SCHOOL OF COMPUTING AND MATHEMATICAL SCIENCES

UNIVERSITY OF LEICESTER

CW Assignment (spring 2023

Module Code: CO7093

Module: Big Data & Predictive Analysis

Group: 88

Assignment: Regression & Clustering

Date of submission: 27/03/2023

GROUP MEMBERS

Anoushka Kudesia: ak1002

Dhrupal Paresh Shah: dps30

Hetal Solanki: hms40

Mayur Sakharam Shinde: mss62

Payal Jetha Modhvadiya : pjm70

TABLE OF CONENTS

1. Introduction

This report presents the development and evaluation of a predictive model aimed at identifying patients with diabetes at risk of readmission to the hospital within 30 days. This work is part of the coursework for the CO7093 - Big Data & Predictive Analytics module, leveraging data from 130 US hospitals.

2. Data understanding and Preparation

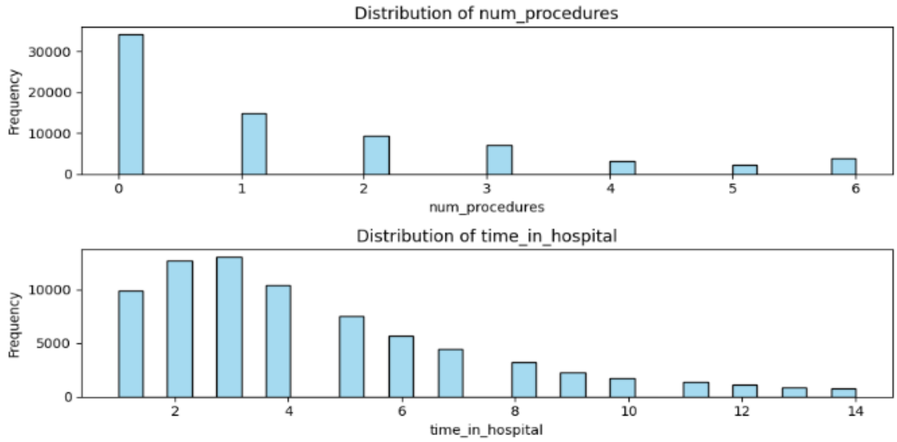
The dataset for this study encompasses records from 101,766 diabetic patients across 130 US hospitals, detailing clinical care over ten years.

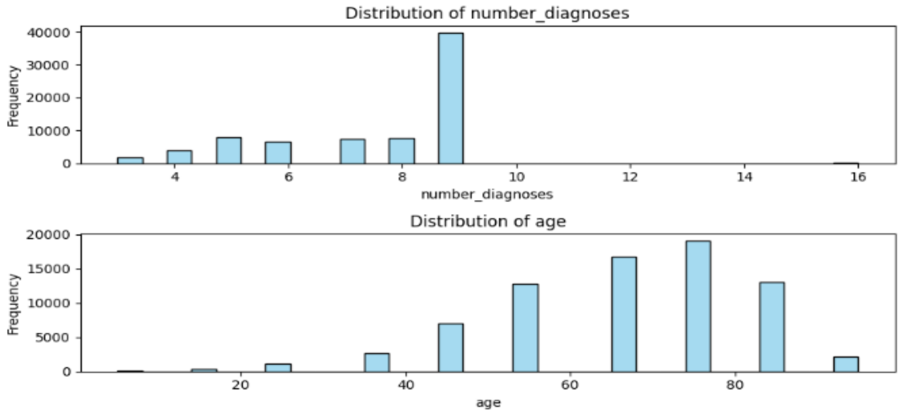
2.a Data Cleaning steps

- Removal of Irrelevant Features - We removed 'encounter_id', 'payer_code', and 'patient_nbr' because these features contained identifiers that are unique to each entry and irrelevant to the analysis, providing no predictive value for the outcome of interest.
- Handling Missing values - Missing values, initially represented by '?', were standardized to NaN. This uniform representation allows for easier identification and handling of missing data, ensuring that subsequent analyses do not misconstrue these missing values as actual data.
- Binary Conversion of Target Variable - The 'readmitted' column was converted to binary format to facilitate our binary classification task. This simplification from three categories to two (readmitted within 30 days or not) aligns with the project's objective and improves model interpretability.
- Data Type Classification - We organized columns by their data types to streamline data handling and ensure appropriate preprocessing techniques were applied to each data type.
- Missing Value Analysis - We evaluated the percentage of missing data per column to determine their viability for inclusion in the analysis. Columns with more than 90% missing values were removed since they are unlikely to contribute meaningful information and could potentially skew the results.
- Removal of Near Zero-Variance Features - We identified and removed features with near-zero variance, such as certain medication indicators, that do not vary significantly across patients. This step simplifies the dataset, reducing dimensionality and computational complexity.
- Filtering out Rows with Unclear or Irrelevant IDs - We identified and filtered out records with ambiguous or irrelevant classifications in 'admission_type_id', 'discharge_disposition_id', and 'admission_source_id'. Categories such as 'Not Available', 'NULL', or 'Not Mapped' were considered non-contributory towards our objective of predicting hospital readmissions due to their unclear nature. This exclusion aligns with our broader data cleaning strategy, aiming to enhance model relevancy and efficiency by concentrating on more definitive and impactful data attributes.
- We streamlined the dataset by excluding rows with ambiguous 'gender' and 'race' values, aligning this step with our overall strategy to enhance clarity and relevance. Additionally, missing 'medical_specialty' values were replaced with 'Unknown', maintaining data integrity while simplifying analysis. Finally, we removed all remaining rows with null values, reducing complexity and ensuring a cleaner dataset for modeling.
- In our dataset normalization process, we encoded medication adjustments, clinical results, and binary attributes numerically to ensure analytical consistency. Medications like 'metformin' and features such as 'A1Cresult' were transformed based on their clinical significance. Additionally, we categorized ICD diagnosis codes to simplify medical conditions into broader groups, aligning with our strategy to reduce complexity. These steps, alongside the removal of records with incomplete information, streamline the dataset, as confirmed by the final structure in normalized_diabetic_data.shape, enhancing our analysis readiness.

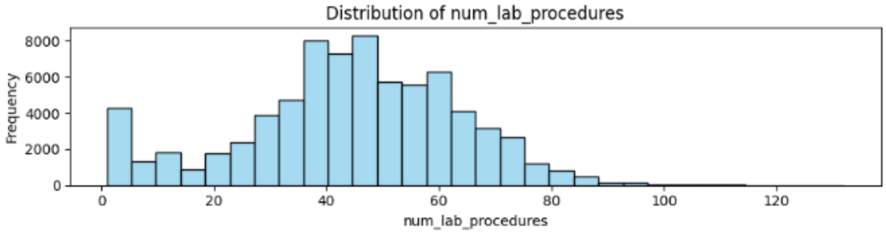
- Removing Outliers - In the data preprocessing phase of our analysis, we adopted various strategies to identify and remove outliers from our dataset. This was essential to ensuring the accuracy and reliability of our subsequent predictive modeling.

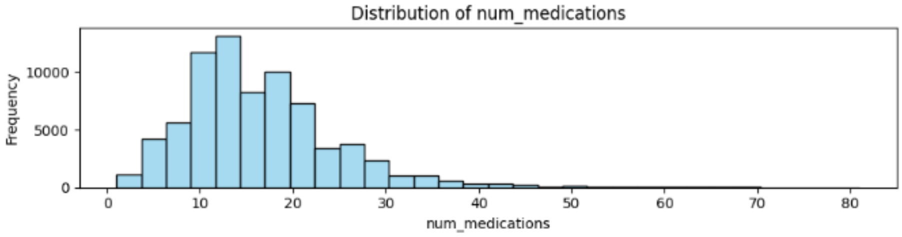
  The Interquartile Range (IQR) method was utilized for features such as 'num_procedures', 'number_diagnoses', 'time_in_hospital', and 'age'. Graphical analysis revealed these variables exhibited skewed distributions with potential extreme values. By employing the IQR method, which calculates the range between the first and third quartiles (Q1 and Q3) of the data, we were able to define a "normal" data range and remove data points that lay outside 1.5 times the IQR above Q3 and below Q1. This ensured the retention of data central to the distribution while discarding outliers that could skew our analysis and model performance.



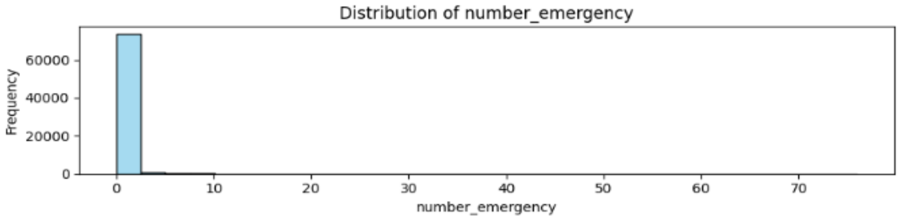


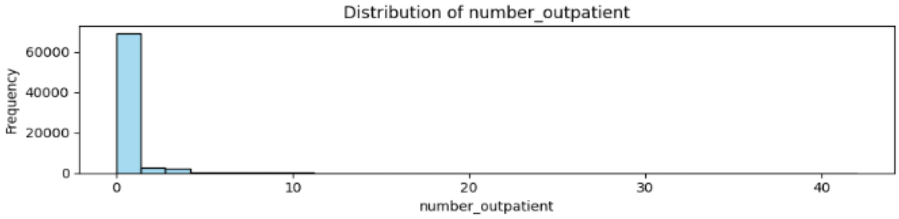
For 'num_lab_procedures' and 'num_medications', our graphical analysis suggested a more symmetrical, bell-shaped distribution, leading us to apply the Z-score method. This approach identified outliers as those data points lying more than three standard deviations from the mean. By removing these outliers, we aimed to negate the influence of extreme variations that are not representative of the general patient population, as indicated by the smoother, more normalized distributions observed in our histograms post-cleanup.

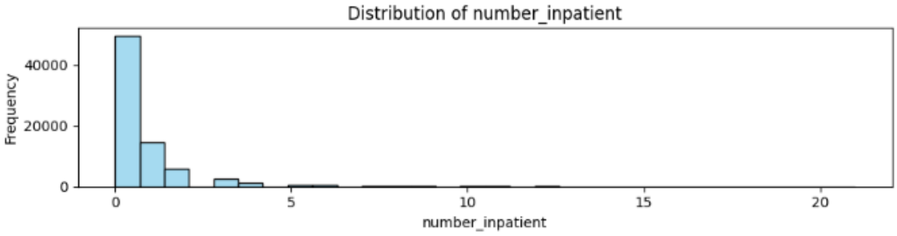




Lastly, the Percentile method was chosen for 'number_emergency', 'number_inpatient', and 'number_outpatient'. The graphical representations for these variables highlighted long tails, suggesting the presence of extreme values concentrated at the higher end of the scale. By removing data beyond the 5th and 95th percentiles, we focused our analysis on the most consistent and representative portion of the data, significantly reducing the skewness and enhancing the robustness of our findings.







Through these strategies, corroborated by our graphical observations, we ensured that our dataset's integrity was maintained while minimizing the potential biases introduced by outliers. This meticulous approach to outlier management has paved the way for more accurate and reliable subsequent analyses, providing a solid foundation for our predictive modeling endeavours.

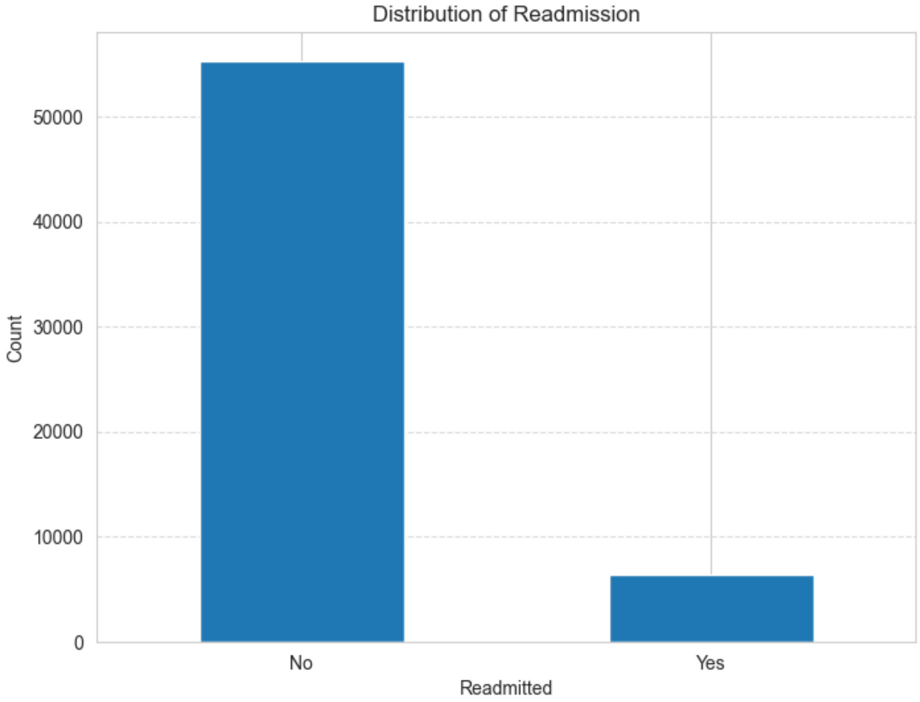
2.b Visualization

In our analysis, we initially plotted graphs to compare the rates of readmission across various categories and features within the dataset. To gain further insight and ensure the robustness of our observations, we then re-plotted these graphs using a subset of the data, specifically the top 80%, to focus on the most representative and recurrent instances.

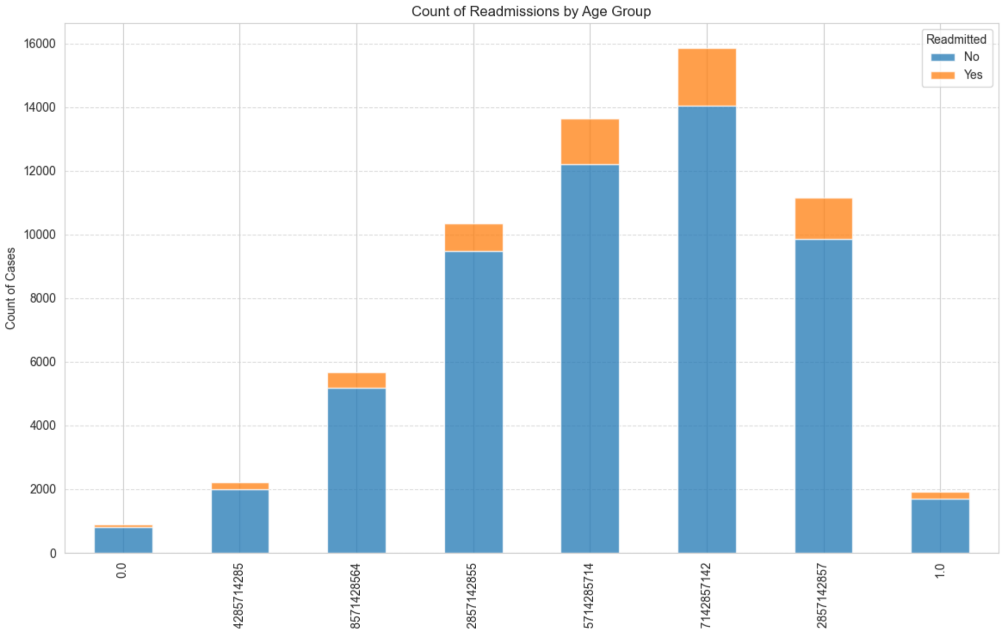
These comprehensive visualizations, which can be found detailed within the accompanying Python notebook, allow us to observe trends and outliers effectively. Here are the distilled observations from this two-tiered graphical analysis:

- **Race Distribution:** Most of the dataset consists of Caucasian patients. African American patients follow but with significantly fewer numbers. However, when viewed as percentages, the readmission rates among different racial groups might paint a different picture, potentially highlighting disparities in readmission rates that are not solely due to population size.
- Gender Distribution: While absolute numbers might show one gender having more readmissions, the percentage distribution could reveal that, relative to their representation in the dataset, both genders have similar readmission rates. This suggests that gender might not be a significant standalone factor in readmission risk when normalized for population size.
- Age Distribution: Analysis illustrates a clear trend; older age groups tend to have higher absolute numbers of readmissions. However, the percentage distribution helps to understand if older age groups are disproportionately affected or if the higher readmission rates are reflective of the larger population sizes within those age brackets. This can inform healthcare providers about the necessity of age-specific healthcare programs to mitigate readmission risks.
- Admission Type ID: Absolute figures might highlight certain admission types as more common; yet percentage analysis reveals which admission types correlate with higher readmission rates. This distinction is vital for understanding how specific types of hospital admissions impact the likelihood of patients returning, guiding improvements in admission practices and patient triage.
- Discharge Disposition ID: In absolute terms, specific discharge dispositions appear linked to higher readmissions, but percentage-wise analysis provides insight into how these dispositions stand relative to the total cases. This could influence post-discharge care plans, indicating that not just the volume but the nature of discharge plays a critical role in patient outcomes.
- Admission Source ID: Admission source 7 dominates the dataset. Though source 7 may send more patients leading to higher absolute readmission numbers, the percentage distribution sheds light on which admission sources pose a higher readmission risk. Understanding this helps tailor interventions aimed at reducing readmissions from high-risk sources.
- Medical Specialty: There is a broad range of readmission rates across medical specialties when viewed as percentages. Some specialties have notably higher percentages of readmissions, indicating areas where patient management may require additional focus or improvement. However, most specialties show a predominant percentage of non-readmitted cases, suggesting effective management in those areas.
- Diagnosis Categories: The bar plots show different patterns for diagnosis categories. Some conditions have higher readmission rates when we look at percentages. This means even if few patients have these conditions, they often come back to the hospital. Other conditions are more common but do not always lead to more readmissions.
- Diabetes Medication: The graphs for Diabetes Medication compare patients taking these drugs to those who are not. The percentage plot helps us see if being on medication affects the chance of coming back to the hospital. It looks like whether patients are on diabetes medicine or not changes their readmission rates.

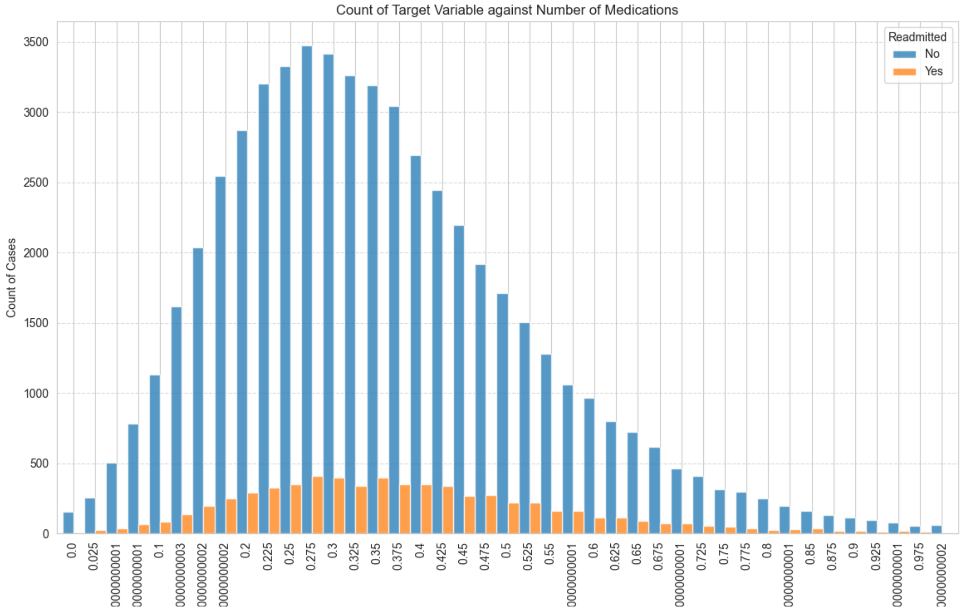
Distribution of Readmission: A clear disparity is evident between the numbers of readmitted and non-readmitted patients, with a significantly higher count of patients not experiencing readmission. This suggests that while a substantial portion of patients manage to avoid readmission, there remains a considerable number who do not, underscoring the need for targeted readmission reduction strategies.



Count of Readmissions by Age Group: This graph shows a significant increase in both readmitted and non-readmitted cases among middle-aged and older patients. The higher readmission rates in these age groups could reflect the increased healthcare needs and complexities associated with aging.



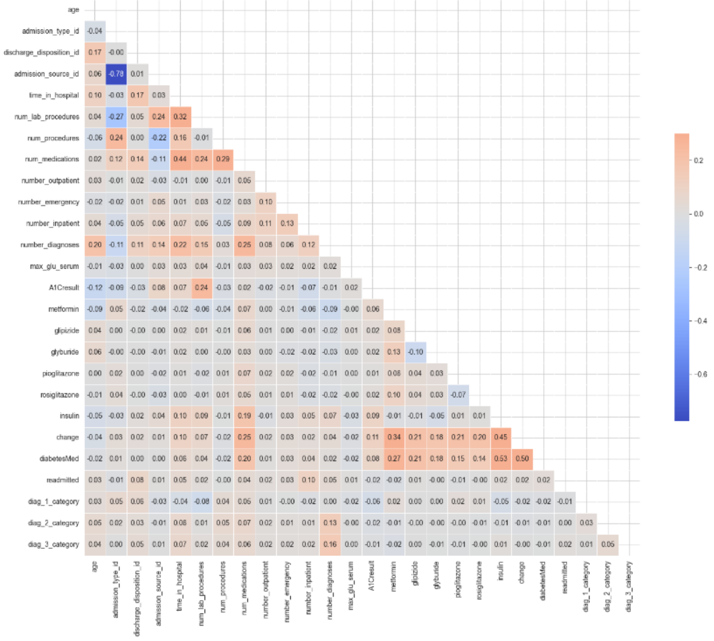
Count of Target Variable against Number of Medications: The distribution indicates that patients taking a higher number of medications tend to have a higher count of non-readmissions than readmissions. However, a considerable number of patients with fewer medications still face readmissions, pointing towards medication management as a potential area for intervention.



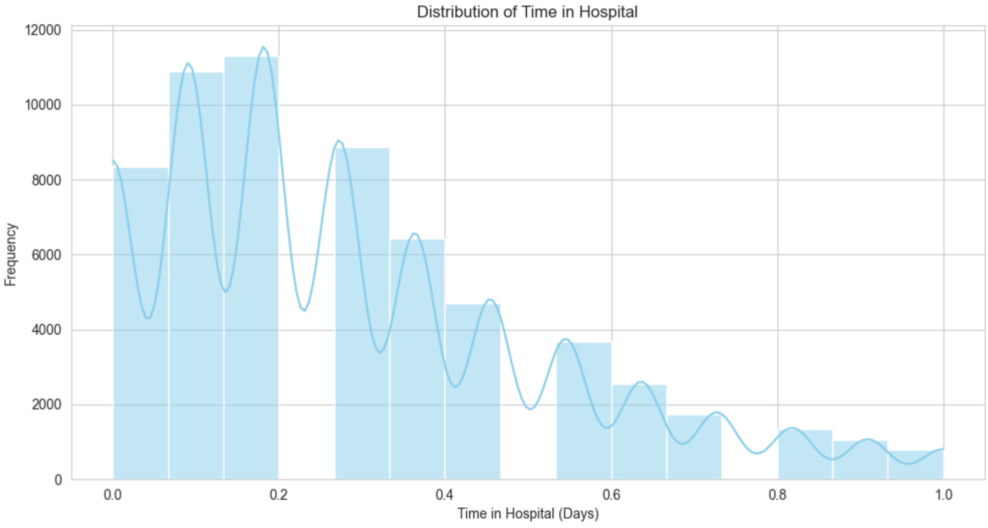
Correlation Matrix: The heatmap shows the correlation coefficients between several factors such as age, different IDs (admission_type_id, discharge_disposition_id, admission_source_id), metrics related to hospital stay (time_in_hospital, num_lab_procedures, etc.), medication usage, and readmission rates. In a heatmap like this, the colour scale typically ranges from blue (indicating negative correlations) to red (indicating positive correlations), with varying intensities reflecting the strength of the relationship.

- **Age and Time in Hospital:** There seems to be a slight positive correlation between age and time spent in the hospital, suggesting older patients might tend to have longer hospital stays.
- **Admission Source and Emergency Visits:** The heatmap indicates a significant positive correlation between the admission source and emergency visits, which might suggest that certain sources of admission are more likely to be associated with emergencies.
- **Number of Lab Procedures and Time in Hospital:** There is a positive correlation here, implying that longer hospital stays could be associated with a higher number of lab tests performed, which could be indicative of more severe or complex conditions.
- **Number of Medications and Number of Diagnoses:** A noticeable positive correlation exists, suggesting that patients with more diagnoses tend to be on more medications, which is logical considering more conditions typically require more treatments.
- **Readmissions:** Several factors show varying degrees of correlation with readmission rates, such as time in hospital, number of inpatient visits, and medication changes (change). These relationships hint at potential areas to explore for reducing readmission rates, such as managing length of stay, reducing inpatient visits, and stabilizing medication regimens.

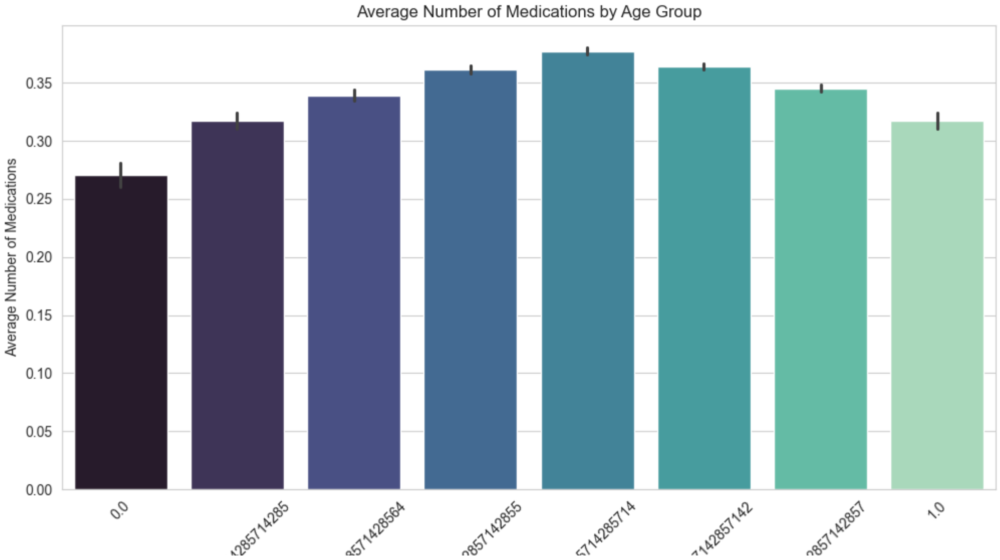
- **Medication Types:** Specific medications like insulin and glyburide show correlations with other factors, potentially indicating common treatments for the patient population within the dataset.



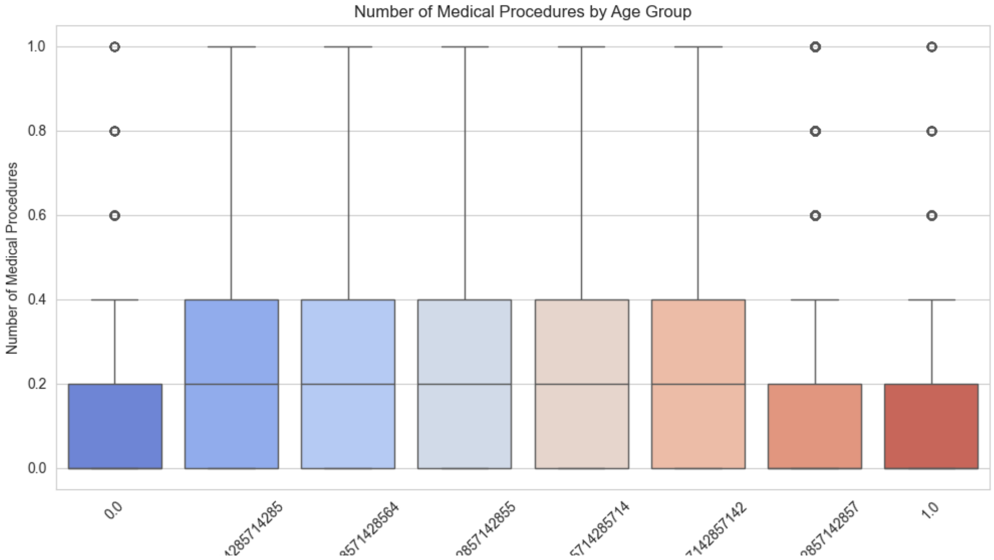
Distribution of Time in Hospital: This shows a varied length of hospital stays among patients, with peaks at certain points. Understanding whether longer or shorter hospital stays are associated with higher readmission rates could be crucial for developing effective discharge planning and post-hospital care.



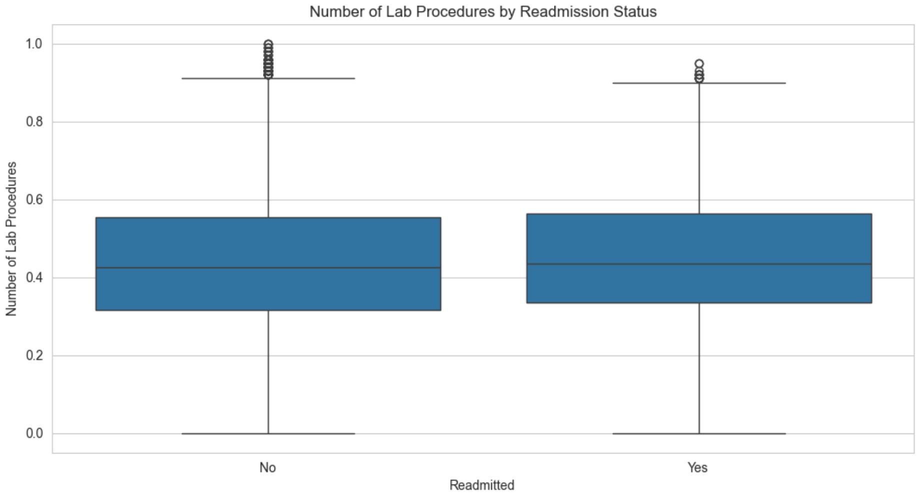
Average Number of Medications by Age Group: The graph indicates that the average number of medications tends to slightly increase with age, suggesting older patients are prescribed more medications. This trend could imply a higher complexity of treatment as age increases, potentially influencing readmission rates.



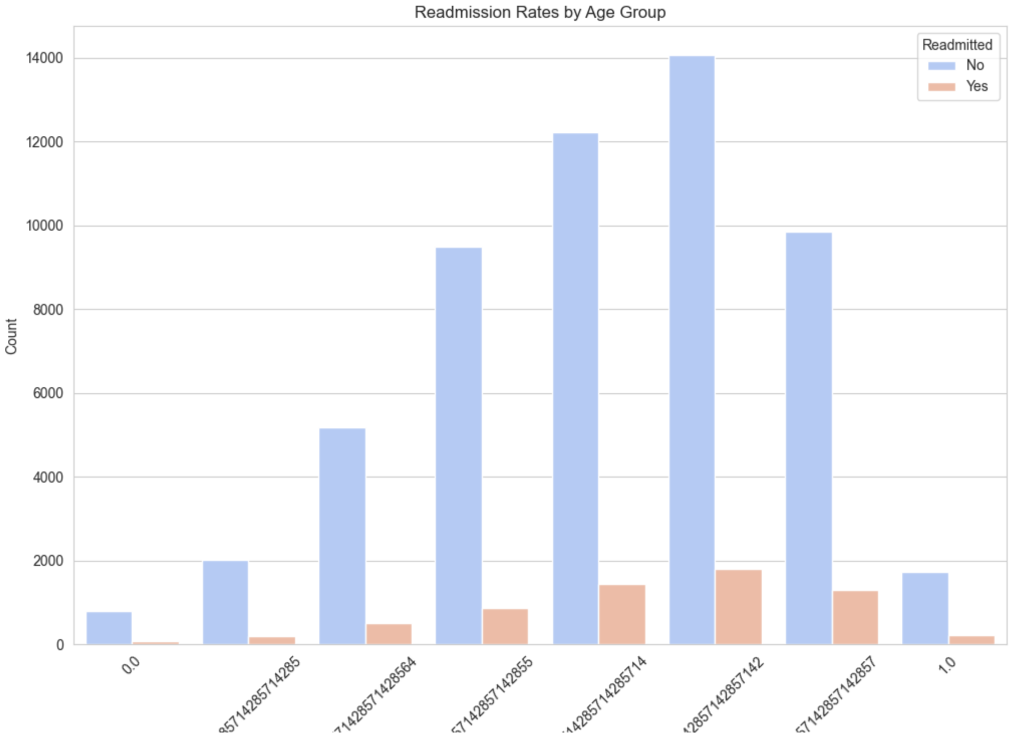
Number of Medical Procedures by Age Group: The distribution across age groups does not show a significant increase in the number of medical procedures with age, which contrasts with the trend seen in medication prescribing. This suggests that while older patients may receive more medications, they do not necessarily undergo more medical procedures.



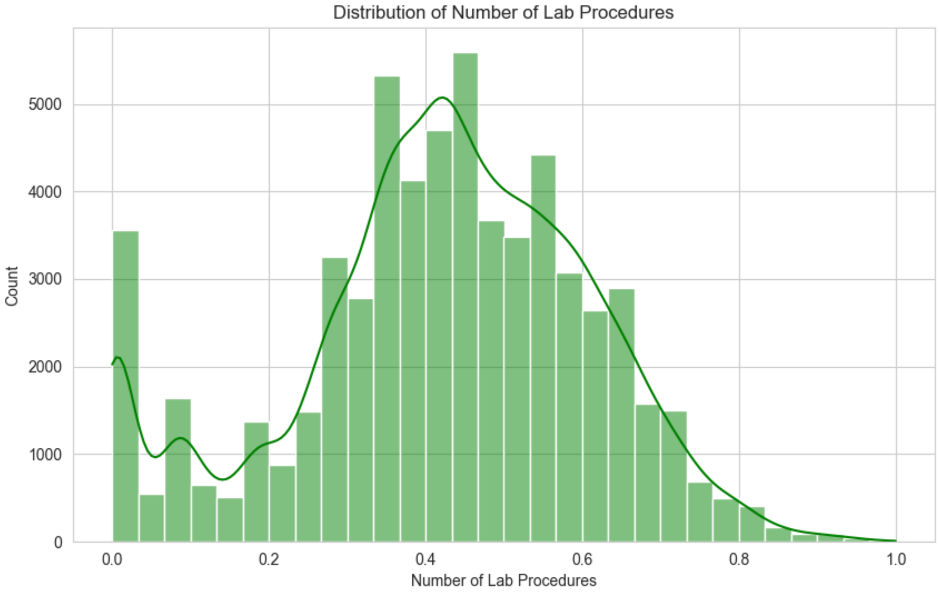
Number of Lab Procedures by Readmission Status: Comparing readmitted to non-readmitted patients, there seems to be a similar distribution in the number of lab procedures, suggesting that the quantity of lab tests alone may not be a significant predictor of readmission.



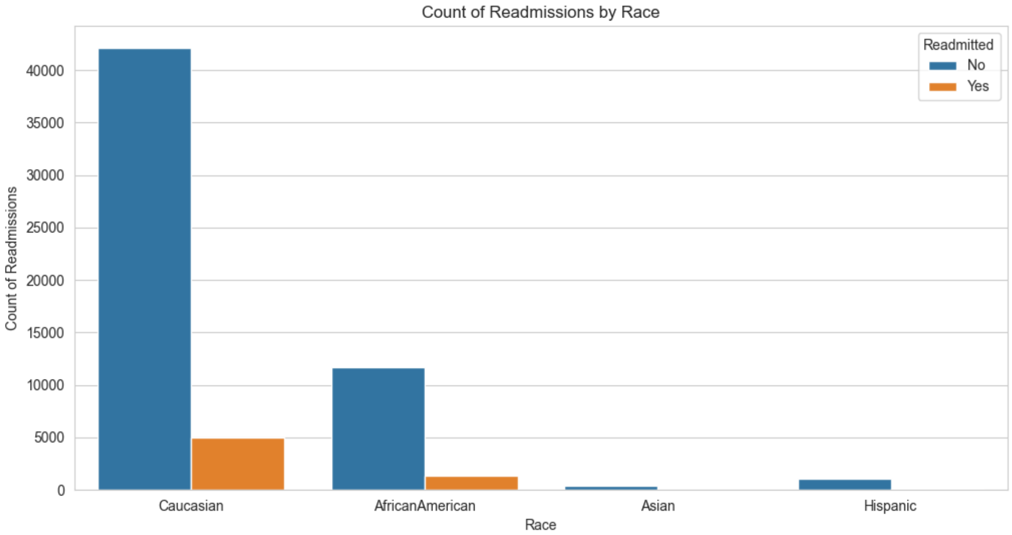
Readmission Rates by Age Group: The rates increase with age, particularly among the older demographics, aligning with the notion that older age groups are more susceptible to readmission, due to more complex health issues.



Distribution of Number of Lab Procedures: The distribution highlights a wide range of lab procedures undertaken by patients, with a peak in lower-range procedures. Understanding the relationship between the number of lab procedures and readmissions could help identify whether extensive testing correlates with better healthcare outcomes.



Count of Readmissions by Race: The data demonstrates that Caucasian patients have a significantly higher count of both readmissions and non-readmissions compared to other races. This disparity may indicate differences in healthcare access, utilization, or underlying social determinants of health affecting these populations.



3. Model Implementation and Evaluation

3.a. Model Selection and Predictors:

For our study, we selected specific predictors based on their correlation with the 'readmitted' variable. These predictors include 'number_inpatient', 'number_diagnoses', 'time_in_hospital', 'num_medications', 'number_emergency', 'age', 'num_lab_procedures', 'number_outpatient', 'A1Cresult', 'metformin', and 'insulin'. We focused on variables with at least 0.02 percent correlation, adjusting our model based on empirical evidence to improve accuracy.

3.b. Data Balancing Technique:

Given the imbalanced nature of our dataset, we implemented an ensemble method to enhance model performance. This involved splitting the data where 'readmitted == 0' into equal subsets, then combining each '0' subset with the '1' subset to create balanced training sets. This approach allowed us to train our model on balanced data, mitigating the skewness towards a particular class.

3.c. Model Training Process:

The training process involved several steps:

- Splitting the data into training and testing sets.
- Scaling the features for normalization.
- Initializing and training a logistic regression model with a maximum iteration limit of 1000.
- Storing and evaluating each model trained on different subsets.
- Aggregating the performance metrics from each trained model.

3.d. Evaluation Metrics:

We evaluated our models using accuracy, precision, recall, and F1-score, considering their importance in understanding model performance, especially in the context of medical predictions. Here are the results aggregated from the models trained on various balanced subsets:

- **Accuracy**: The models achieved an average accuracy of 60.15% with a standard deviation of 0.91%. While this accuracy may seem moderate, it is significant given the balanced nature of the dataset used for training.
- **Precision**: The models demonstrated an average precision of 61.37% for class '0' and 58.09% for class '1', with standard deviations of 0.53% and 1.50%, respectively. This indicates a higher ability to identify non-readmitted cases correctly.
- **Recall**: The average recall was 71.44% for class '0' and 46.77% for class '1', with standard deviations of 1.94% and 0.65%, respectively. This suggests the models are more effective at identifying true positives in the non-readmitted group than in the readmitted group.
- **F1 Score**: The mean F1 scores were 66.01% for class '0' and 51.81% for class '1', with standard deviations of 1.11% and 0.56%, respectively. F1 scores combine precision and recall, providing a balanced view of model performance.

3.e. Conclusions and Future Work:

The model evaluation shows a respectable performance considering the challenges associated with imbalanced medical datasets. Future work could explore more sophisticated ensemble methods, incorporate additional predictors, or apply alternative algorithms to enhance model accuracy and reliability, especially for the readmitted class.

Metric	Mean	Standard Deviation
Accuracy	60.15	0.91
Precision Class 0	61.37	0.53
Precision Class 1	58.09	0.15
Recall Class 0	71.44	0.19
Recall Class 1	46.77	0.65
F1 Score Class 0	66.01	1.11
F1 Score Class 1	51.81	0.56

These metrics represent the performance of the logistic regression models trained on the balanced subsets from the dataset, with each metric providing a unique perspective on the models' performance.

4. Improved Model Report: Diabetic Readmission Prediction

4.a. Data Preprocessing:

- The dataset initially contained patient admission data relevant to diabetes.
- Irrelevant columns such as 'encounter_id', 'payer_code', and 'patient_nbr', as well as columns with near-zero variance, were removed.
- Missing values represented as '?' were identified and appropriate handling strategies were applied, such as replacing them with NaN or specific values.
- Columns with more than 90% missing values were dropped, and specific strategies were employed for handling missing data in 'race' and 'medical_specialty' columns.
- The 'age' column was normalized to reflect the average of the age ranges.
- Diagnostic codes (diag_1, diag_2, diag_3) were categorized into broader categories for analysis.
- Outlier detection and removal were performed using methods based on IQR, Z-score, and percentile for various columns.

4.b. Data Transformation and Feature Engineering:

- Certain medications and other categorical variables were numerically encoded to reflect changes in medication dosage and other relevant conditions.
- The target variable 'readmitted' was converted into a binary format to distinguish between patients readmitted within 30 days and those who were not.
- One-hot encoding was applied to the dataset to convert categorical variables into a format suitable for model training.

4.c. Model Building and Evaluation:

- The processed data was balanced using the SMOTE technique to address class imbalance.
- The dataset was split into training and testing sets, with 80% used for training and 20% for testing.
- A K-Nearest Neighbors (KNN) classifier was chosen as the prediction model, with the number of neighbours set to 4.
- The model's performance was evaluated based on accuracy, precision, recall, F1-score, and ROC AUC score.

4.d. Results:

- The final model showed a balance between sensitivity and specificity as indicated by the classification metrics.
- The heatmap of the correlation matrix provided insights into the relationships between different features and the target variable.

Metric	Score
Accuracy	85.00
Precision (Class 0)	98.00
Precision (Class 1)	78.00
Recall (Class 0)	72.00
Recall (Class 1)	98.00
F1 Score (Class 0)	83.00
F1 Score (Class 1)	87.00
ROC AUC Score	85.34

4.e. Conclusion and Future Work:

- The improved model demonstrated a systematic approach to addressing data imbalance and enhancing model performance.
- Future work could focus on experimenting with different classifiers, hyperparameter tuning, and further feature engineering to improve prediction accuracy.