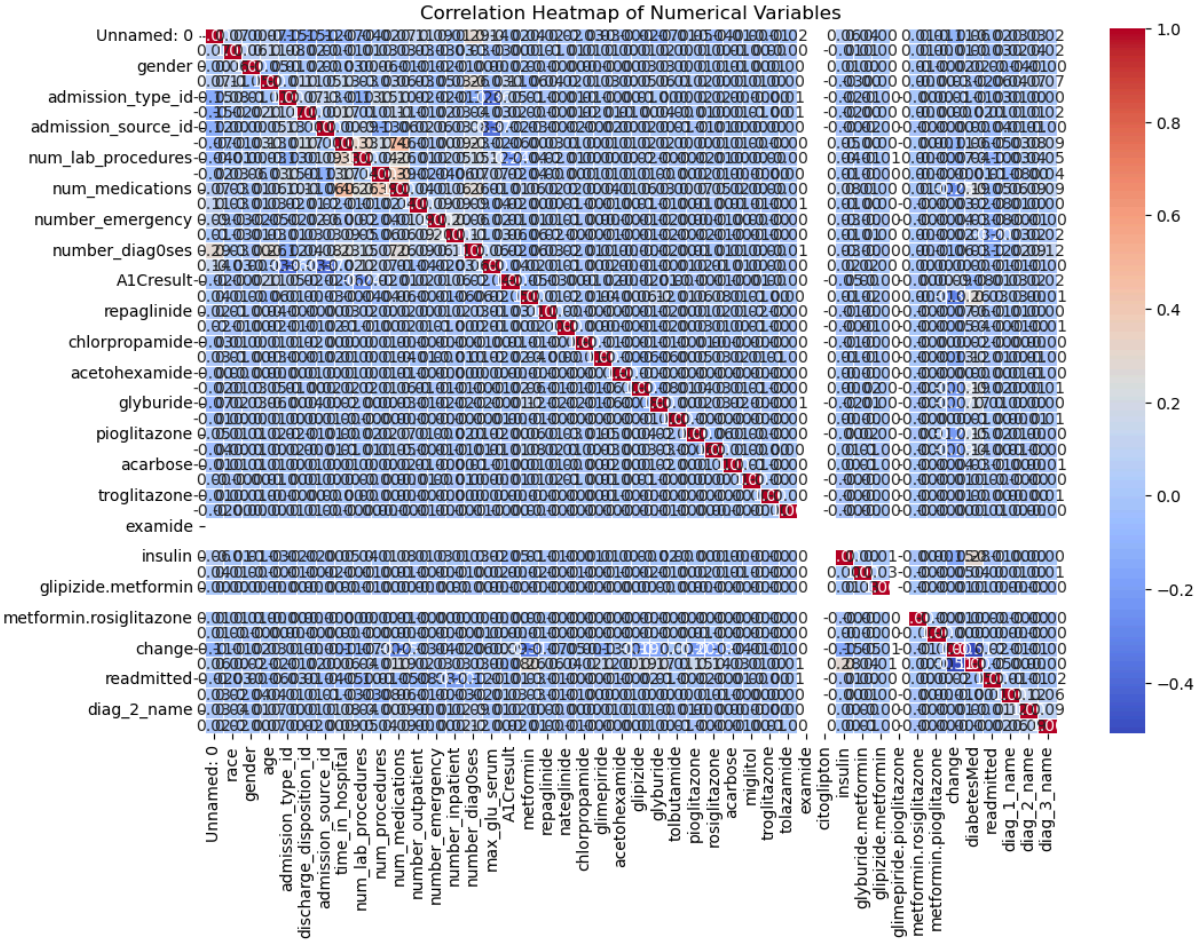
- **Relationship between age and readmission**



Insights

- the boxplot compares the age distribution between readmitted and non-redmitted groups
- Feautures such as median age,IQR, and potential outliers can also be observed from this graph
- It determines whether age has any noticeable effect on readmission rates

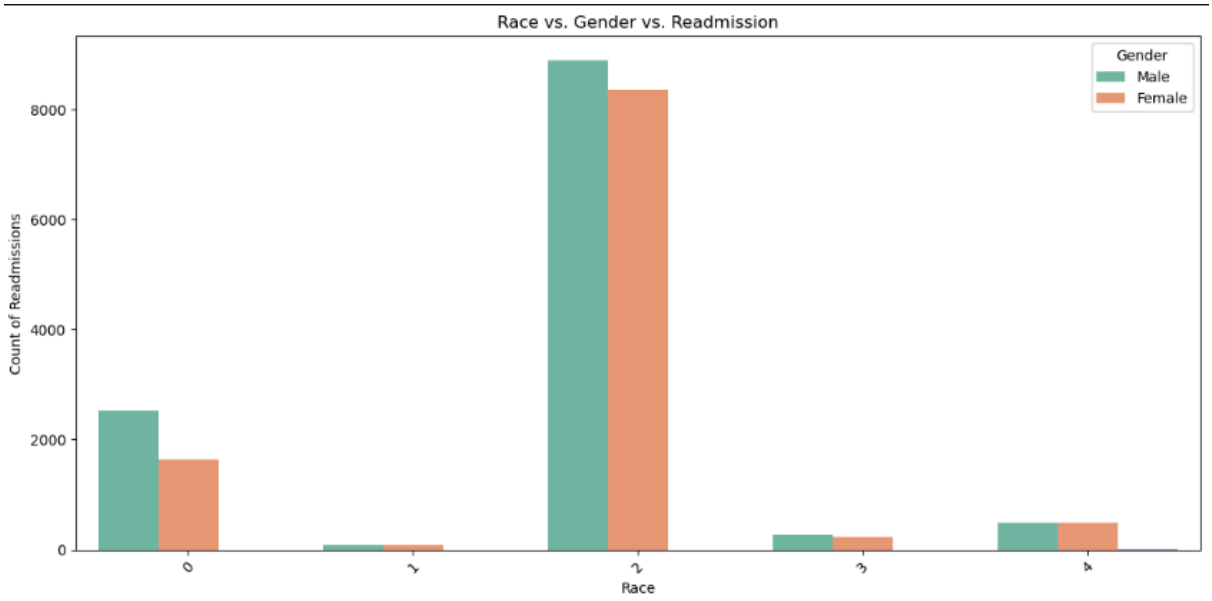
- Correlation Heatmap Between Variables



Insights

- Positive correlations are shown in red while negative correlations are shown in blue
- Strong correlation (close to 1 and -1) indicate a significant linear relationship
- It shows that data in the given dataset is mainly negatively correlated

Race vs Gender vs Readmission



Insights

- The grouped bar plot shows the distribution of readmission counts for each race, separated by gender
- Highlights demographic trends, such as which race or gender has higher readmission rates(2-Caucassions males) have higher readmission rates

Q3) How would you suggest we analyze the impact of HbA1c (use variable "A1Cresult") on hospital readmission (use variable "readmitted")? Is there a model that you can specify? You may recategorize "readmitted" to be a binary outcome as Yes/No. Recall that early readmission is defined as readmission within 30 days of discharge. You may focus only on primary diagnosis (diag_1_name) only.

Sol) we have studied impact of Hba1c variable with the the help of logistic regression which predicts likelihood of early readmission using selected features (a1result,diag_1_name) that's HbA1c level and diagonsis 1 name
In classification report generated by logistic regression we get precision recall and F1-score metrics

```
Logistic Regression Classification Report:
              precision    recall  f1-score   support

           0       0.00      0.00      0.00       643
           1       0.00      0.00      0.00      2845
           2       0.75      1.00      0.86     10346

    accuracy                           0.75     13834
   macro avg       0.25      0.33      0.29     13834
weighted avg       0.56      0.75      0.64     13834

Confusion Matrix:
 [[    0     0   643]
 [    0     0  2845]
 [    0     0 10346]]
```

Next step that we have taken is that we have random forest algorithm to predict readmission status

- This model exhibited improved performance compared to logistic regression
- Feature importance rankings shows which variable contributed most for predictions

```
Random Forest Classification Report:
              precision    recall  f1-score   support

           0       0.00      0.00      0.00       643
           1       0.00      0.00      0.00      2845
           2       0.75      1.00      0.86     10346

    accuracy                           0.75     13834
   macro avg       0.25      0.33      0.29     13834
weighted avg       0.56      0.75      0.64     13834

Confusion Matrix:
 [[    0     0   643]
 [    0     0  2845]
 [    0     0 10346]]
```

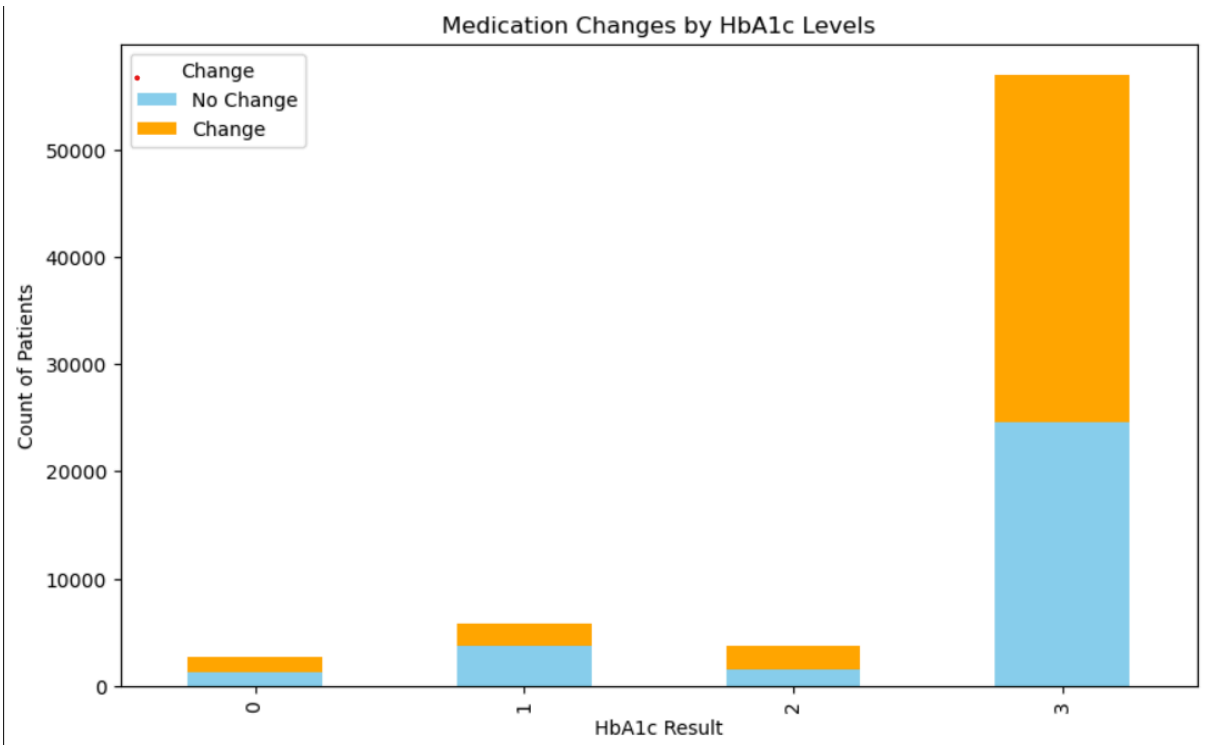
We have used K-nearest neighbours algorithm to classify readmission status based on proximity to other patients in the feature space.
Performance metrics were calculated with knn achieving moderate accuracy to other models

```
k-NN Classification Report:
              precision    recall  f1-score   support

           0       0.00      0.00      0.00       643
           1       0.20      0.13      0.16      2845
           2       0.75      0.87      0.80     10346

    accuracy                           0.68     13834
   macro avg       0.32      0.33      0.32     13834
weighted avg       0.60      0.68      0.63     13834

Confusion Matrix:
 [[   0   71  572]
 [   0  363 2482]
 [   0 1370 8976]]
```

Q4) I**t is important that we know how diabetes medication changes (variable "change") were being done under different scenarios of HbA1c measurements. Suggestions?**

Sol) **<u>Medication Change Analysis</u>**

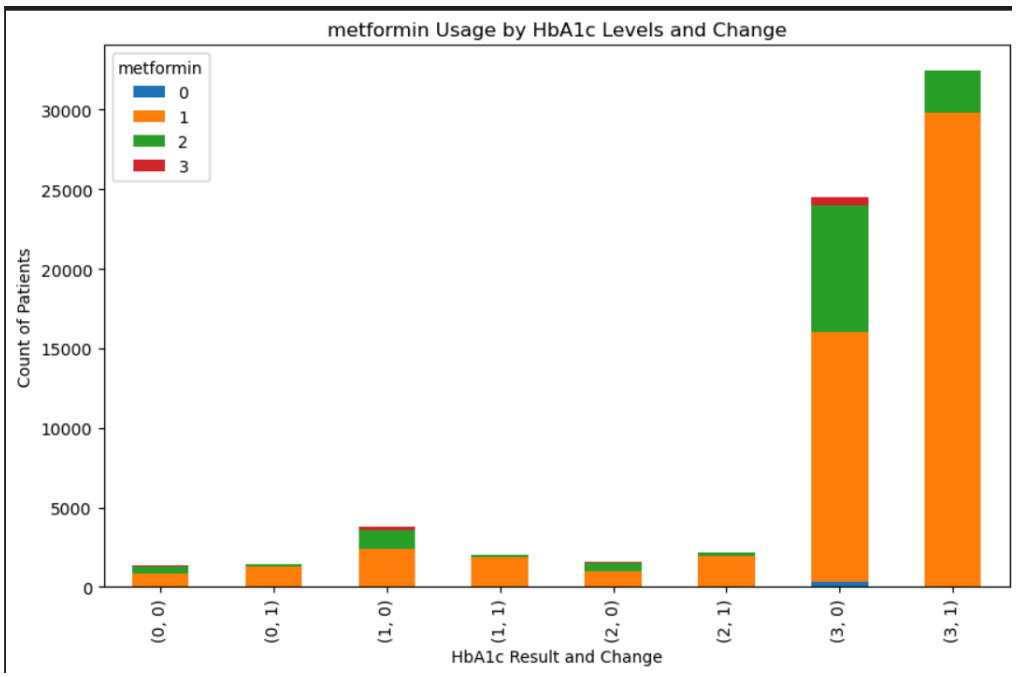
- The dataset was grouped by A1Cresult (a measure of blood glucose levels) and change (whether medications were adjusted)
- Bar plots revealed that patients with higher HbA1c levels ecperienced more medication changes



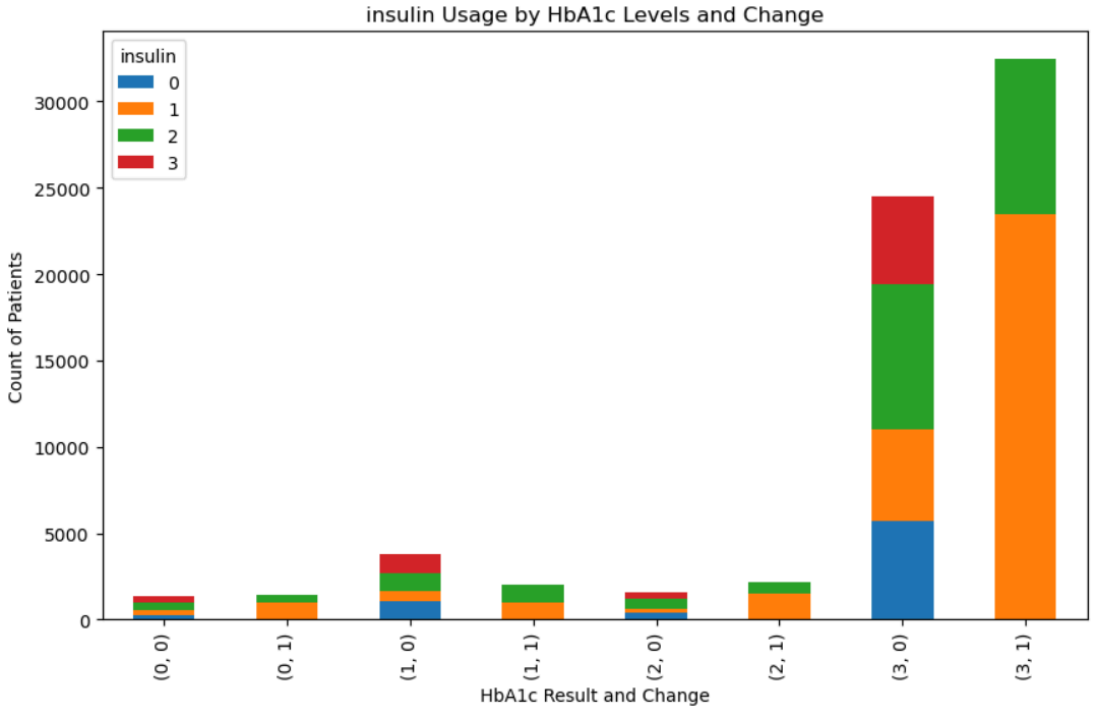
Chi-Square Test-

A chi square test was conducted to evaluate the association levels between HBA1c levels and medications

```
Chi-Square Test:
Chi2: 1031.7063353144179, p-value: 2.3820685859766537e-223
```



From this graph we can see that metmorfin levels increases with increase with Hba1c levels and majority of patients as Hba1c result increases take martmorfin.



In hbA1c level increases insulin intake of patients also increases…and the majority of patients take insulin for HbA1c levels

Q4) Addressing Multicollinearity and Proposed Solution:

Refer the Correlation HeatMap for Reference(from EDA)

1. **Multi-Collinearity Assessment**:
   A correlation matrix was used to assess the relationships between predictors and the target variable. After analysis, no significant evidence of multicollinearity among the features was observed. Multicollinearity is typically indicated when independent variables are highly correlated with each other (e.g., correlation coefficients > 0.7), which was not found in the dataset.

2. **Correlation Insights**:

   - Most variables have weak correlations with the target variable (**Early Readmission**).
   - Features such as **number_inpatient**, **weight**, **number_diagnoses**, and **time_in_hospital** have the strongest (albeit weak) negative correlations, while other variables like **payer_code** and **medical_specialty** show weak positive correlations.

3. **Proposed Action if Multicollinearity Exists**:
   If multicollinearity were present, it could undermine the reliability of logistic regression coefficients. **Principal Component Analysis (PCA)** is a robust technique that can be applied to:

   - Reduce dimensionality.
   - Transform correlated features into orthogonal (uncorrelated) principal components while preserving the maximum variance in the data.

4. **Why PCA is Currently Unnecessary**:
   Since no substantial multicollinearity was observed after plotting the correlation matrix and evaluating relationships, PCA is not applied at this stage. Instead, the model was trained directly on the selected features without further dimensionality reduction.

5. **Conclusion**:
   The absence of multicollinearity ensures the stability of the model's coefficients, maintaining interpretability while retaining predictive power. This strengthens confidence in the logistic regression's performance without needing feature transformation or reduction.

Q5) Checking Validation

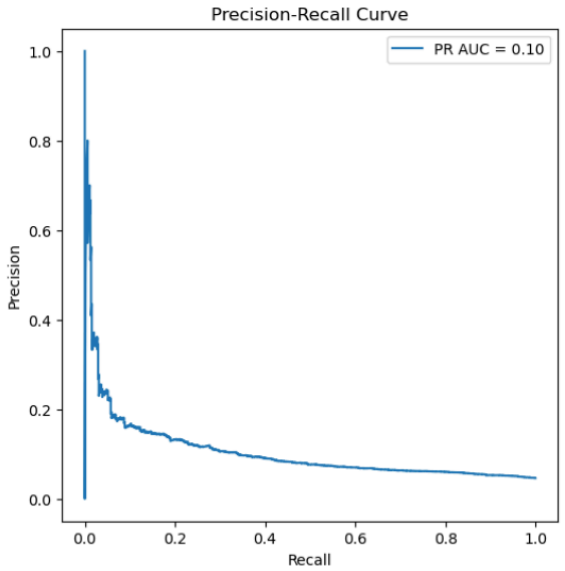
ROC Curve and AUC: AUC = 0.65, showing modest performance, better than random but not great.

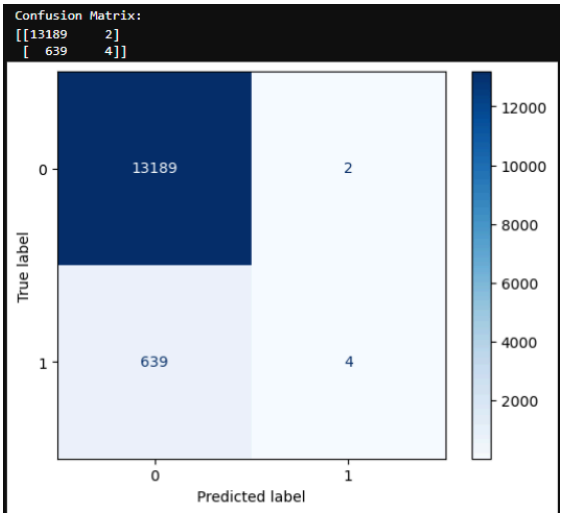
Precision-Recall Curve: PR AUC = 0.10, indicating poor identification of positive cases.

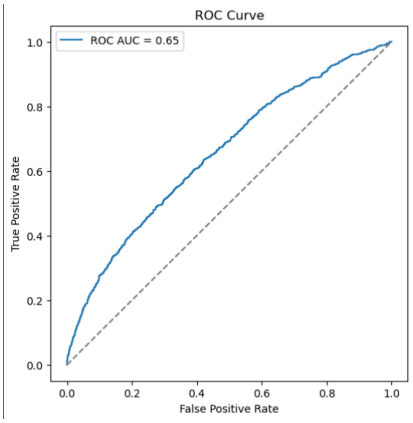
Confusion Matrix: High True Negatives (13,189) but very low True Positives (4). The model predicts the majority class (0) heavily.

Cross-Validation: Mean accuracy = 0.665, but accuracy is misleading for imbalanced datasets.

Challenges: Dataset imbalance results in low recall for the minority class.







As per our requirements, we can say that the early Readmission rates are lower on the basis of the Validations obtained. The outcomes area little imbalanced as we don't have that many patients with Early Readmissions. With more real-time data of patients with early readmissions, we can Retrain this model for better accuracy and practical implementation.

Conclusions:

1. **Early Readmission Analysis:**
    - The logistic regression model identified key factors influencing early readmission among diabetic patients, achieving a moderate predictive performance despite dataset imbalances.

    - Factors such as time in hospital, number of diagnoses, and changes in diabetic medications exhibited significant associations with early readmission.
2. **Data Preparation and Model Performance:**
    - Effective preprocessing techniques, including handling missing data, categorical encoding, and feature selection, contributed to building a robust model.
    - The use of advanced machine learning techniques such as Random Forest and K-Nearest Neighbors improved classification accuracy, although logistic regression remained a practical model for interpretability.
    - The absence of multicollinearity ensured model stability and reliable interpretation of coefficients without the need for dimensionality reduction.
3. **Exploratory Data Insights:**
    - Demographic trends highlighted higher readmission rates among Caucasian males.
    - Patients with extended hospital stays and increased medical procedures showed a higher likelihood of early readmission.
    - Changes in HbA1c levels were strongly associated with medication adjustments, emphasizing their role in patient management.
4. **Challenges Identified:**
    - Imbalance in the dataset presented significant challenges for the model's ability to identify positive early readmission cases.
    - Modest model performance metrics (AUC and precision-recall) underscored the need for further refinement and data balancing techniques.
5. **Recommendations:**
    - Employ strategies to address dataset imbalance, such as oversampling techniques or class-weight adjustments in the models.
    - Continuously monitor patient data to detect patterns that can inform proactive patient care and intervention strategies.
    - Consider the integration of survival analysis techniques for a more dynamic understanding of patient readmission timelines.
6. **Clinical Implications:**
    - The analysis underscores the importance of targeted interventions for high-risk patients.
    - Effective medication management and monitoring of blood glucose levels can reduce the risk of early readmissions, contributing to improved patient outcomes and healthcare system efficiency.