Using Machine Learning for Prediction of Early Readmission of Diabetic Patients

Final Report

CIND 820- Big Data Analytics Project

Supervised by: Ceni Babaoglu

Presented by: Nehal Gamal Mohamed (501278190)

Ryerson University

Table of Contents

Revised Abstract

Diabetes management poses a significant challenge in healthcare, with patient readmission within 30 days of discharge serving as a critical metric for assessing care quality. Despite advances in preventive interventions, many diabetic patients experience readmissions due to suboptimal glycemic control and inadequate care.

This study will try to predict the likelihood of early readmission (within 30 days) for patients diagnosed with diabetes using clinical data collected over a ten-year period (1999-2008) from 130 US hospitals and integrated delivery networks.

The study objective is to answer the following questions:

- By using machine learning models can we accurately predict early readmission of diabetic patients?
- What are the patient and hospital factors which strongly influence early readmission?
- By using predictive models how can we improve diabetes management and reduce readmission rates?

The dataset has patient records from diabetic encounters, including demographic information, medical history, admission details, laboratory results, medications given, and hospital outcomes. It has 101,766 instances and 50 features, the data is multivariate, consisting of categorical and integer variables.

The study will commence with preparing the dataset, then employing classification techniques to predict early readmission outcomes based on patient and hospital features. Machine learning algorithms including logistic regression, decision trees, and random forests will be utilized.

After splitting the dataset and training the model, the performance will be evaluated using standard metrics such as accuracy, precision, recall and F1 score. Moreover, feature selection and dimensionality reduction techniques will be applied to identify the most informative factors and enhance model performance. Additionally, clustering algorithms will be explored to check if there are hidden patterns that could reveal any patient subgroups with different readmission risks.

Python programming will be used for implementation, to take advantage of various python libraries and tools.

Introduction

Hospital readmission, especially among diabetic patients, creates a challenge due to its implications for patient outcomes and healthcare resources. According to the Centers for Medicare & Medicaid Services (CMS), nearly one in five Medicare patients discharged from a hospital is readmitted within 30 days, resulting in approximately $26 billion in annual costs. Diabetes, affecting over 34 million Americans, frequently leads to complications that necessitate hospital readmissions.

The COVID-19 pandemic has deepened this issue, as diabetic patients are at a higher risk for severe complications from COVID-19. Studies have shown that the pandemic has led to an increase in hospital admissions and readmissions among diabetic patients due to the virus's direct and indirect effects. For instance, research indicates that diabetic patients with COVID-19 are more likely to experience severe outcomes, including hospital readmission, due to worsened glycemic control and increased inflammation. A study published in Diabetes Care found that COVID-19 patients with diabetes had a 30-day readmission rate of 10.4%, significantly higher than the general population.

Readmissions not only indicate potential issues in the quality of care but also place a significant financial burden on the healthcare system. Studies have shown that diabetic patients are twice as likely to be readmitted compared to non-diabetic patients. Common factors contributing to readmissions include inadequate post-discharge care, medication non-adherence, and the presence of comorbid conditions. The pandemic has further highlighted these issues, as the disruptions in routine care, delayed medical attention, and increased psychological stress have contributed to poorer health outcomes for diabetic patients.

Addressing these issues through effective management and predictive modeling could potentially reduce the high rates of readmission, thereby improving patient outcomes and reducing healthcare costs. By utilizing machine learning techniques to predict readmissions, healthcare providers can identify high-risk patients early and implement targeted interventions to improve care quality, particularly during challenging times like the COVID-19 pandemic.

Methodology and Contribution of Current Study

This study builds on the existing body of research by integrating advanced Machine Learning (ML) techniques and comprehensive data preprocessing methods to enhance predictive accuracy. By using a robust, large dataset and employing comprehensive feature selection, along with exploring clustering algorithms to check for hidden patterns that could reveal patient subgroups with different readmission risks, the study aims to provide actionable insights for healthcare providers.

Moreover, previous studies have addressed the same problem, with some utilizing the same dataset. These studies employed a variety of tools such as R statistical software, Hadoop and Spark. However, this study utilizes Python programming and its libraries, providing a unique approach to the analysis. The study is valuable in the context of previous studies because it addresses existing gaps and leverages Python programming for enhanced predictive accuracy and healthcare impact.

Comprehensive Findings from Reviewed Research Papers

As mentioned before diabetic patients have higher readmission rates compared to the general population, often due to poor disease control and complications. Reducing readmissions can significantly lower healthcare costs and improve patient outcomes. Various ML models, such as Random Forest (RF), Naive Bayes (NB), and decision tree, have been used to predict 30-day readmissions. RF models have consistently shown superior performance in predicting readmissions. Important features include patient demographics (age, sex, race), clinical factors (number of diagnoses, length of stay, medication use), and healthcare utilization patterns (number of inpatient admissions, emergency visits).

Although RF models outperformed other algorithms in predicting readmissions, there was a potential for overfitting which needs careful management. RF models were robust in handling various types of data, and their ability to provide feature importance measures made them a preferred choice. Thorough data preprocessing is crucial for building accurate predictive models, including handling missing values, normalizing data, and selecting relevant features. Studies emphasized removing attributes with high missing values or irrelevant features to improve model performance. Techniques like down-sampling and over-sampling Synthetic Minority Oversampling Technique (SMOTE) are employed to address class imbalances in the datasets, enhancing model accuracy.

Overview of Previous Research Papers

Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records

This study highlighted the importance of HbA1c measurement in predicting readmissions among diabetic patients. The analysis revealed that better diabetes management, reflected in HbA1c levels, could potentially reduce readmission rates. ML models were used to identify high-risk patients, focusing on the influence of HbA1c levels. However, the study focused on a single predictor (HbA1c) and lacked the need for incorporating multiple predictors to enhance model accuracy.

Implementation of Big Data Analytics on Diabetes 130-US Hospitals for Year 1999-2008 for Predicting Patient Readmission & The 30-days Hospital Readmission Risk in Diabetic Patients: Predictive Modeling with Machine Learning Classifiers

Both studies utilized the same Health Facts Database to develop predictive models for 30-day readmissions. Various ML algorithms, including RF, NB, and decision trees, were employed, with RF showing the best performance. Key predictors identified include race, sex, age, admission type, admission location, length of stay, and drug use. Both studies concluded that RF is more suitable for making readmission predictions and emphasized the importance of identifying high-risk patients to reduce the probability of readmission within 30 days. These studies had a comprehensive approach and used a large dataset, but the potential for overfitting with RF models needs to be considered carefully. The focus on multiple predictors enhances the model's accuracy, making it more applicable in real-world settings. However, findings need to be validated in clinical settings.

Hospital Readmission and Length-of-Stay Prediction Using an Optimized Hybrid Deep Model

The study introduced the Genetic Algorithm-Optimized Convolutional Neural Network GAOCNN model, a hybrid deep learning model, for predicting readmissions and length of stay. GAOCNN proved robust to missing values and performed well on imbalanced data. Future improvements suggested including optimizing feature extraction and classifier training time. The complexity and computational intensity of the model may limit its practical application in clinical settings, and further research is needed to streamline the model for real-time predictions.

Effective Hospital Readmission Prediction Models Using Machine-Learned Features

The study showed that combining machine-learned features with manual features improved prediction accuracy over traditional models. The ML model was effective in identifying high-risk patients for targeted interventions. The integration of manual and machine-learned features highlights the importance of comprehensive feature selection. Further validation with larger datasets is necessary to generalize its findings.

Forecasting Hospital Readmissions with Machine Learning

The study employed various ML techniques, including support vector machines and random forests, to predict readmissions using data from a Greek hospital. Balanced RF models achieved the best performance in terms of sensitivity and generalization. However, the study's focus on a single hospital limits the generalizability of the results.

Application of Machine Learning in Predicting Hospital Readmissions: A Scoping Review of the Literature

The study found that tree-based methods, neural networks, and regularized logistic regression are commonly used for predicting hospital readmissions. The performance of these algorithms varies due to different factors, emphasizing the need for external validation. The review highlights the variability in model performance and the need for standardized evaluation metrics. Future efforts should focus on optimizing ML algorithms for clinical integration to improve care quality and reduce costs.

Descriptive Statistics of the Selected Dataset

Total Records: 101,767

Key Attributes: Age, sex, race, number of diagnoses, length of stay, medication use

Missing Values: During data preprocessing, missing values were identified in several attributes, including race, weight, payer_code, medical_specialty, max_glu_serum, A1Cresult, diag_1, diag_2, and diag_3. To manage these missing values, they were categorized into a separate "Missing" subgroup. However, attributes with a high percentage of missing values, such as Weight and Payer_code, will be excluded from the analysis to preserve data quality and ensure robust results.

Total Number of Attributes: 53

Numeric Attributes: 16

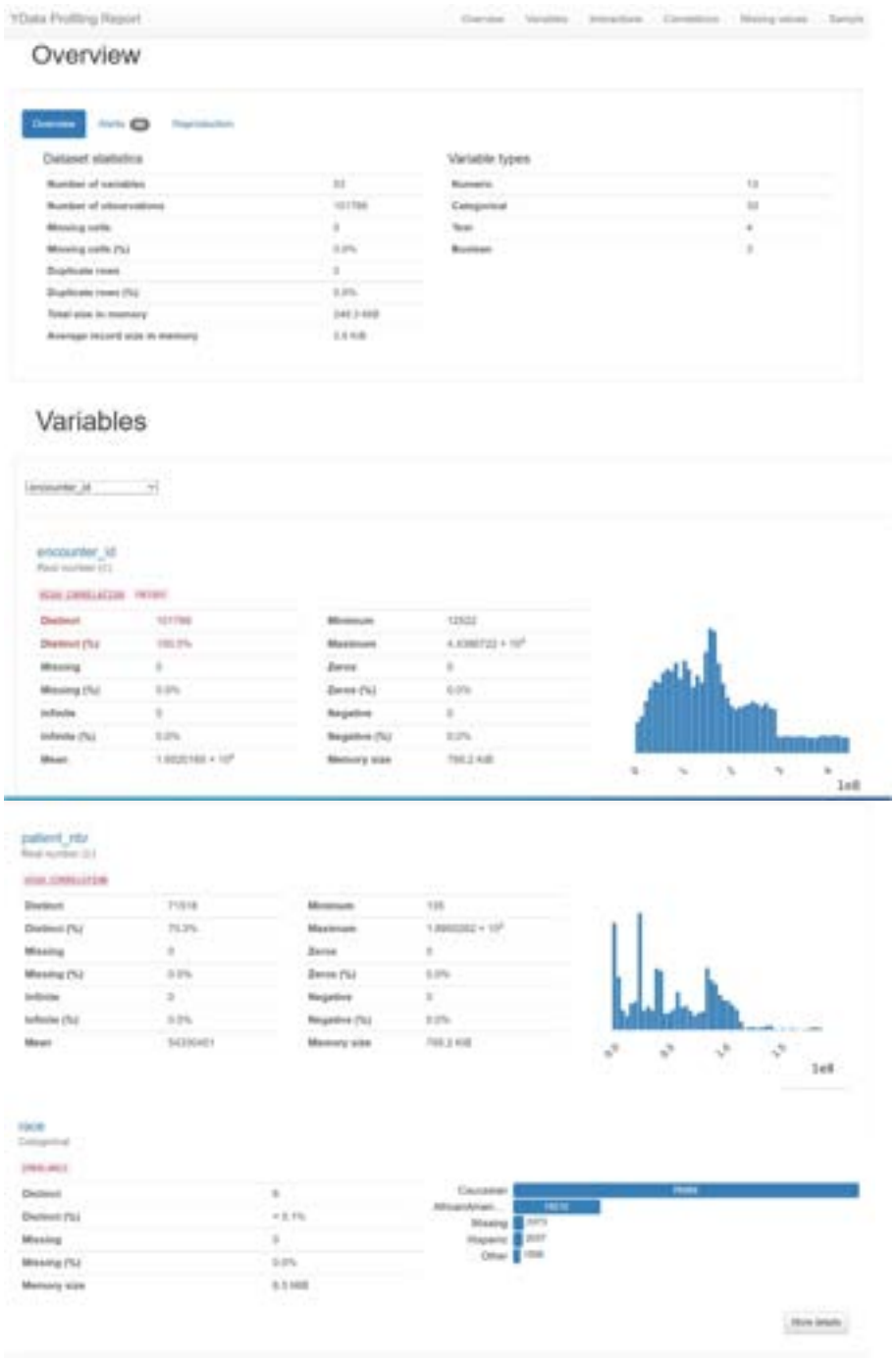
Categorical Attributes: 33

Text Attributes: 1

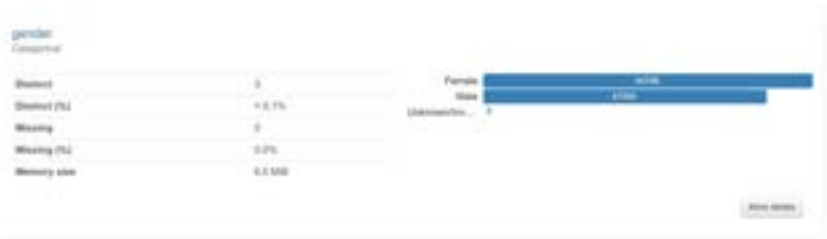
Boolean Attributes: 3

Duplicated Records: 0

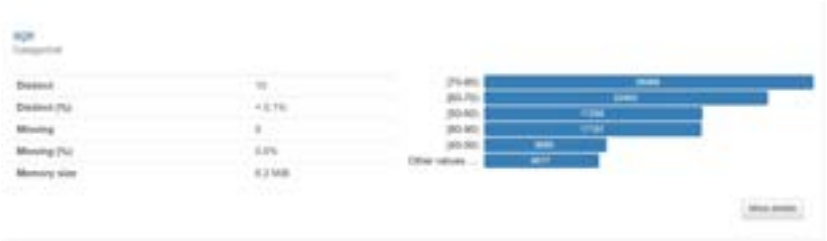
Dataset Overview & variables distribution



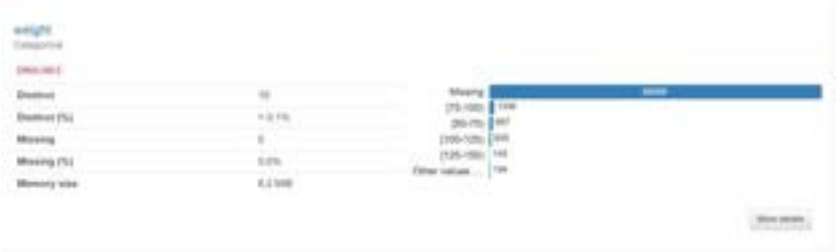
The race distribution shows the frequency of hospital admissions among different racial groups. It indicates that most patients are Caucasian.



The gender distribution illustrates that the number of female patients is higher than male patients in the dataset.



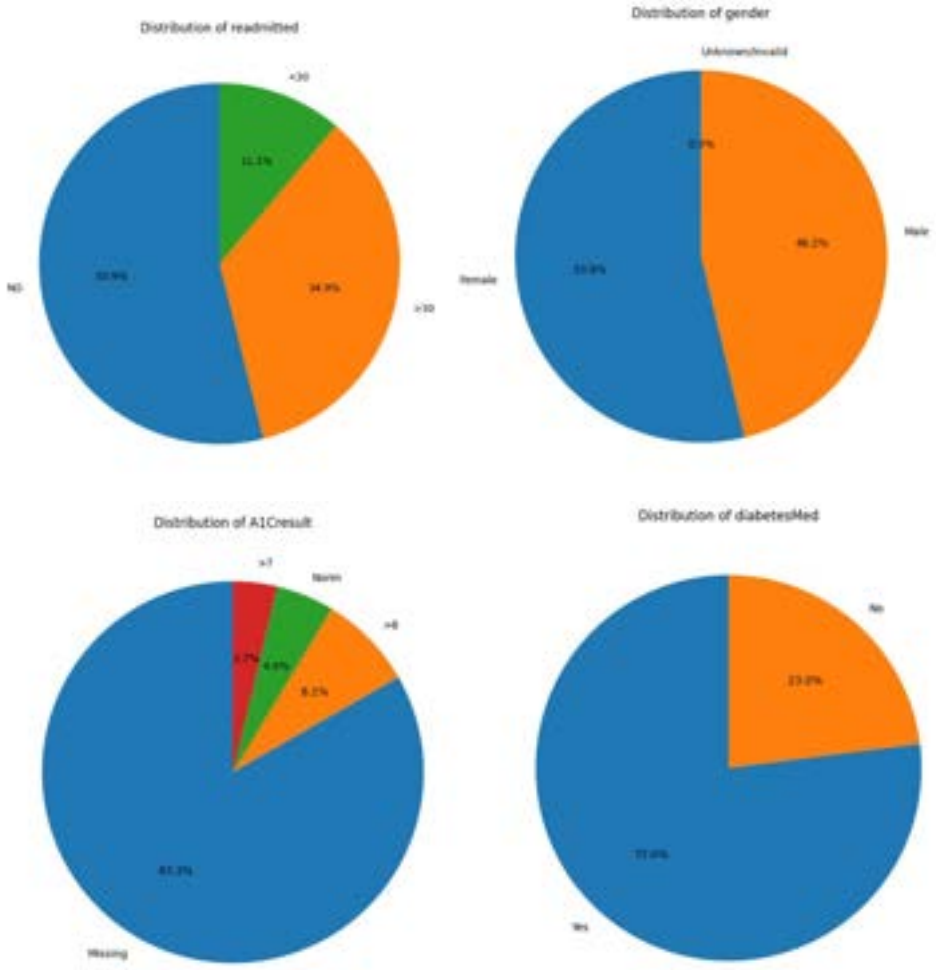
The age distribution indicates that a significant proportion of admissions occur in older adults, with a peak in the age group of 70-80 years, this highlights the increased vulnerability of elderly diabetic patients to hospital readmissions.



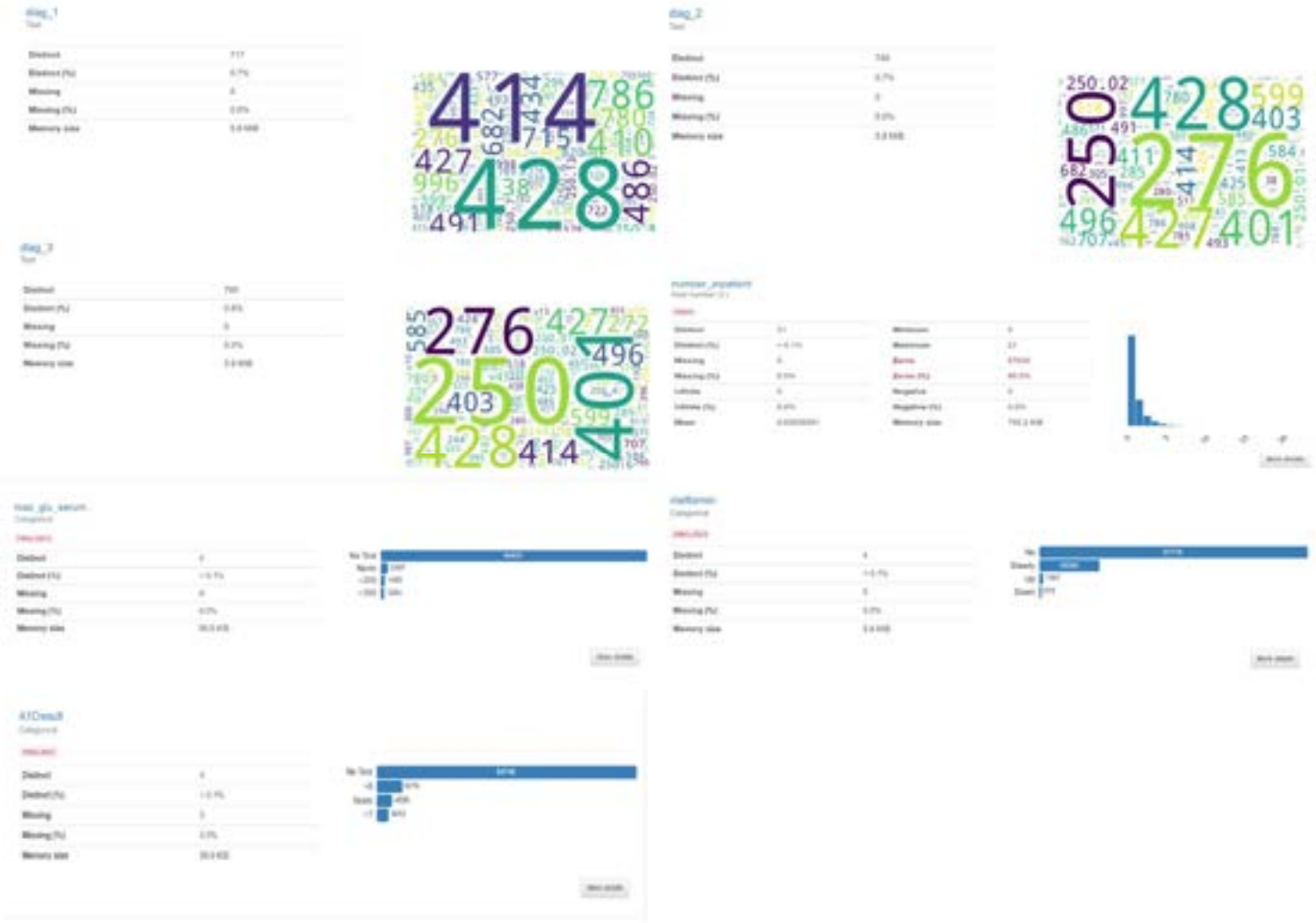
The weight distribution illustartes that we have 98% of missing values so it lacks sufficient representation and reliability. As a result, it will be excluded from further analysis.

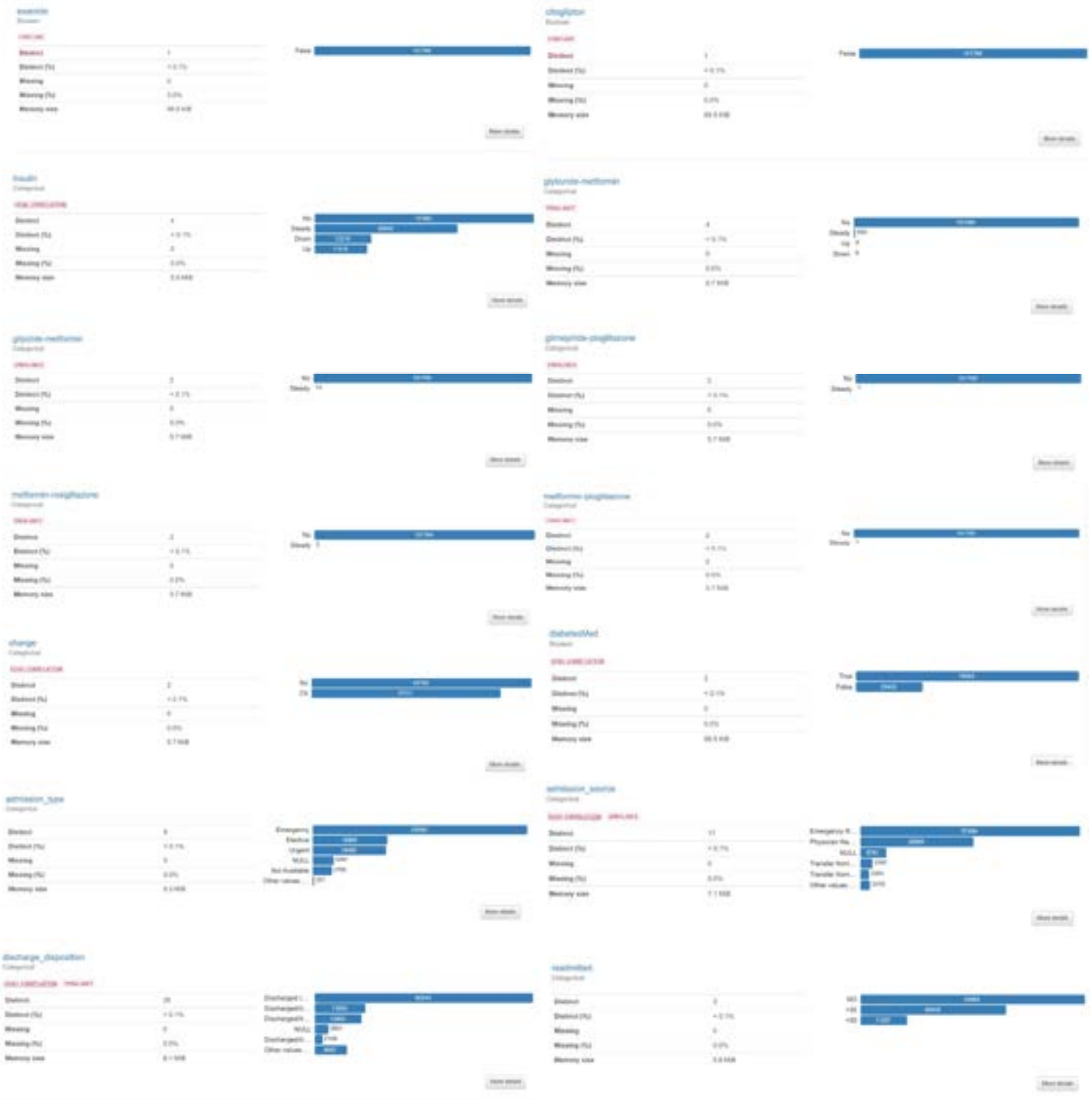
The time spent in hospital distribution reveals that most hospital stays are relatively short, with many patients staying between 2 to 5 days. However, there are a few cases with extended hospital stays.

The number of diagnoses distribution indicates that many patients have multiple diagnoses, with a concentration around 5 to 10 diagnoses. This suggests that diabetic patients often have several coexisting health issues, contributing to their risk of readmission.

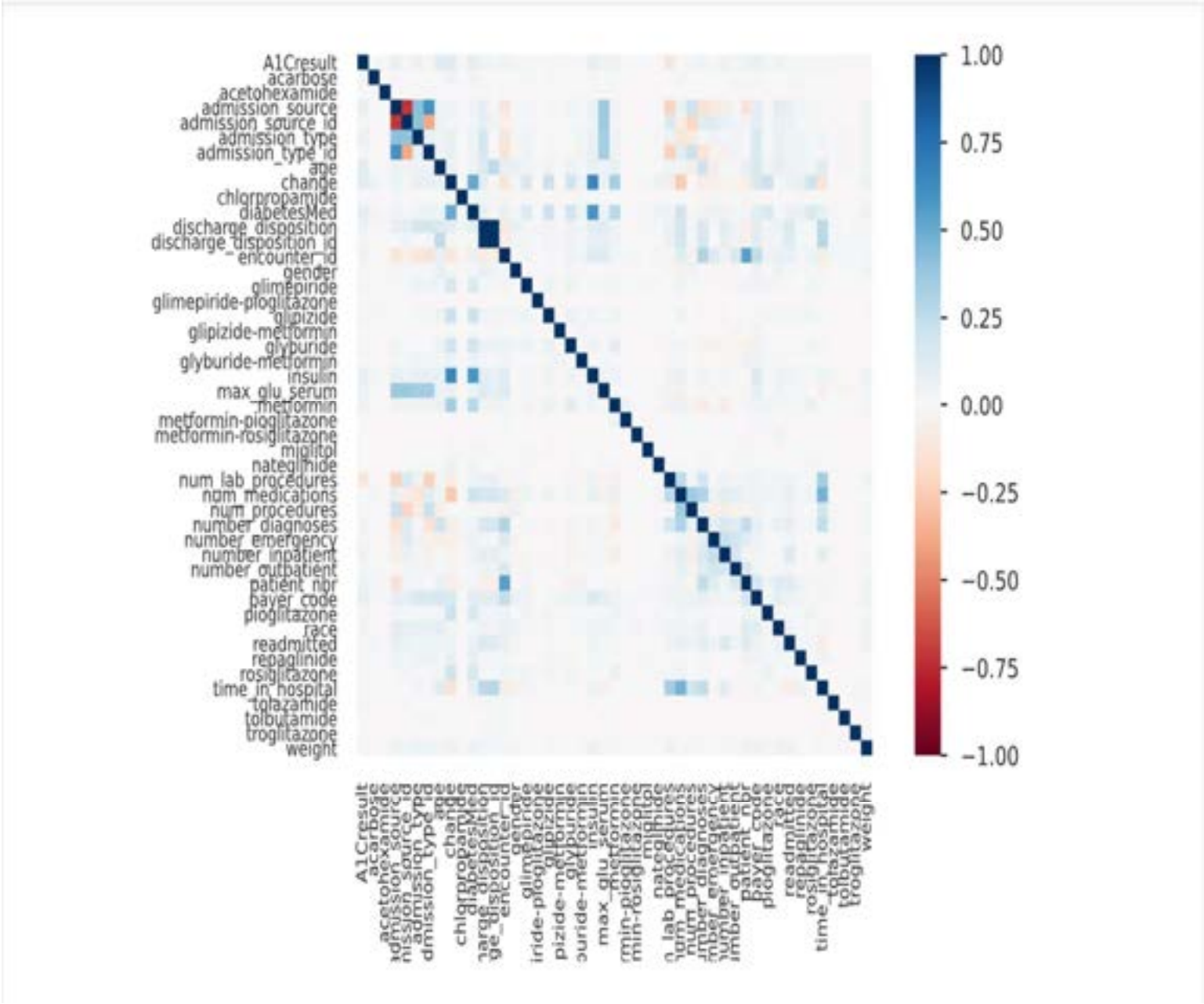

Distribution of readmitted
<30
11.2%
NO
50.9%
34.9%
>30
Distribution of gender
Unknown/Invalid
0.0%
Male
46.2%
Female
53.8%
Distribution of A1Cresult
>7
Norm
>8
4.8%
8.2%
83.3%
Missing
Distribution of diabetesMed
No
23.0%
77.0%
Yes


Distribution of insulin
No
Down
Steady





Correlation matrix



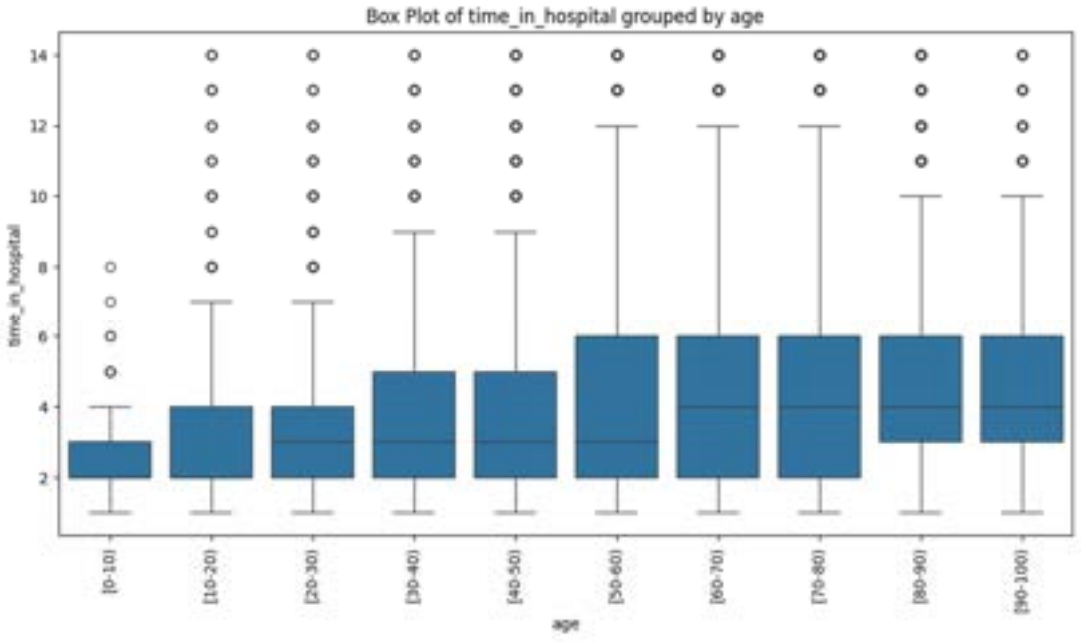
The correlation matrix illustrates the relationships between:

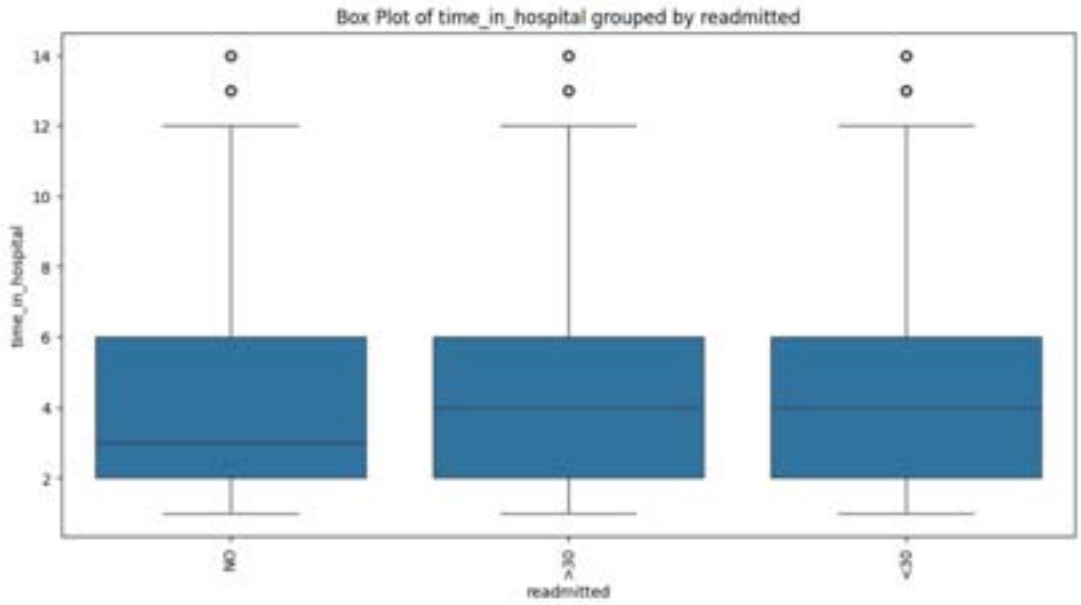
- the number of diagnoses and length of stay indicating that patients with more diagnoses tend to have longer hospital stays.

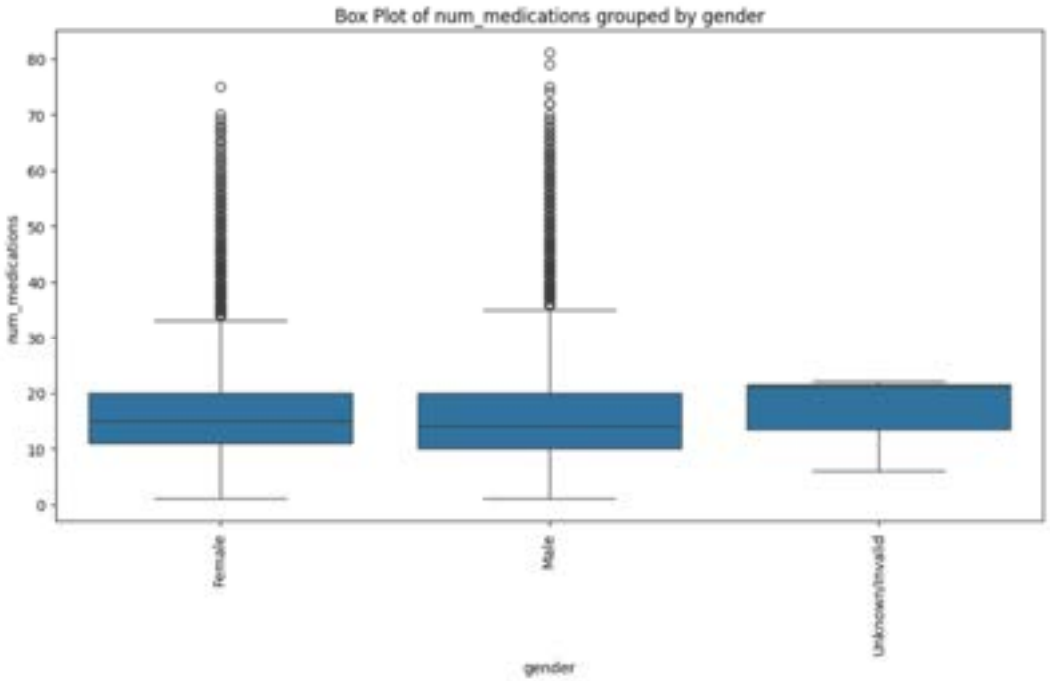
- medication use and the number of diagnoses might indicate that patients with more health issues require more intensive medication management.

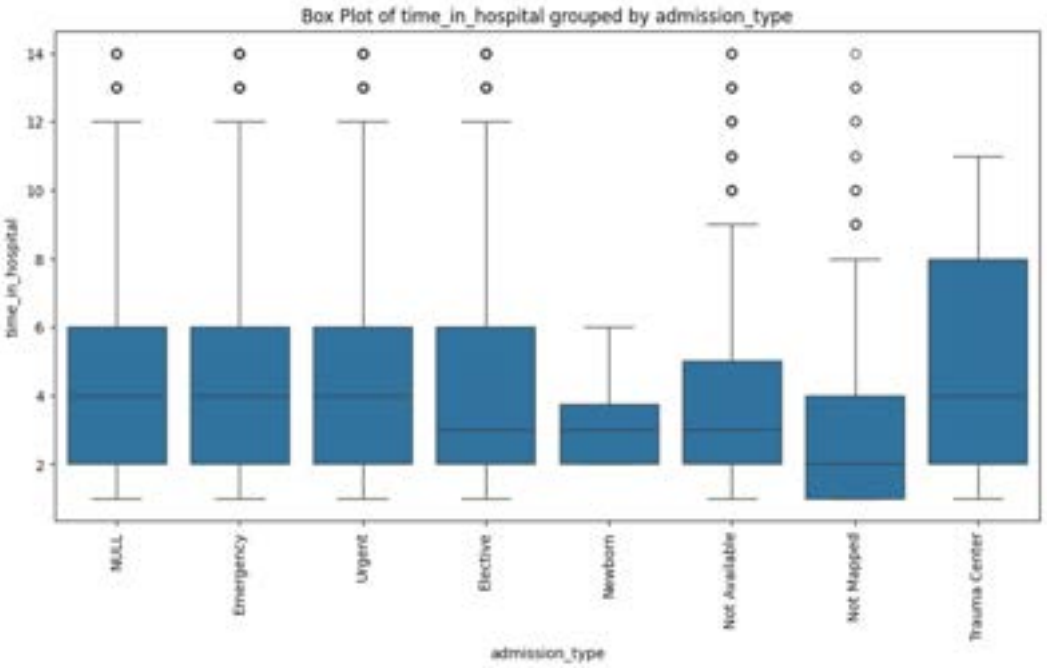
- age and length of stay may reveal that older patients tend to have longer hospital stays, which reflects the increased complexity of managing health conditions in elderly populations.

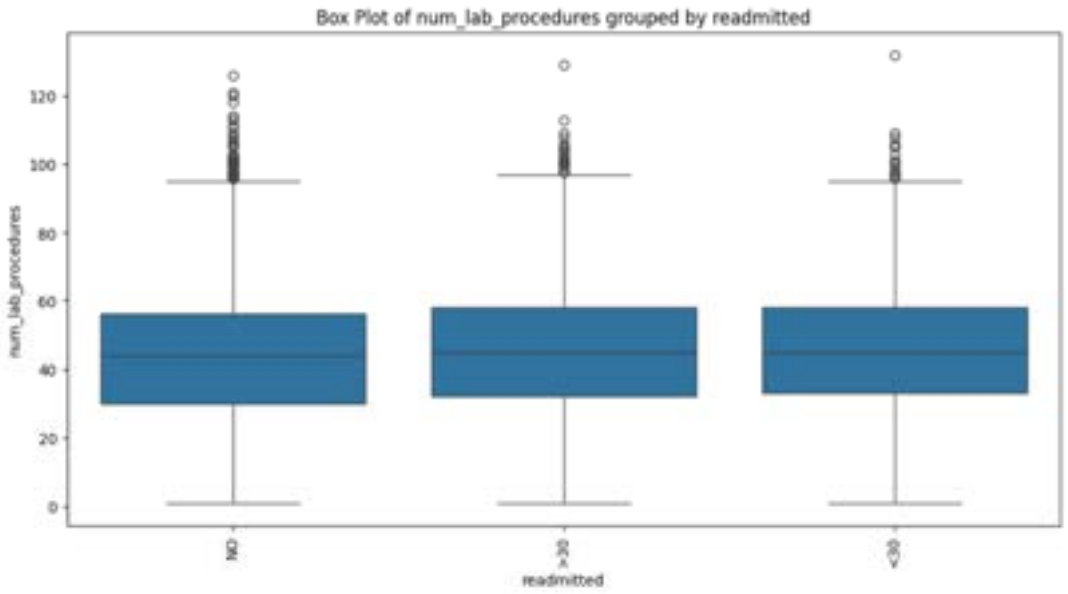
Boxplots & Outliers

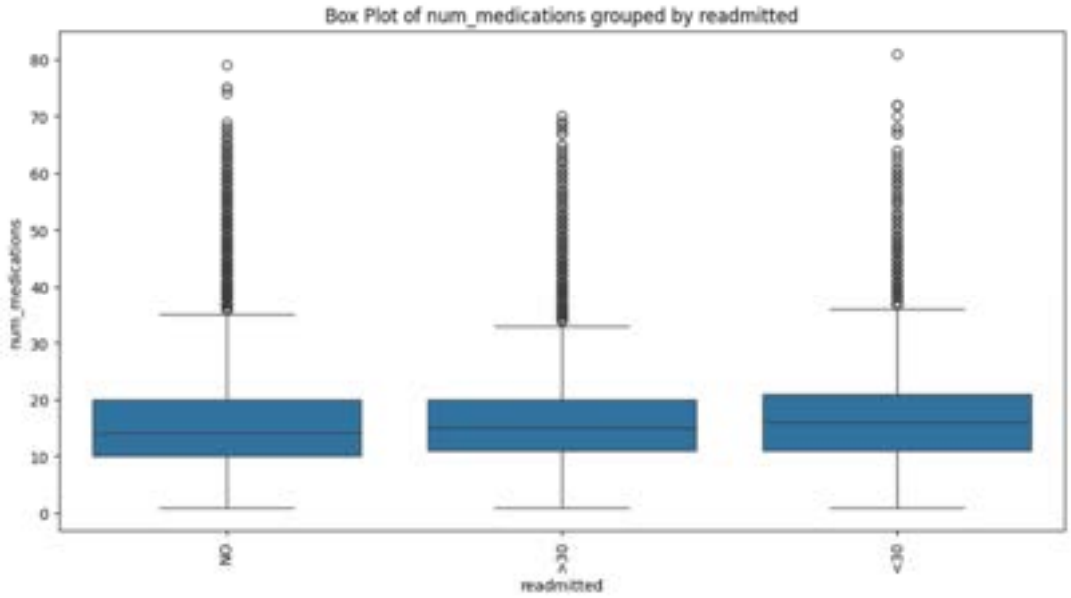


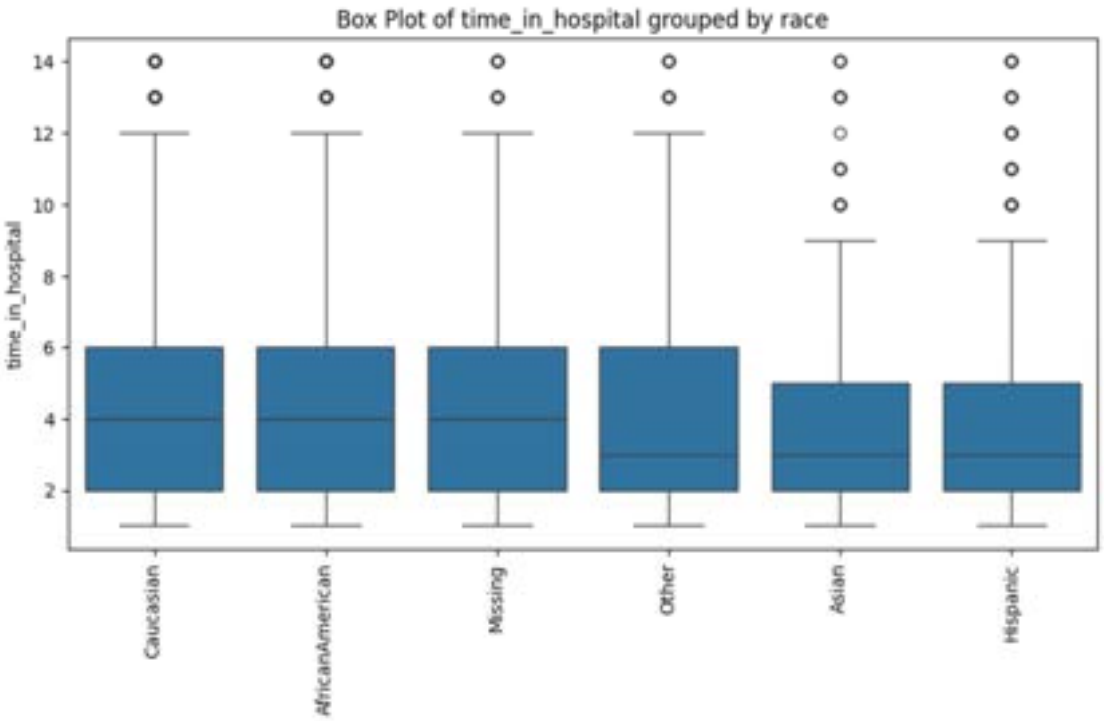


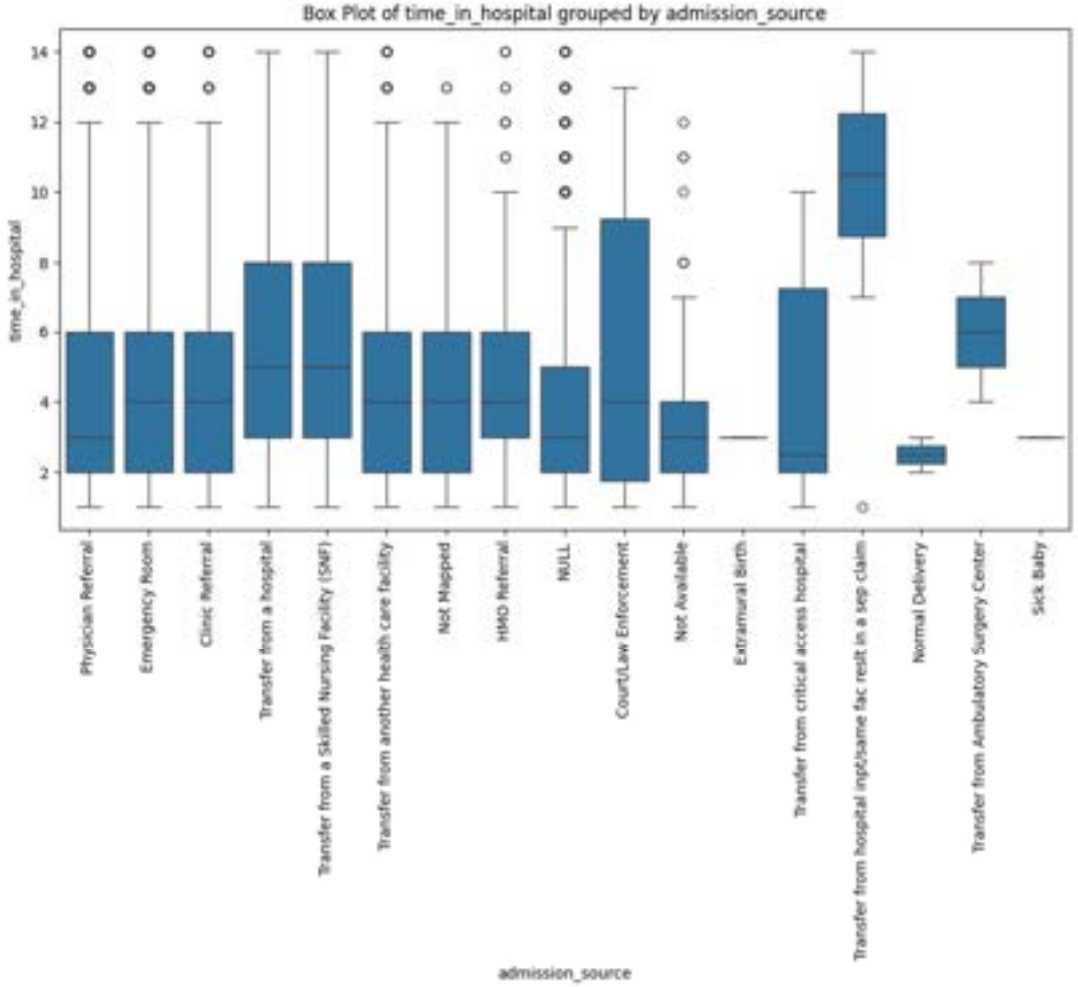

Box Plot of num_medications grouped by gender
num_medications
0
10
20
30
40
50
60
70
80
Female
Male
Unknown/Invalid
gender

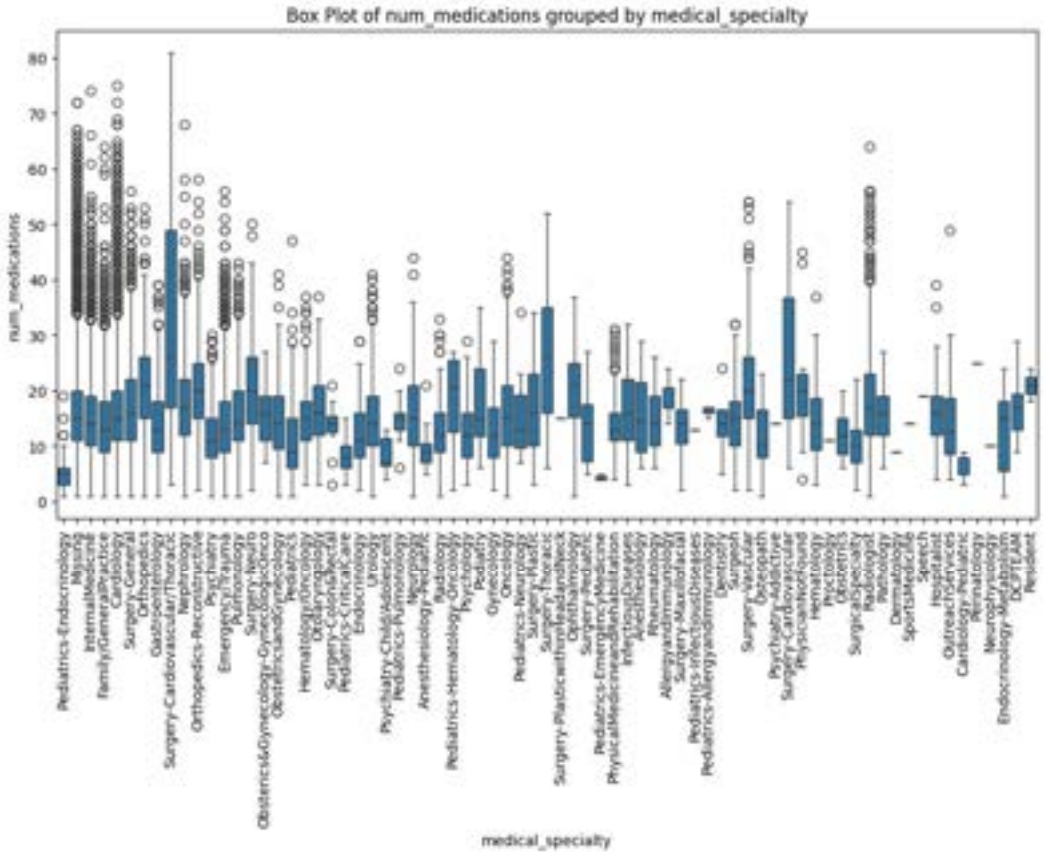

Box Plot of time_in_hospital grouped by admission_type
time_in_hospital
2
4
6
8
10
12
14
NULL
Emergency
Urgent
Elective
Newborn
Not Available
Not Mapped
Trauma Center
admission_type

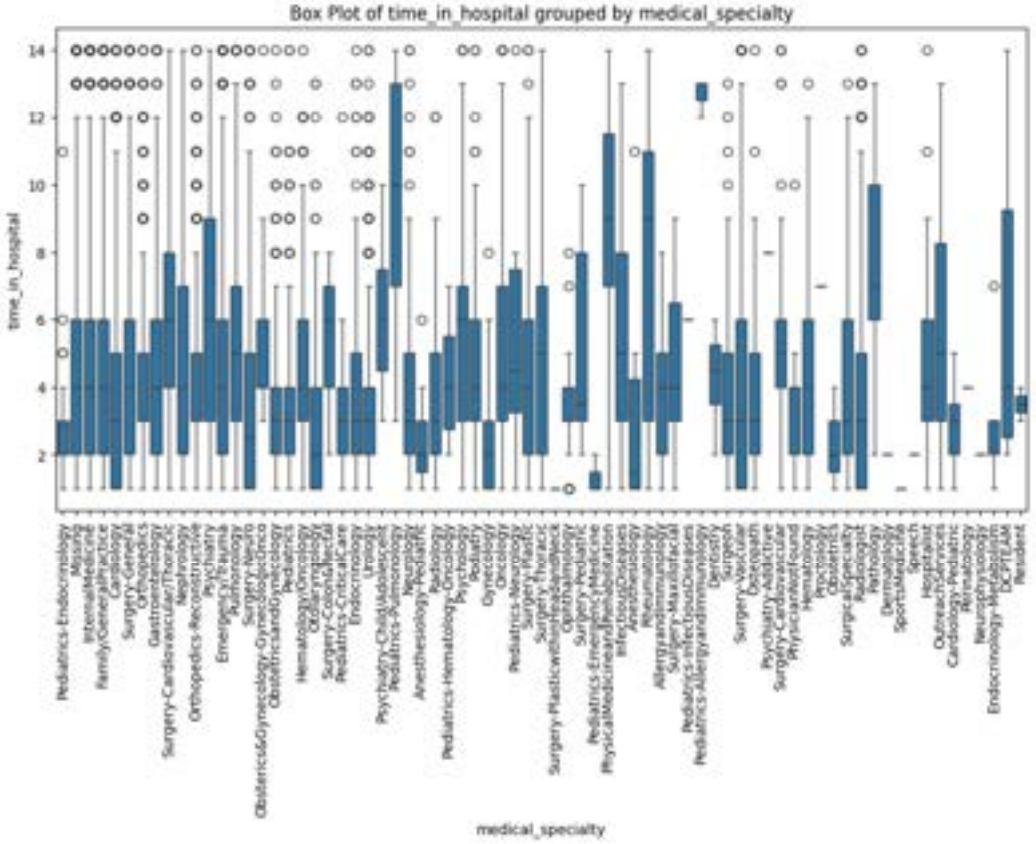

Box Plot of num_lab_procedures grouped by readmitted
num_lab_procedures
0
20
40
60
80
100
120
NO
>30
<30
readmitted

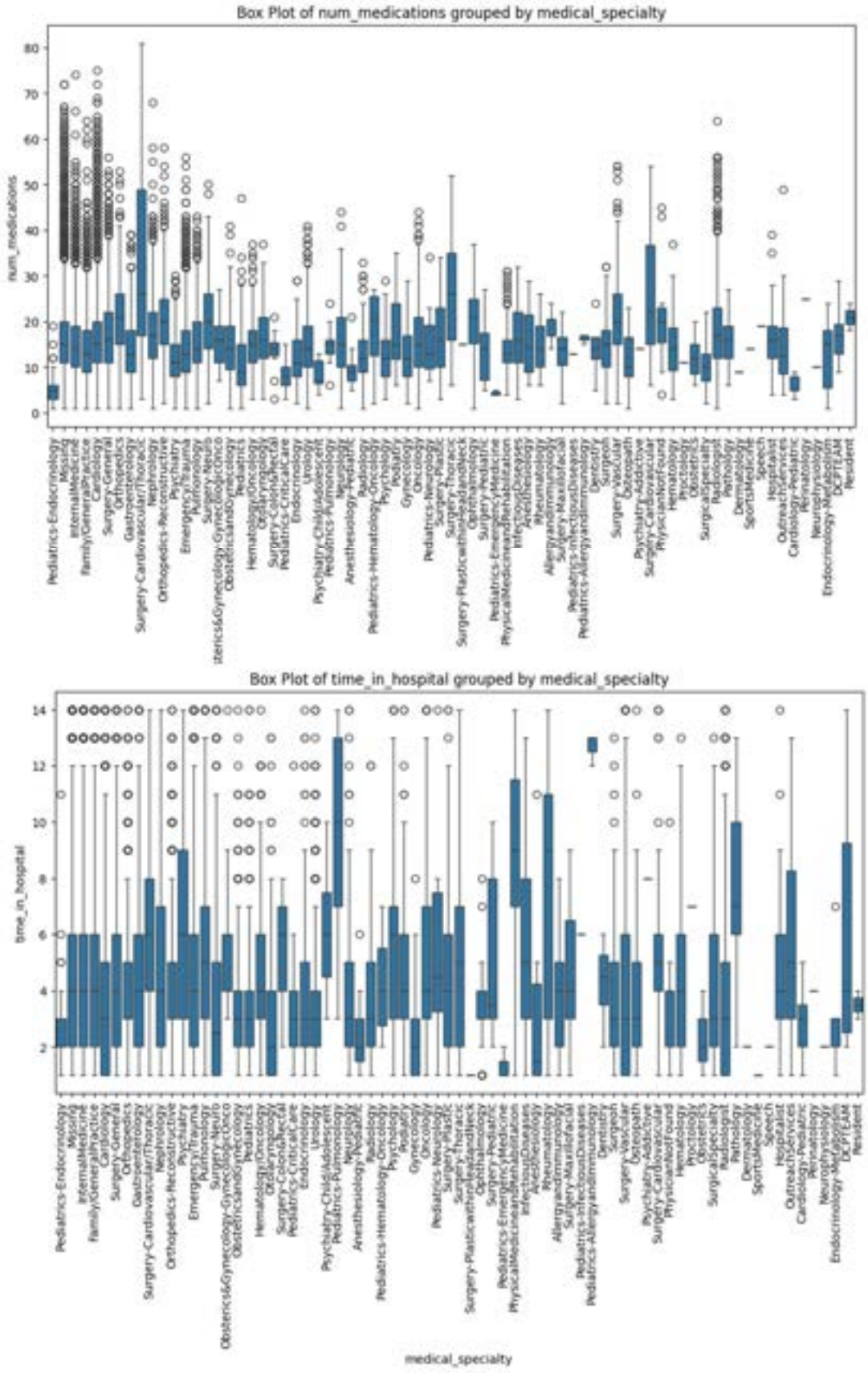

Box Plot of num_medications grouped by readmitted
num_medications
0
10
20
30
40
50
60
70
80
NO
>30
<30
readmitted

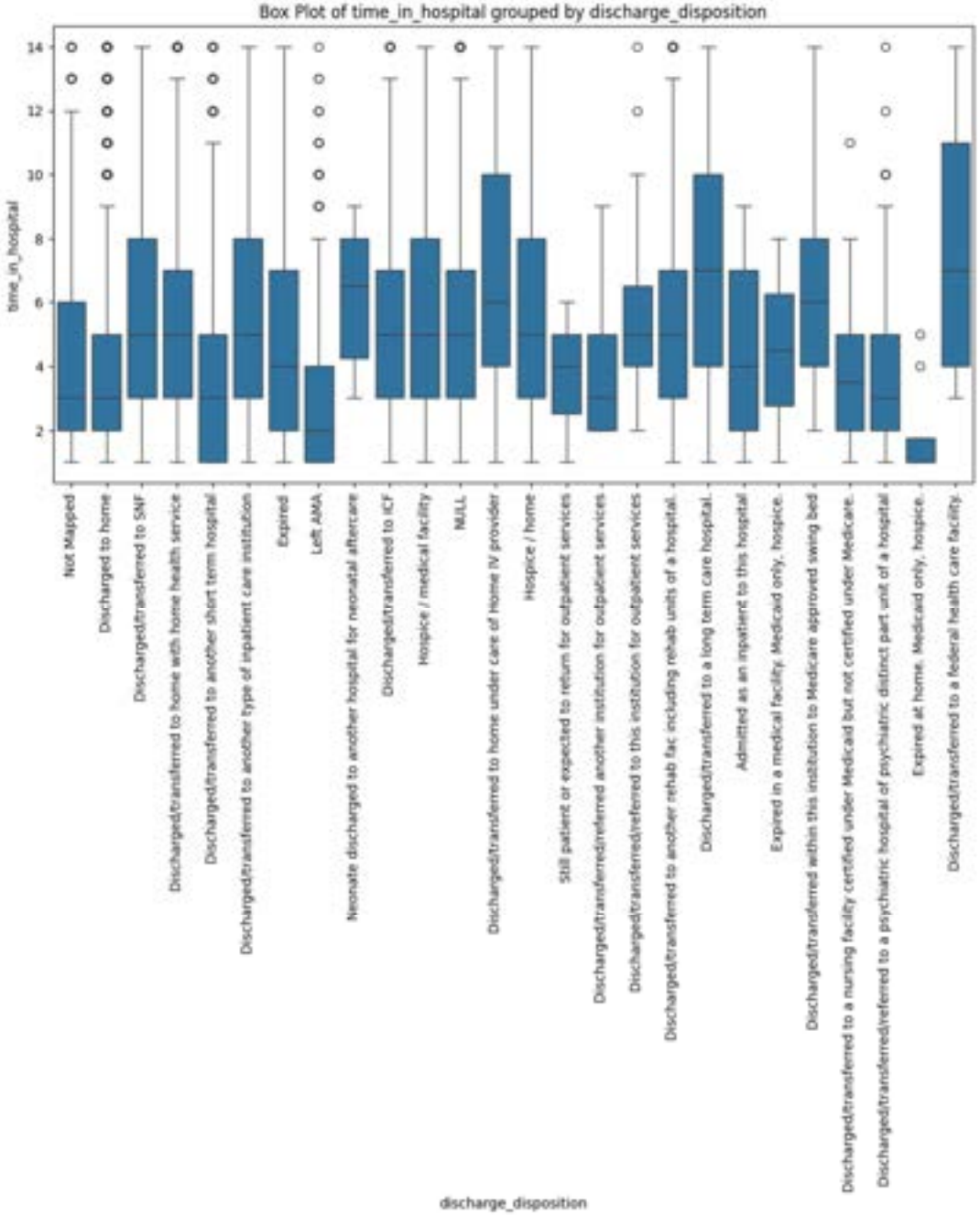

Box Plot of time_in_hospital grouped by race
time_in_hospital
14
12
10
8
6
4
2
Caucasian
AfricanAmerican
Missing
Other
Asian
Hispanic


Box Plot of time_in_hospital grouped by admission_source
time_in_hospital
14
12
10
8
6
4
2
Physician Referral
Emergency Room
Clinic Referral
Transfer from a hospital
Transfer from a Skilled Nursing Facility (SNF)
Transfer from another health care facility
Not Mapped
HMO Referral
NULL
Court/Law Enforcement
Not Available
Extramural Birth
Transfer from critical access hospital
Transfer from hospital inpt/same fac resit in a sep claim
Normal Delivery
Transfer from Ambulatory Surgery Center
Sick Baby
admission_source


Box Plot of num_medications grouped by medical_specialty
num_medications
medical_specialty


Box Plot of time_in_hospital grouped by medical_specialty
time_in_hospital
medical_specialty


Box Plot of num_medications grouped by medical_specialty
num_medications
Box Plot of time_in_hospital grouped by medical_specialty
time_in_hospital
medical_specialty


Box Plot of time_in_hospital grouped by discharge_disposition
time_in_hospital
14
12
10
8
6
4
2
Not Mapped
Discharged to home
Discharged/transferred to SNF
Discharged/transferred to home with home health service
Discharged/transferred to another short term hospital
Discharged/transferred to another type of inpatient care institution
Expired
Left AMA
Neonate discharged to another hospital for neonatal aftercare
Discharged/transferred to ICF
Hospice / medical facility
NULL
Discharged/transferred to home under care of Home IV provider
Hospice / home
Still patient or expected to return for outpatient services
Discharged/transferred another institution for outpatient services
Discharged/transferred to this institution for outpatient services
Discharged/transferred to another rehab fac including rehab units of a hospital.
Discharged/transferred to a long term care hospital.
Admitted as an inpatient to this hospital
Expired in a medical facility. Medicaid only, hospice.
Discharged/transferred within this institution to Medicare approved swing bed
Discharged/transferred to a nursing facility certified under Medicaid but not certified under Medicare.
Discharged/transferred/referred to a psychiatric hospital of psychiatric distinct part unit of a hospital
Expired at home. Medicaid only, hospice.
Discharged/transferred to a federal health care facility.
discharge_disposition

```python
#Identifying outliers in numeric columns
#Function to calculate and print outliers
def outliers(df, numeric_columns):
    outliers_count = {}

    for col in numeric_columns:
        Q1 = df[col].quantile(0.25)
        Q3 = df[col].quantile(0.75)
        IQR = Q3 - Q1
        lower_bound = Q1 - 1.5 * IQR
        upper_bound = Q3 + 1.5 * IQR

        outliers = df[(df[col] < lower_bound) | (df[col] > upper_bound)]
        outliers_count[col] = outliers[col]

        print(f"{col}:")
        print(f"  Q1: {Q1}")
        print(f"  Q3: {Q3}")
        print(f"  IQR: {IQR}")
        print(f"  Lower Bound: {lower_bound}")
        print(f"  Upper Bound: {upper_bound}")
        print(f"  Outliers: {len(outliers)}")
        print()

    return outliers_count

#print outliers
outliers_count = outliers(df, numeric_columns)
```

```
encounter_id:
  Q1: 84961194.0
  Q3: 230270887.5
  IQR: 145309693.5
  Lower Bound: -133003346.25
  Upper Bound: 448235427.75
  Outliers: 0

patient_nbr:
  Q1: 23413221.0
  Q3: 87545949.75
  IQR: 64132728.75
  Lower Bound: -72785872.125
  Upper Bound: 183745042.875
  Outliers: 247

admission_type_id:
  Q1: 1.0
  Q3: 3.0
  IQR: 2.0
  Lower Bound: -2.0
  Upper Bound: 6.0
  Outliers: 341

discharge_disposition_id:
  Q1: 1.0
  Q3: 4.0
  IQR: 3.0
  Lower Bound: -3.5
  Upper Bound: 8.5
  Outliers: 9818

admission_source_id:
  Q1: 1.0
  Q3: 7.0
  IQR: 6.0
  Lower Bound: -8.0
  Upper Bound: 16.0
  Outliers: 6956

time_in_hospital:
  Q1: 2.0
  Q3: 6.0
  IQR: 4.0
  Lower Bound: -4.0
  Upper Bound: 12.0
  Outliers: 2252

num_lab_procedures:
  Q1: 31.0
  Q3: 57.0
  IQR: 26.0
  Lower Bound: -8.0
  Upper Bound: 96.0
  Outliers: 143

num_procedures:
  Q1: 0.0
  Q3: 2.0
  IQR: 2.0
  Lower Bound: -3.0
  Upper Bound: 5.0
  Outliers: 4954

num_medications:
  Q1: 10.0
  Q3: 20.0
  IQR: 10.0
  Lower Bound: -5.0
  Upper Bound: 35.0
  Outliers: 2557

number_outpatient:
  Q1: 0.0
  Q3: 0.0
  IQR: 0.0
  Lower Bound: 0.0
  Upper Bound: 0.0
  Outliers: 16739

number_emergency:
  Q1: 0.0
  Q3: 0.0
  IQR: 0.0
  Lower Bound: 0.0
  Upper Bound: 0.0
  Outliers: 11383

number_inpatient:
  Q1: 0.0
  Q3: 1.0
  IQR: 1.0
  Lower Bound: -1.5
  Upper Bound: 2.5
  Outliers: 7049

number_diagnoses:
  Q1: 6.0
  Q3: 9.0
  IQR: 3.0
  Lower Bound: 1.5
  Upper Bound: 13.5
  Outliers: 281
```

The outliers count shows that:

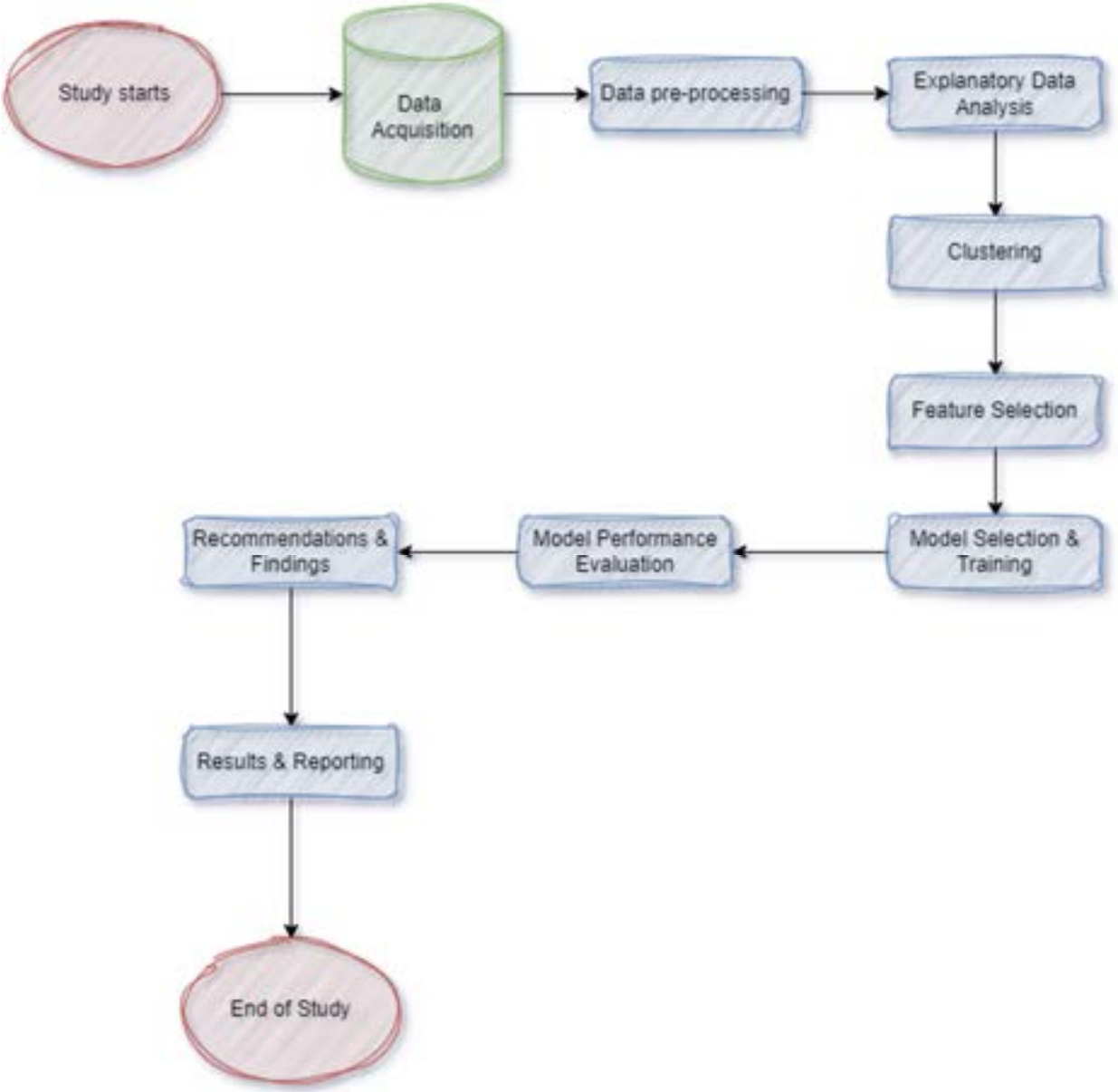
- Most patients have a stay duration within the interquartile range of approx. 4 days, but there are several outliers with stays extending beyond 14 days. These outliers may represent patients with severe complications or those requiring extended care, which are important to consider for accurate modeling.

- There are patients with number of diagnoses more than 13.5 diagnoses indicate patients with exceptionally complex medical histories, these patients are at a significantly higher risk for readmissions due to the overcomplex nature of their health conditions. These patients may require more comprehensive care plans and closer monitoring to prevent/reduce readmissions.

Link to GitHub Repository

https://github.com/NehalTMU/TMU-Capstone-Project

Methodology Flowchart



The study begins with data preprocessing. Following that will be applying clustering algorithms to uncover hidden patterns and identify patient subgroups with different readmission risks. This step aims to enhance the understanding of patient categories and tailor the predictive models accordingly.

Subsequently, classification techniques will be applied to predict early readmission outcomes based on patient and hospital features. Machine learning algorithms including logistic regression, decision trees, and random forests will be utilized. The dataset will be split into training and

testing sets, and the models will be trained and evaluated using standard metrics such as accuracy, precision, recall.

Feature selection and dimensionality reduction techniques, such as Recursive Feature Elimination (RFE) and Principal Component Analysis (PCA), will be applied to identify the most informative factors and improve model performance. Python programming will be used for the implementation, leveraging libraries and tools such as pandas for data manipulation, scikit-learn for machine learning, NumPy for numerical operations, Matplotlib and Seaborn for data visualization.

Data Preprocessing

Handling Missing Values

- Missing values were handled as follows:
- Replaced '?' and blank cells with NaN.
- Columns with a high percentage of missing values or irrelevant to the study scope were excluded e.g., ('weight', 'payer_code', 'medical_specialty', 'encounter_id', 'patient_nbr').
- Missing values in categorical columns such as 'race', 'diag_1', 'diag_2', and 'diag_3' were replaced with the mode of each column.

Data Cleaning and Feature Engineering

- Simplified and categorized diagnosis codes.
- Combined 'NO' and '>30' categories into a single '>30' category for the target variable 'readmitted'.
- Dropped columns dominated by "No" responses and those with low variance.
- Converted numerical columns to numeric data types and categorical columns to category data types.
- One-hot encoded categorical features.

Normalization and Resampling

- Normalized the data using StandardScaler.
- Addressed class imbalance using SMOTETomek.

Classification Analysis

Model Training and Evaluation

Four models were trained and evaluated: Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting.

Cross-Validation

10-fold cross-validation was performed for each model to ensure robustness and reliability.

Results (Cross-Validation)

Model	Accuracy	Precision	Recall	Confusion Matrix
Logistic Regression	61.95%	63.35%	56.72%	TN: 42,480, FP: 20,750, FN: 27,369, TP: 35,861
Decision Tree	87.66%	87.05%	88.48%	TN: 54,907, FP: 8,323, FN: 7,283, TP: 55,947
Random Forest	94.22%	99.49%	88.90%	TN: 62,943, FP: 287, FN: 7,019, TP: 56,211
Gradient Boosting	91.28%	98.69%	83.66%	TN: 62,530, FP: 700, FN: 10,329, TP: 52,901

Feature Selection

Recursive Feature Elimination (RFE) was used to identify the top features contributing to the predictions. The features selected were:

Logistic Regression: number_inpatient, number_diagnoses, age_[70-80), age_[80-90), admission_type_id_7, discharge_disposition_id_11, discharge_disposition_id_12, discharge_disposition_id_16, discharge_disposition_id_17, discharge_disposition_id_22.

Random Forest: number_inpatient, num_procedures, time_in_hospital, diag_3, number_diagnoses, diag_2, diag_1, num_medications, num_lab_procedures, gender_Male.

Clustering Analysis

Determination of Optimal Clusters (Elbow Method)

The Elbow Method was used to determine the optimal number of clusters (k). The optimal k was found to be 4 based on the elbow observed in the WCSS plot.

Clustering Results

K-Means clustering was performed with k=4 clusters. Each data point was assigned to one of the 4 clusters.

Cluster Characteristics and Analysis

Cluster 0: Lower average time in the hospital and fewer lab procedures, indicating less severe or less complicated cases.

Cluster 1: Higher average time in the hospital and more medications, suggesting more severe cases or cases requiring intensive treatment.

Cluster 2: Moderate lab procedures and time in the hospital, representing cases with average severity.

Cluster 3: Like Cluster 0, with lower values across several features, possibly representing routine or mild cases.

Silhouette Score

The silhouette score for k=4 was 0.126, indicating that the clusters are not very well separated and overlap to some extent.

The highest silhouette score was achieved with k=2, yielding a score of 0.181, suggesting better-defined clusters with less overlap.

Clustering Results with k=2:

K-Means clustering was performed with k=2 clusters.

Each data point was assigned to one of the 2 clusters.

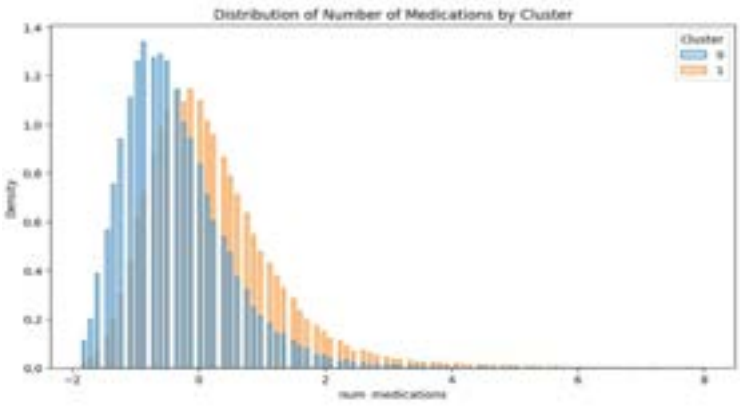
Cluster Characteristics and Analysis (k=2):

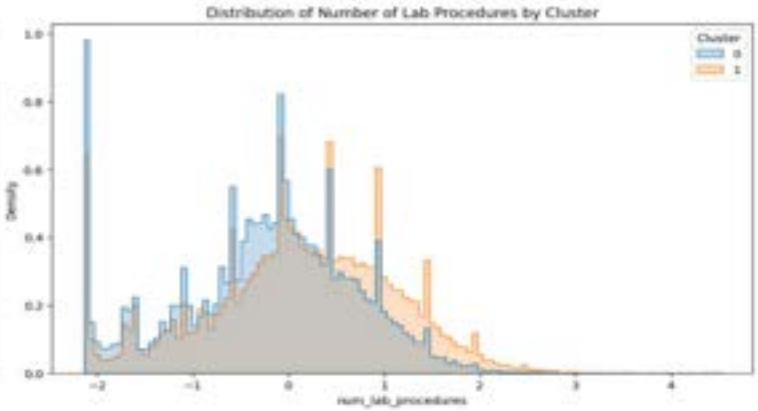
Cluster 0: This cluster had a lower average time in the hospital and fewer lab procedures, indicating less severe or less complicated cases.

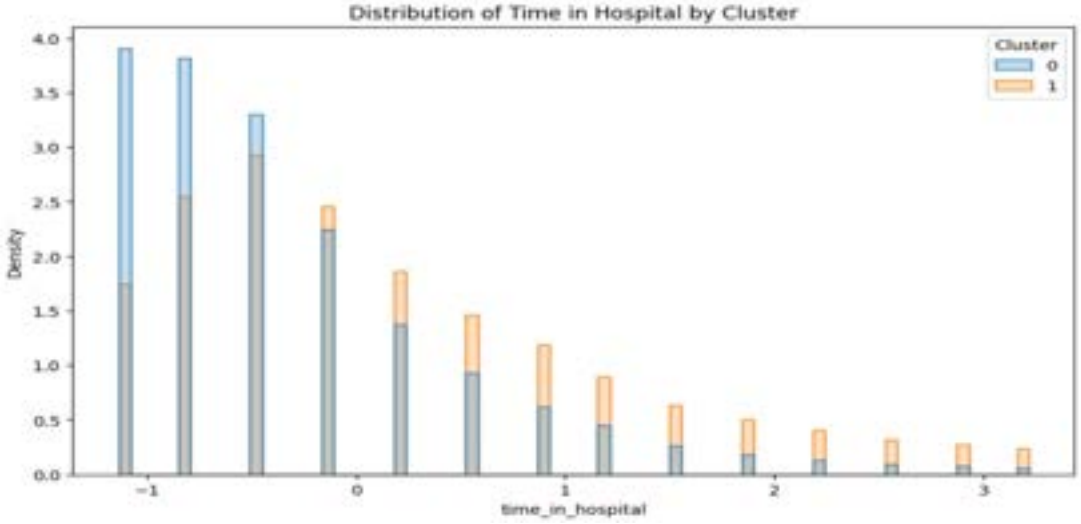
Cluster 1: This cluster had a higher average time in the hospital and more medications, suggesting more severe cases or cases requiring intensive treatment.

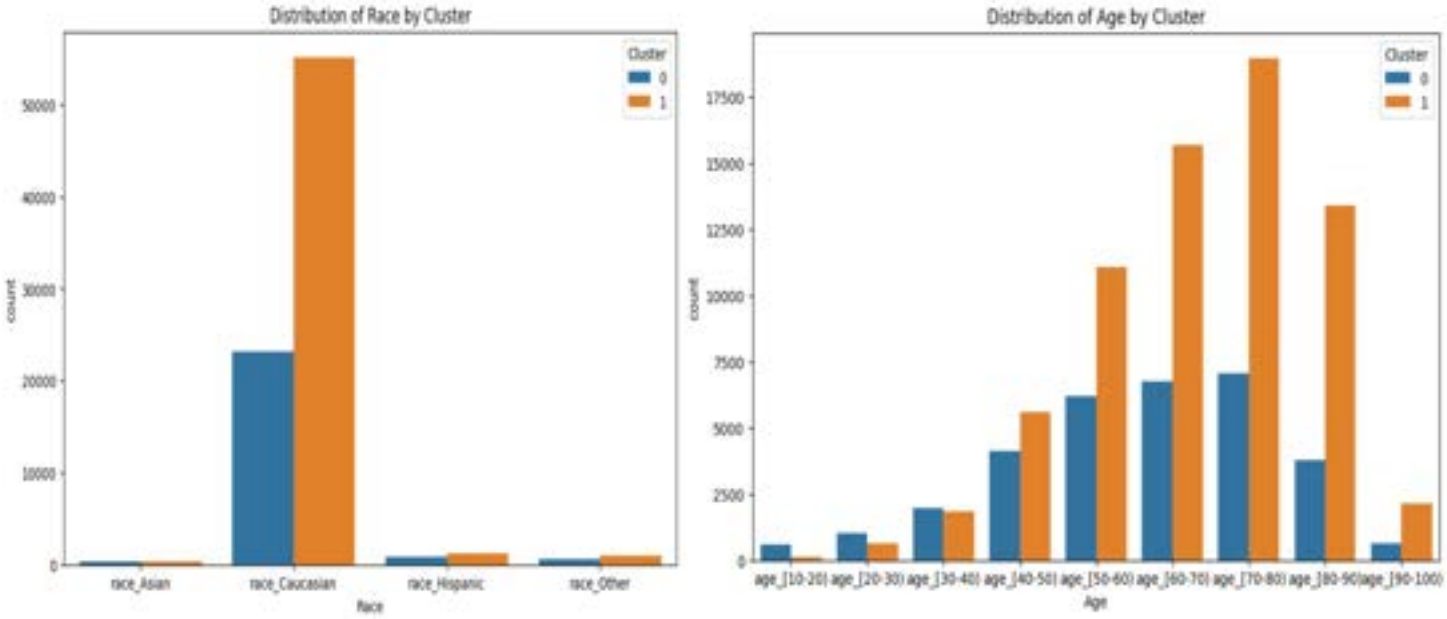
Visualization Insights

Visualizations of feature distributions across clusters highlighted distinct patterns, showing differences in patient profiles among the clusters. The distribution of categorical features like race and age provided insights into demographic patterns within each cluster.









Conclusions

- **Cluster Interpretation**: Different segments of the patient population characterized by varying levels of medical complexity and resource utilization.
- **Clinical Relevance**: Understanding these clusters can help in designing targeted interventions and resource allocation strategies.
- **Model Evaluation**: The clusters provided meaningful insights despite the modest silhouette score. Future work could explore other clustering methods and further tune parameters to improve cluster quality.
- **Data Insights**: Most patients fall into clusters representing routine or mild cases, while a smaller proportion falls into clusters requiring intensive treatment and longer hospital stays.

Overall Conclusions

- The study successfully applied classification and clustering techniques to predict early readmissions for diabetic patients.
- Cross-validation results were robust, providing reliable models for predicting readmissions.
- Clustering analysis offered valuable insights into patient segmentation, which can be used to optimize resource allocation and targeted interventions.
- Future work could involve exploring additional models, tuning parameters, and applying alternative clustering methods for further improvement.

Future Work

1. **Enhanced Model Tuning & Feature Engineering**:

- Acquire more updated datasets with fewer anomalies and missing values to improve model accuracy and reliability.
- Collect data on patients' health behaviors and lifestyle choices, such as diet, weight, exercise, and smoking status, to identify their impact on readmission rates.
- Explore advanced hyperparameter tuning techniques such as GridSearchCV.

2. **Advanced Clustering Techniques**:

- Implement hierarchical clustering to explore the possibility of a more granular grouping of patients based on their readmission characteristics.
- Explore DBSCAN Optimization to refine parameters to reduce the sensitivity to noise and improve cluster quality.

3. **Validation and Interpretation**:

- Engage healthcare professionals to validate the identified clusters and ensure their clinical relevance.

4. **Integration with Predictive Modeling**:

- Design targeted intervention strategies for high-risk clusters to reduce the likelihood of readmissions, such as personalized follow-up plans and post-discharge support.

5. **Addressing Class Imbalance**:

- Apply advanced resampling techniques like SMOTE, ADASYN, or cost-sensitive learning to address class imbalance and enhance model performance.

- Implement anomaly detection algorithms to identify rare but significant patterns that may contribute to readmissions.

6. **Scalability and Generalization**:

- Test the scalability of the clustering and predictive models on larger datasets to ensure they can handle increased data volume without compromising performance.
- Evaluate the generalizability of the findings across different hospitals and healthcare settings to ensure the robustness and applicability of the results.

References

Dataset Source

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

Strack, B., DeShazo, J. P., Gennings, C., Olmo, J. L., Ventura, S., Cios, K. J., & Clore, J. N. (2014). Impact of HbA1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records. *BioMed Research International, 2014*, 781670–781611. https://doi.org/10.1155/2014/781670

URL: https://www.hindawi.com/journals/bmri/2014/781670/

Kumar Sah, D., & Khanal, M. (2023, November). Implementation of big data analytics on diabetes 130-US hospitals for the year 1999-2008 for predicting patient readmission. Preprint. https://doi.org/10.13140/RG.2.2.18564.30081

URL: https://www.researchgate.net/publication/375690075_Implementation_of_Big_Data_Analytics_on_Diabetes_130-US_Hospitals_for_year_1999-2008_for_predicting_patient_readmission

Shang, Y., Jiang, K., Wang, L., Zhang, Z., Zhou, S., Liu, Y., Dong, J., & Wu, H. (2021). The 30-days hospital readmission risk in diabetic patients: Predictive modeling with machine learning classifiers. *BMC Medical Informatics and Decision Making, 21*(Suppl 2), 57. https://doi.org/10.1186/s12911-021-01423-y

URL: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-021-01423-y

Tavakolian, A., Rezaee, A., Hajati, F., & Uddin, S. (2023). Hospital readmission and length-of-stay prediction using an optimized hybrid deep model. *Future Internet, 15*(9), Article 304. https://doi.org/10.3390/fi15090304

URL: https://www.mdpi.com/1999-5903/15/9/304

Davis, S., Zhang, J., Lee, I., Rezaei, M., Greiner, R., McAlister, F. A., & Padwal, R. (2022). Effective hospital readmission prediction models using machine-learned features. *BMC Health Services Research*. https://doi.org/10.1186/s12913-022-08748-y

URL: https://bmchealthservres.biomedcentral.com/articles/10.1186/s12913-022-08748-y

Michailidis, P., Dimitriadou, A., Papadimitriou, T., & Gogas, P. (2022). Forecasting hospital readmissions with machine learning. *Healthcare, 10*, 981. https://doi.org/10.3390/healthcare10060981

URL: https://www.mdpi.com/2227-9032/10/9/981

Huang, Y., Talwar, A., Chatterjee, S., & Aparasu, R. R. (2021). Application of machine learning in predicting hospital readmissions: A scoping review of the literature. *BMC Medical Research Methodology*. https://doi.org/10.1186/s12874-021-01284-z

URL: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12874-021-01284-z

Centers for Medicare & Medicaid Services. (n.d.). CMS hospital readmissions reduction program (HRRP). Retrieved from https://www.cms.gov

American Diabetes Association. (2022). Statistics about diabetes. Retrieved from https://www.diabetes.org

Zhu, Z., Hinton, N., Zhao, Y., & He, Y. (2021). 30-Day readmission risk for diabetic patients with COVID-19. *Diabetes Care, 44*(7), 1530-1533. https://doi.org/10.2337/dc21-0104

URL: https://care.diabetesjournals.org/content/44/7/1530

Hirsch, J. S., Ng, J. H., Ross, D. W., Sharma, P., Shah, H. H., Barnett, R. L., ... & Northwell COVID-19 Research Consortium. (2020). Acute kidney injury in patients hospitalized with COVID-19. *Kidney International, 98*(1), 209-218. https://doi.org/10.1016/j.kint.2020.05.006

URL: https://www.kidney-international.org/article/S0085-2538(20)30612-9/fulltext

Dandachi, D., Geiger, G., Montgomery, M. W., Kharfen, M., & Rodriguez-Barradas, M. C. (2021). Characteristics, outcomes, and mortality among persons with diabetes hospitalized with COVID-19: Experience from a large New York City health system. *Journal of Diabetes and Its Complications, 35*(10), 107966. https://doi.org/10.1016/j.jdiacomp.2021.107966

URL: https://www.journalofdiabetesanditscomplications.com/article/S1056-8727(21)00140-2/fulltext