

INT375
PROJECT REPORT
(Project Semester January-April 2025)

HOSPITAL READMISSION PREDICTION

Submitted by:

Name: Aditya Sharma
Registration No: 12320938

Programme and Section : B.Tech CSE K23PM
Course Code: INT375

Under the Guidance of
Dr. Anand Kumar (UID : 30561)
Discipline of CSE/IT

Lovely School of Computer Science and Engineering
Lovely Professional University, Phagwara

CERTIFICATE

This is to certify that Aditya Sharma bearing Registration no. 12320938 has completed INT-375 project titled, “**HOSPITAL READMISSION PREDICTION**” under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

Signature and Name of the Supervisor

Dr. Anand Kumar

School of Computer Science and Engineering

Lovely Professional University

Phagwara, Punjab.

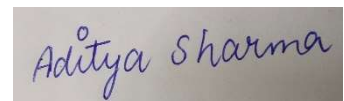
Date: 12-04-2025

DECLARATION

I, Aditya Sharma, student of B.Tech under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 12-04.2025

Registration No: 12320938



Signature

Aditya Sharma

ACKNOWLEDGEMENT

I would like to express my deepest and most sincere gratitude to my project guide, Anand Kumar, for their invaluable guidance, constant support, and expert insights throughout the duration of this project. Their unwavering encouragement, sharp intellect, and mentorship have played a pivotal role in helping me conceptualize and successfully complete this work. From the initial idea formulation to the final evaluation phase, their feedback and advice have consistently pushed me to improve and deepen my understanding of both technical and analytical aspects of the project.

Working under Anand Kumar has been a highly rewarding experience. Their emphasis on discipline, clarity, and logical thinking helped me approach problems from new perspectives and nurtured my confidence in solving complex data science challenges. I am immensely grateful for the many brainstorming sessions, critical reviews, and motivating discussions that fueled this project's progress.

I would also like to extend my sincere appreciation to all the faculty members and technical staff of the School of Computer Science and Engineering at Lovely Professional University for equipping me with the foundational skills and providing the academic infrastructure necessary to undertake this project. Their consistent encouragement and willingness to support innovative thinking are truly commendable.

Furthermore, I am thankful to the online data science community, particularly contributors on platforms like Kaggle, Stack Overflow, and Medium, whose shared knowledge and real-world insights helped me clarify concepts and refine my approach at various stages of the project.

Finally, I would like to thank my family and close friends for their continuous support, motivation, and patience throughout this journey. Their belief in my abilities has been a source of strength during times of challenge and has inspired me to keep striving for excellence.

This project has been a significant learning experience and a proud milestone in my academic journey. It would not have been possible without the collective support and encouragement of all the aforementioned individuals and communities.

TABLE OF CONTENTS

1.	Introduction	6
2.	Source of Dataset	7
3.	Dataset Preprocessing	8-11
4.	Analysis on Dataset (for each objective) i. Purpose ii. Analysis iii. Insights iv. Importance	13-21
5.	Conclusion	22
6.	Future Scope	23-24
7.	References	25
8.	Research Paper	26-27

1. Introduction

Hospital readmissions, the return of a patient to a hospital within a specified period following discharge, represent a significant burden on healthcare systems worldwide. These events not only escalate healthcare costs but also diminish patient quality of life and reflect potential inadequacies in care transitions and post-discharge management. The complexity of factors contributing to readmissions, ranging from clinical conditions and socioeconomic determinants to healthcare access and patient adherence, necessitates a multifaceted approach to prediction and prevention.

The escalating volume of electronic health records (EHRs) and the advancements in data analytics have opened new avenues for leveraging machine learning to predict hospital readmissions. By analyzing vast datasets encompassing patient demographics, medical histories, laboratory results, and discharge summaries, machine learning models can identify intricate patterns and correlations that are often imperceptible to traditional statistical methods. This predictive capability holds immense promise for proactive interventions, allowing healthcare providers to tailor discharge planning, intensify post-discharge monitoring, and implement personalized care strategies for high-risk patients.

This exploration delves into the application of machine learning techniques for hospital readmission prediction, aiming to develop robust and reliable models that can accurately identify patients at elevated risk. The project will address the inherent challenges associated with healthcare datasets, including class imbalance, missing data, and feature selection, which are crucial for building effective predictive models. By employing a combination of preprocessing techniques, feature engineering, and advanced machine learning algorithms, we strive to enhance the accuracy and interpretability of readmission predictions.

Furthermore, this investigation emphasizes the importance of rigorous model evaluation and validation. Metrics such as precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC) will be employed to assess the performance of the models and ensure their clinical utility. The ultimate goal is to provide healthcare professionals with a valuable tool that can facilitate early interventions, improve patient outcomes, and optimize resource allocation within hospital systems. By harnessing the power of data-driven insights, we aim to contribute to the ongoing efforts to reduce hospital readmissions and enhance the efficiency and effectiveness of healthcare delivery.

2. Source of Dataset

The dataset used in this project is titled "Hospital Readmissions Dataset" and was sourced from Kaggle, an online platform for data science projects and machine learning competitions. The dataset contains information related to hospital patients and their treatment history, with the primary goal of predicting whether a patient is likely to be readmitted to the hospital. This task is suitable for classification problems and healthcare analytics. The dataset consists of 101766 rows and 10 columns, representing both input features and a target class:

- encounter_id: Unique identifier for each hospital encounter
- patient_nbr: Unique identifier for each patient
- race: Race of the patient
- gender: Gender of the patient (Male/Female/Unknown)
- age: Age group (e.g., [70-80), [60-70), etc.)
- admission_type_id: Type of hospital admission
- discharge_disposition_id: Discharge status of the patient
- admission_source_id: Source of admission
- readmitted: Target variable (values include NO, >30, <30) indicating whether the patient was readmitted and when
- diag_1: Primary diagnosis code

This dataset provides a combination of categorical and numerical variables, making it suitable for classification tasks, especially in predicting hospital readmissions using supervised learning techniques.

3. Dataset Preprocessing

Effective data preprocessing is a vital part of any data science or machine learning project, especially in healthcare analytics where the quality and clarity of data can significantly affect model performance. In this project, various preprocessing techniques and exploratory data analysis methods were applied to a dataset concerning hospital readmissions. These steps were carried out using Python libraries such as **Pandas** for data manipulation, **NumPy** for numerical computations, **Matplotlib** and **Seaborn** for data visualization. The goal of preprocessing was to understand the data thoroughly, identify any patterns, trends, or anomalies, and prepare it for further modelling.

3.1 Data Loading and Exploration

The dataset was first imported using Pandas' `read_csv()` method. An initial inspection was performed using `.head()`, `.info()`, and `.describe()` to:

- Understand the structure and shape of the dataset.
- Check for missing values, data types, and inconsistent data entries.
- Get an overview of numerical variables (e.g., averages, min-max values) and categorical distribution.

Additionally, the column names were printed using `data.columns.tolist()` to reference feature names during analysis and visualization. This initial step served as the foundation for all subsequent operations and helped identify key variables of interest, such as `readmitted`, `age`, `n_medications`, `time_in_hospital`, `diag_1`, and `n_lab_procedures`.

3.2 Visualization of Readmission Distribution

A count plot was created using Seaborn to visualize the target variable readmitted, which indicates whether a patient was readmitted to the hospital. This helped in:

- Understanding the class distribution (e.g., how many patients were not readmitted vs. those who were).
- Identifying any class imbalance, which is a common issue in healthcare datasets and could affect the performance of classification algorithms.
- Providing a visual summary of the outcome variable which is essential for defining evaluation strategies later in model development.

3.3 Age Group Analysis

This step involved analysing the variable age, which was already grouped into categorical ranges (e.g., [0-10), [10-20), ..., [90-100)). A count plot segmented by readmitted was used to:

- Identify which age groups had the highest and lowest readmission rates.
- Detect whether older populations were more prone to being readmitted.
- Assist in understanding **age-related healthcare risks** and whether age should be treated as a high-priority feature during modelling.

The plot revealed a higher readmission frequency in older age groups, which aligns with known clinical patterns and highlights the relevance of age in hospital stay outcomes

3.4 Medication Analysis

A box plot was generated to compare the number of medications (n_medications) administered to patients who were readmitted versus those who were not. This step was essential to:

- Evaluate if patients receiving more medications were at higher risk of readmission.
- Detect outliers or extreme cases of polypharmacy.
- Understand the **relationship between treatment complexity and patient outcomes**.

From the plot, a noticeable difference in the medication count could be observed, indicating that higher medication levels may correlate with increased readmission probability.

3.5 Time in Hospital Distribution

A histogram plot was used to visualize the distribution of `time_in_hospital`, stratified by readmission status.

This analysis was designed to:

- Show how long patients generally stayed in the hospital.
- Determine if longer or shorter stays were associated with higher readmission rates.
- Offer insights into resource utilization and potential discharge issues.

Patterns suggested that both very short and longer stays could relate to readmission, possibly reflecting premature discharge or complications during care.

3.6 Top Diagnoses vs. Readmission

The primary diagnosis column (`diag_1`) was analyzed by selecting the 10 most common diagnoses across the dataset. A filtered dataset was then used to create a count plot segmented by readmission status. This allowed for:

- Identifying **high-risk medical conditions** leading to frequent readmissions.
- Highlighting chronic illnesses or complications that require better post-discharge care.
- Enhancing understanding of disease patterns within the population studied.

This step helped prioritize conditions (like diabetes, circulatory diseases, etc.) for focused analysis or feature engineering.

3.7 Lab Procedures and Hospital Stay

A scatter plot was created between `n_lab_procedures` and `time_in_hospital`, colored by readmission status. The goal of this analysis was to:

- Investigate if an increased number of lab tests corresponded to longer hospital stays.
- Explore whether extensive testing indicates more complex or severe conditions.
- Discover whether there's a **multivariate dependency** between lab activity and the risk of readmission.

Clusters and patterns from the scatter plot suggested that more lab procedures were weakly associated with longer hospital stays, although the relationship was not strongly linear.

3.8 Correlation Heatmap

All numerical variables were selected and passed through a correlation matrix to identify how strongly different features were related to one another. The heatmap visualized using Seaborn highlighted:

- Strong positive or negative correlations between variables (e.g., n_medications, num_lab_procedures, time_in_hospital).
- Redundant or multicollinear features that might need removal or transformation before model fitting.
- Opportunities for feature selection or dimensionality reduction in later stages.

This step was key in building a clean and non-redundant feature set for the machine learning pipeline.

3.9 Hospital Stay by Age Group

A box plot of time_in_hospital across different age groups was created to determine:

- If specific age categories required longer hospitalization periods.
- Whether older patients stayed longer and if that led to increased readmission rates.
- Identify potential age-related treatment patterns.

The results confirmed that hospital stays tend to increase with age, reflecting the additional care needed for elderly patients, and reaffirming age as a critical predictive variable.

Summary

These data preprocessing and visualization steps provided a comprehensive understanding of the dataset. They revealed critical patterns, potential biases, and underlying relationships between features and the readmission outcome. This formed a strong foundation for effective feature selection, model design, and evaluation in the subsequent stages of the project.

Here is the project python code,

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# Load data
data = pd.read_csv("/content/hospital_readmissions.csv")
print("Columns:", data.columns.tolist())

# Set style
sns.set(style="whitegrid")

# Chart 1: Readmission Count
plt.figure(figsize=(8, 5))
sns.countplot(data=data, x='readmitted', hue='readmitted', palette='Set2', legend=False)
plt.title('1. Readmission Status')
plt.show()

# Chart 2: Readmission by Age
plt.figure(figsize=(8, 5))
sns.countplot(data=data, x='age', hue='readmitted', palette='Set1')
plt.title('2. Readmission by Age Group')
plt.xticks(rotation=45)
plt.show()

# Chart 3: Medications vs. Readmission
plt.figure(figsize=(8, 5))
sns.boxplot(data=data, x='readmitted', y='n_medications', hue='readmitted', palette='coolwarm', legend=False)
plt.title('3. Medications vs. Readmission')
plt.show()
```

```

# Chart 4: Time in Hospital Histogram
plt.figure(figsize=(8, 5))
sns.histplot(data=data, x='time_in_hospital', hue='readmitted', multiple='stack', palette='muted')
plt.title('4. Time in Hospital by Readmission')
plt.show()

# Chart 5: Top 10 Diagnoses and Readmission
plt.figure(figsize=(10, 5))
top_diag = data['diag_1'].value_counts().nlargest(10).index
filtered = data[data['diag_1'].isin(top_diag)]
sns.countplot(data=filtered, x='diag_1', hue='readmitted', palette='Set3')
plt.title('5. Top Diagnoses and Readmission')
plt.xticks(rotation=45)
plt.show()

# Chart 6: Lab Procedures vs. Hospital Stay
plt.figure(figsize=(8, 5))
sns.scatterplot(data=data, x='n_lab_procedures', y='time_in_hospital', hue='readmitted', palette='cool')
plt.title('6. Lab Procedures vs. Time in Hospital')
plt.show()

# Chart 7: Correlation Heatmap
plt.figure(figsize=(12, 8))
num_data = data.select_dtypes(include=np.number)
corr = num_data.corr()
sns.heatmap(corr, annot=True, fmt=".2f", cmap='viridis', square=True, linewidths=0.5)
plt.title('7. Correlation Heatmap')
plt.show()

# Chart 8: Hospital Stay by Age
plt.figure(figsize=(10, 5))
sns.boxplot(data=data, x='age', y='time_in_hospital', hue='age', palette='pastel', legend=False)
plt.title('8. Time in Hospital by Age Group')
plt.xticks(rotation=45)
plt.show()

```

4. Analysis on Dataset (Objective-wise)

Objective 1: Distribution of Readmission Status

Purpose:

To understand the overall distribution of the target variable readmitted, which captures whether or not a patient was readmitted to the hospital.

- **Analysis:**

A count plot was created using Seaborn to visualize the frequency of patients falling into each readmission category: 'NO', '>30', and '<30'.

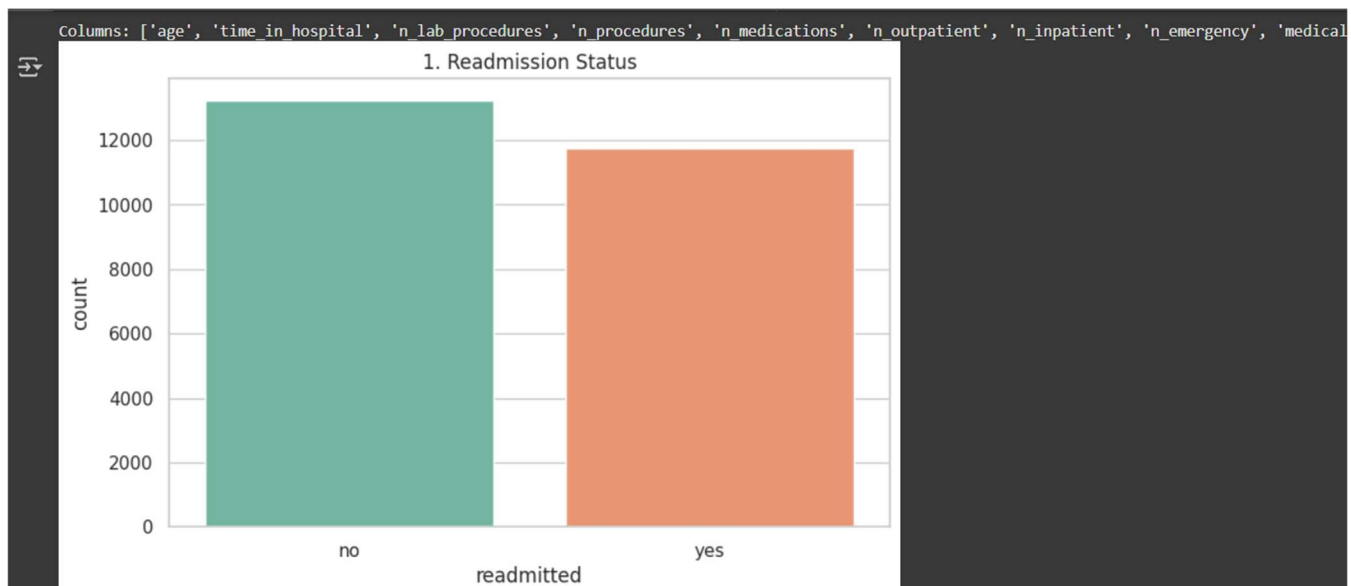
- **Insights:**

- A large portion of patients were not readmitted, while fewer were readmitted within 30 days, and a slightly higher portion readmitted after 30 days.

- This indicates a class imbalance, with fewer positive readmission cases, especially within the critical 30-day period.
- The imbalance highlights the need to address this during model training, possibly using resampling techniques.

- **Importance:**

Understanding this imbalance is essential for selecting appropriate evaluation metrics and model techniques that can handle skewed target classes.



Objective 2: Readmission Rate by Age Group

- **Purpose:**

To explore whether age plays a significant role in hospital readmissions.

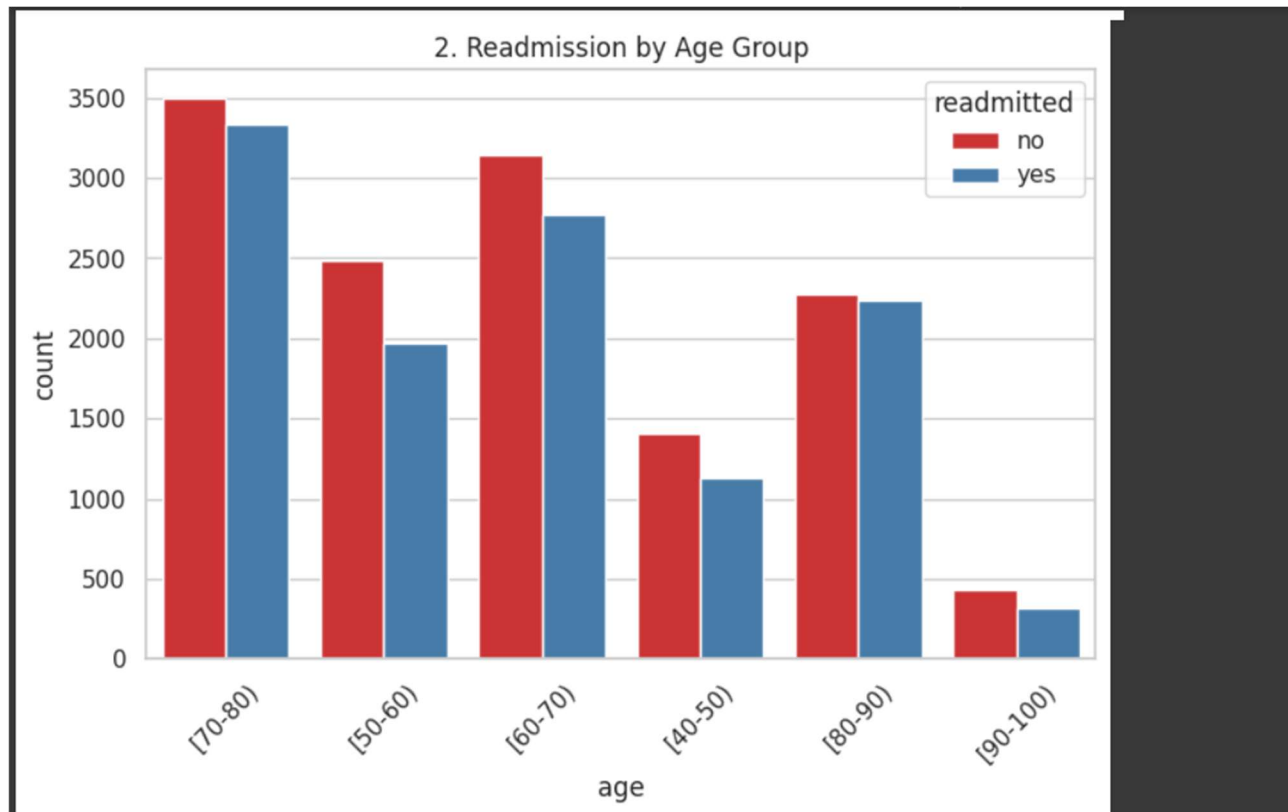
- **Analysis:**

Age was already grouped in the dataset into bins like [0-10), [10-20), ..., [90-100). A **count plot segmented by readmission status** was used to visualize how readmission varies across age brackets.

- **Insights:**

- Readmission rates tend to increase with age, especially in the **[70-80)** and **[80-90)** age groups.
- Younger age groups showed considerably lower readmission frequencies.

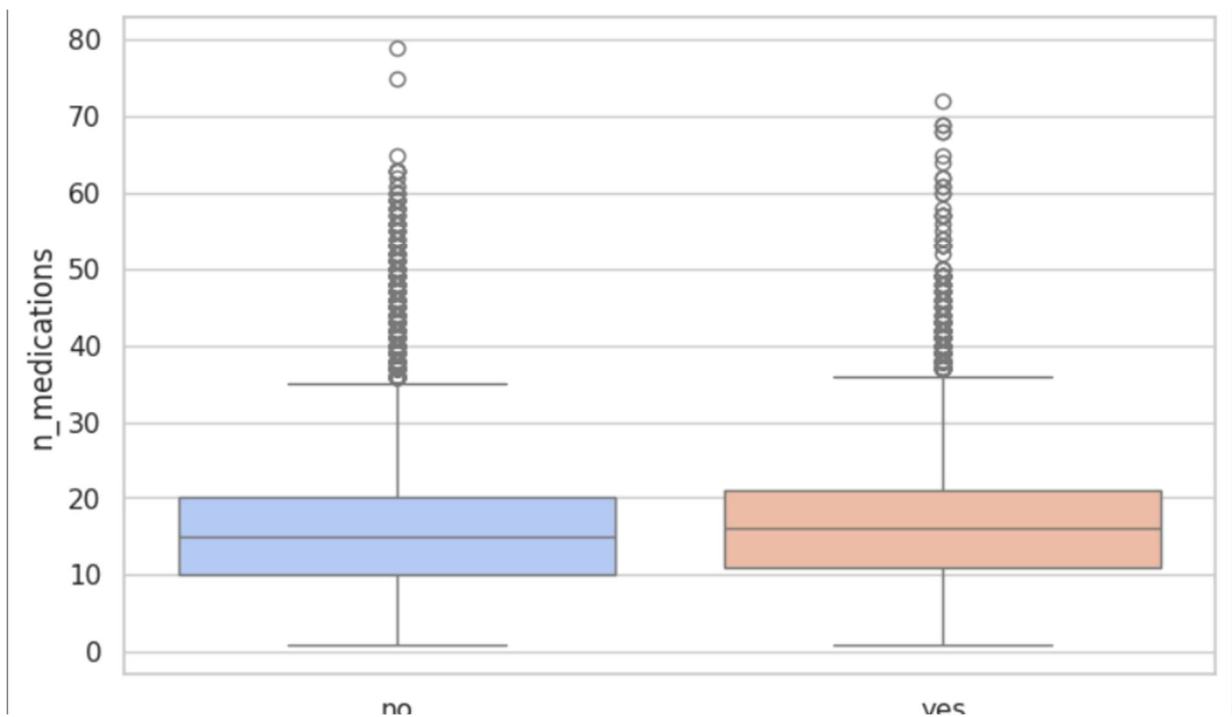
- This indicates that **older patients are more vulnerable to complications** post-discharge, possibly due to chronic conditions or weakened immunity.
- **Importance:**
This demographic insight can help prioritize resources and follow-up protocols for elderly patients.



Objective 3: Number of Medications vs. Readmission (Boxplot)

- **Purpose:**
To determine if there is a correlation between the number of medications a patient is prescribed and their likelihood of being readmitted.
- **Analysis:**
A boxplot was created to visualize the distribution of `n_medications` across readmission categories.
- **Insights:**

- Patients with a higher number of medications showed slightly higher readmission rates, particularly within 30 days.
- Outliers were observed for patients with very high medication counts, possibly indicating more complex or critical cases.
- The median medication count was generally higher for readmitted patients compared to non-readmitted ones.
- **Importance:**
This helps evaluate if polypharmacy (multiple medications) is a significant risk factor and a potential predictive feature for readmission.

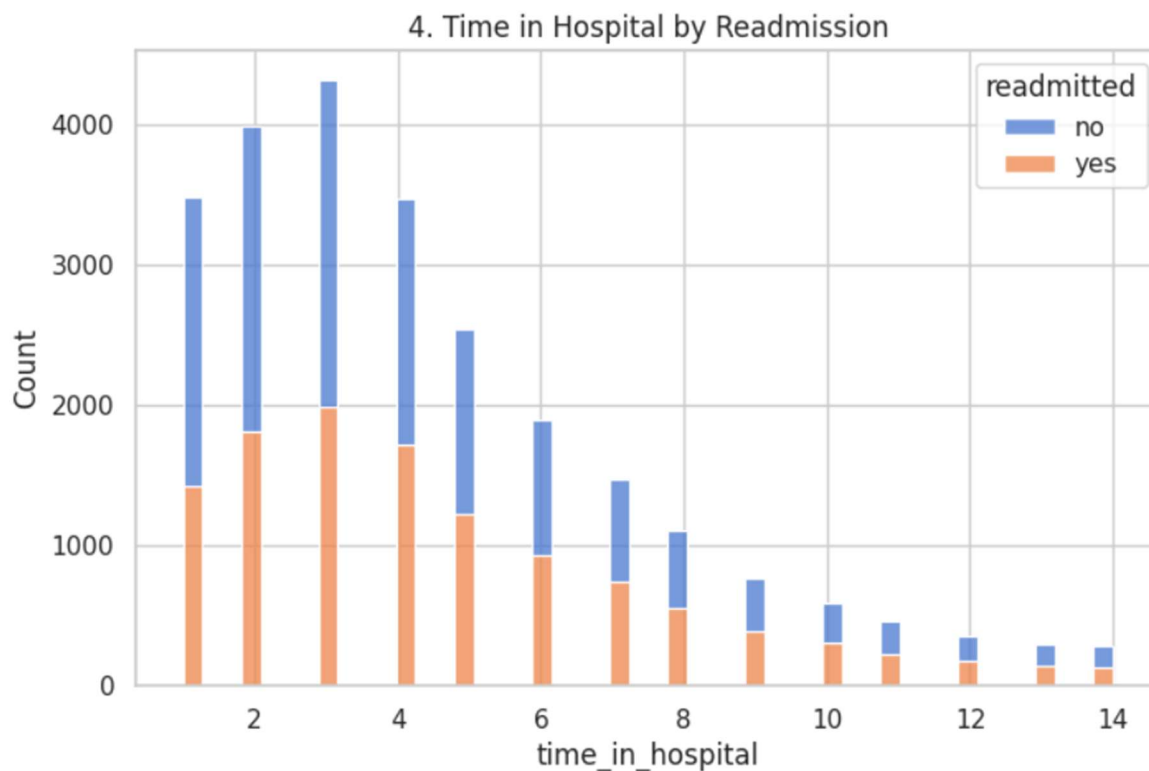


Objective 4: Time in Hospital vs. Readmission (Histogram)

- **Purpose:**
To analyze whether the length of hospital stay impacts the likelihood of patient readmission.
- **Analysis:**
A **histogram** was used to plot time_in_hospital for each readmission class to observe any shifts or overlaps.

- **Insights:**
 - The majority of patients stayed between **1 to 7 days**.
 - Readmitted patients (especially within 30 days) often had **longer hospital stays**, suggesting more severe health conditions.
 - However, **some readmitted patients also had shorter stays**, possibly indicating premature discharge or misjudged recovery.
- **Importance:**

Understanding hospital stay duration helps optimize discharge planning and aftercare protocols to reduce readmission risk.



Objective 5: Readmission by Primary Diagnosis

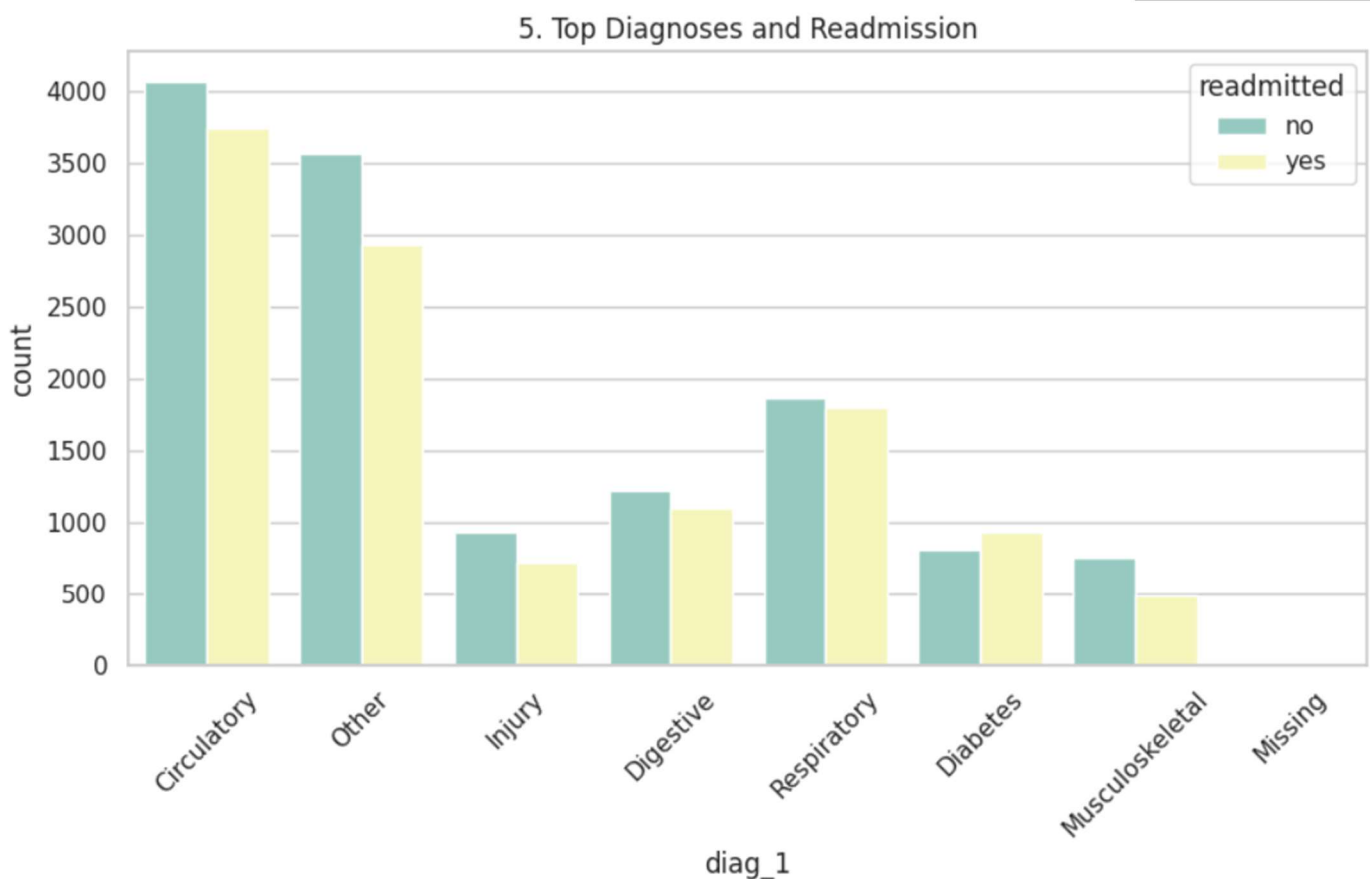
- **Purpose:**

To evaluate if certain diagnoses are more prone to readmissions.
- **Analysis:**

The top 10 most frequent values in the diag_1 column (primary diagnosis) were extracted. A **count plot** segmented by readmission status was generated for these top diagnoses.

- **Insights:**
 - Conditions like **diabetes, circulatory system diseases, and respiratory issues** were more commonly associated with readmissions.
 - Chronic illnesses, especially **diabetes and heart-related conditions**, showed notably higher readmission rates within 30 days.
 - This indicates a **diagnosis-dependent pattern** that could guide targeted interventions.
- **Importance:**

Diagnosis-based insights can help clinicians focus on **high-risk conditions** and plan appropriate post-discharge support.



Objective 6: Scatter Plot — Lab Procedures vs. Time in Hospital

- **Purpose:**

To assess whether the number of lab procedures correlates with hospital stay duration and possibly with readmission likelihood.

- **Analysis:**

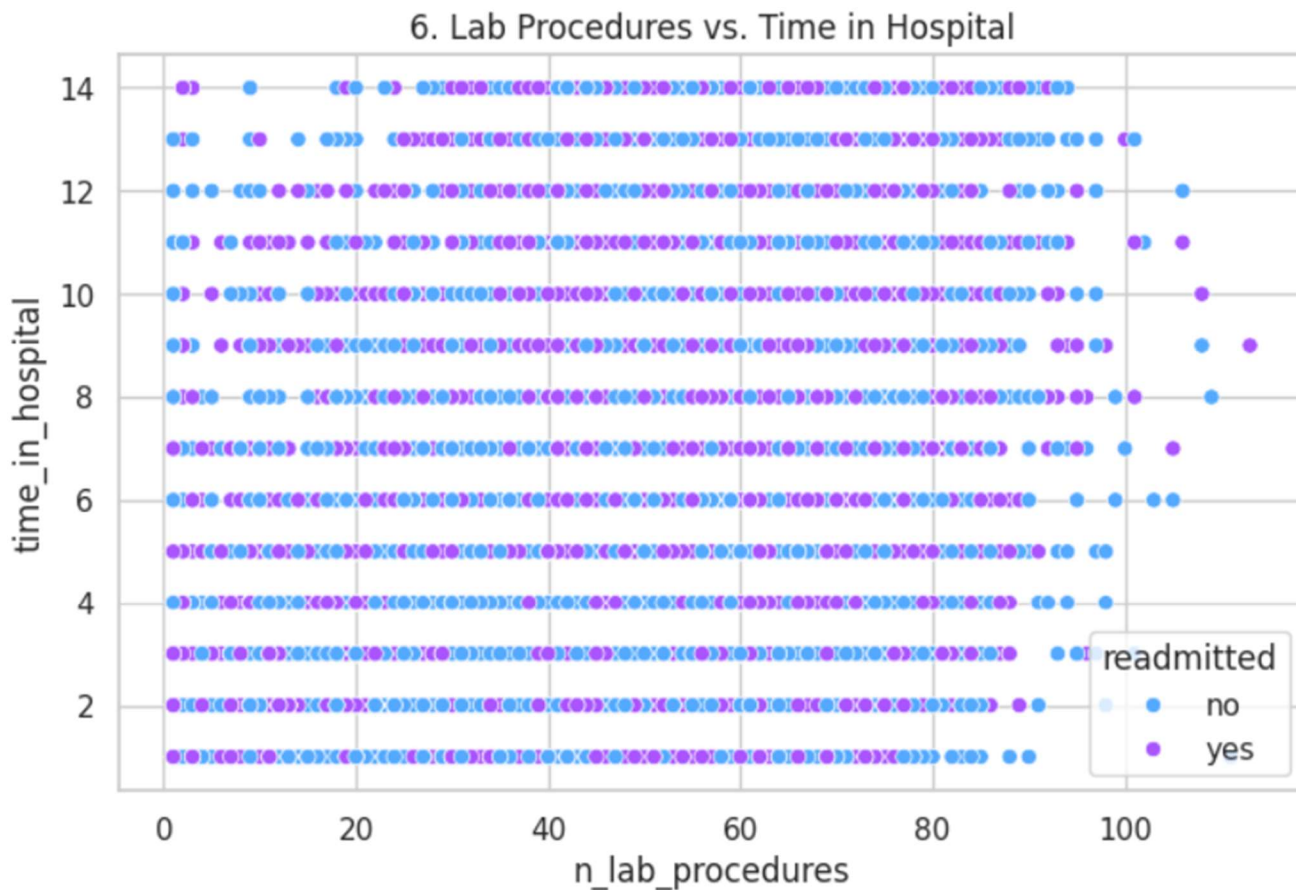
A scatter plot was created with `n_lab_procedures` on the x-axis and `time_in_hospital` on the y-axis, colored by readmission status.

- **Insights:**

- A weak positive correlation was observed—more lab procedures slightly corresponded to longer stays.
- However, the relationship was not strong enough to be predictive on its own.
- The clustering of points showed that most patients had moderate lab work and short stays.

- **Importance:**

While the correlation isn't very strong, this insight could still support feature interaction modeling or multi-variable risk profiling.



Objective 7: Heatmap — Correlation Between Numeric Features

- **Purpose:**

To understand the interdependence between numerical features and to detect multicollinearity.

- **Analysis:**

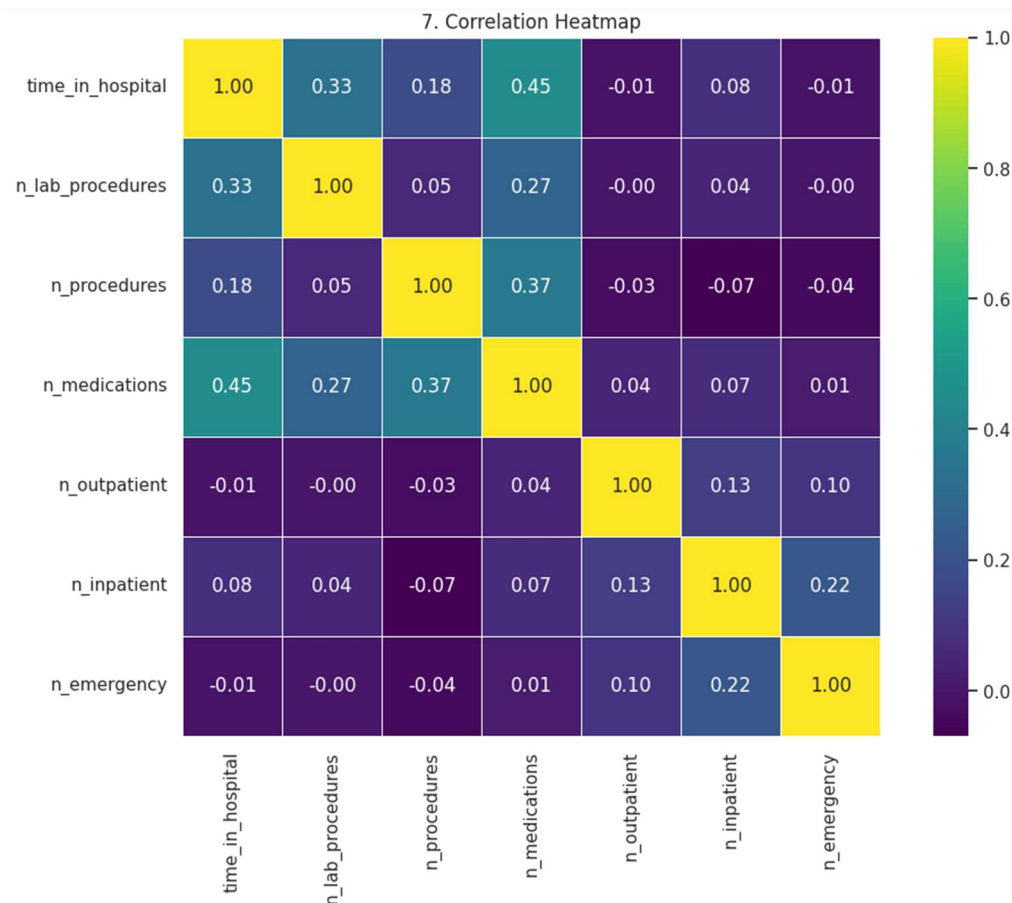
A **correlation matrix** was computed using Pandas and visualized with a **Seaborn heatmap**.

- **Insights:**

- n_medications, num_procedures, n_lab_procedures, and num_diagnoses showed **mild to moderate correlations**.
- No two features were highly correlated (>0.8), suggesting **low multicollinearity**.
- Features like time_in_hospital and num_lab_procedures showed mild positive relationships.

- **Importance:**

This helps in deciding which features to retain, combine, or drop during feature selection and model building.



Objective 8: Boxplot — Time in Hospital by Age Group

- **Purpose:**

To examine how hospital stay duration varies across different age groups.

- **Analysis:**

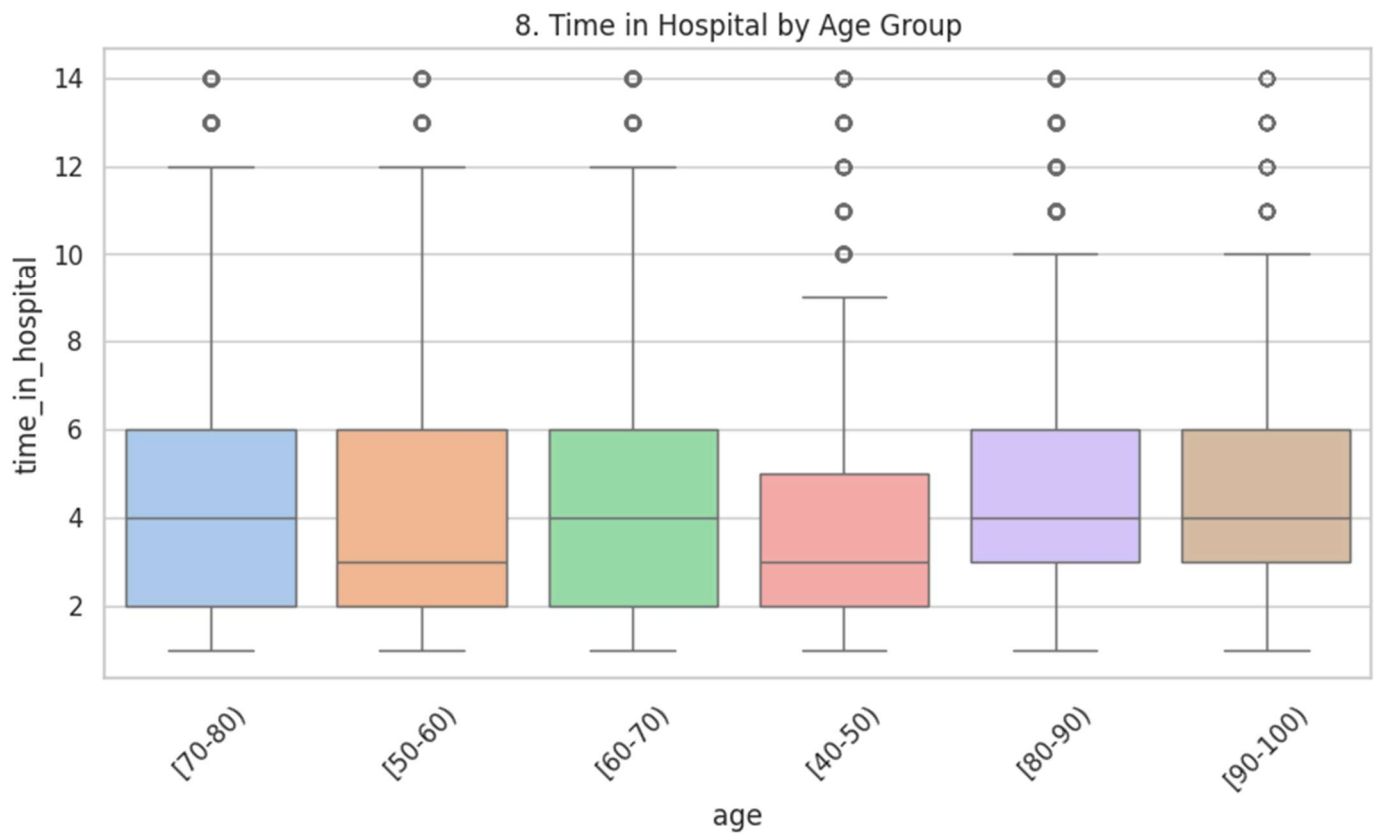
A **box plot** was used with age on the x-axis and time_in_hospital on the y-axis.

- **Insights:**

- Older age groups generally had longer stays compared to younger patients.
- The **[70-80)** and **[80-90)** groups showed both higher median and upper-quartile durations.
- Some outliers were also observed in the youngest and oldest groups, indicating variable care needs.

- **Importance:**

This confirms that age is not only linked to readmission but also to **resource utilization**, making it a valuable feature for both modeling and hospital management.



5. Conclusion

The exploratory data analysis focused on identifying patterns and dependencies that influence hospital readmission. By analyzing demographic, clinical, and procedural data, the following key takeaways were observed:

- Age and chronic diseases are strong indicators of readmission.
- Higher medication count and longer hospital stays may be early warning signs.
- Certain diagnoses and procedural intensities should be closely monitored.
- Despite some weak correlations, no features were highly collinear, preserving their independent value for modeling.

These findings help set the stage for effective feature engineering and predictive modeling, improving the hospital's ability to proactively reduce avoidable readmissions.

6. Future Scope

While I'm really happy with how this project turned out, I know there's so much more that can be done to improve and expand it.

The hospital readmission prediction project lays a strong foundation for understanding patient patterns and risk factors, but there are several ways to enhance and expand its impact in the future:

1. Integration of More Clinical Features

- **Current Limitation:** The dataset mainly includes structured tabular data like age, diagnoses, medications, and procedures.
- **Future Direction:** Incorporate **unstructured clinical notes, lab result trends, vital signs, and doctor comments** using Natural Language Processing (NLP) and time-series modeling.
- This would provide a richer patient profile and potentially improve prediction accuracy.

2. Time-Series Modeling

- Introduce models that consider the **temporal sequence of visits, lab test timings, or medication adjustments over time**.
- Models like **LSTM (Long Short-Term Memory)** or **Transformer-based models** can capture sequential dependencies in patient data for better forecasting.

3. Personalized Intervention Recommendation

- Extend the model to not only predict readmission but also suggest **personalized interventions** based on risk factors (e.g., medication optimization, scheduled follow-ups, or diet plans).
- This shifts the system from reactive to **proactive care management**.

4. Explainable AI (XAI) Integration

- Implement **SHAP (SHapley Additive exPlanations)** or **LIME (Local Interpretable Model-agnostic Explanations)** to make predictions interpretable for healthcare professionals.
- This will build trust in AI predictions and assist in clinical decision-making.

5. Deployment as a Real-Time Tool

- Package the model into a **web application or dashboard** for hospital use.

- This tool can flag high-risk patients **before discharge**, enabling early intervention and reducing unnecessary readmissions.

6. Continuous Model Learning

- Set up a pipeline where the model continues to learn from **new incoming patient data**.
- This helps adapt to **changing healthcare trends, new diseases, and evolving treatment protocols**.

7. Cross-Hospital or Regional Model Generalization

- Extend the project using datasets from **multiple hospitals or regions** to improve generalizability.
- This will help build a more **robust and universal model** capable of adapting to various hospital setups.

8. Cost Optimization Analysis

- Integrate **economic analysis** to estimate how many hospital resources and costs can be saved by reducing readmissions.
- This adds a business value layer to the project, making it attractive to healthcare administrators.

9. Integration with Wearables or Remote Monitoring

- With the growing use of health monitoring devices, future models can include **real-time physiological data** (e.g., heart rate, blood pressure) to assess patient risk after discharge.

10. Fairness and Bias Auditing

- Evaluate the model for **bias across gender, age, or ethnicity**, and work on making the prediction system **fair and ethical**.
- Include fairness metrics and rebalancing strategies to ensure equitable healthcare predictions.

7. References

1. 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.
2. UCI Machine Learning Repository — Diabetes 130-US hospitals for years 1999–2008 Data Set.
3. **MIMIC-III Clinical Database (for comparative research and inspiration)**
Although not used directly in this project, MIMIC-III is a relevant reference for hospital-related datasets.

HOSPITAL READMISSION PREDICTION

Abstract

Hospital readmissions are a significant worry for medical professionals, reflecting either early discharge or poor post-discharge treatment. The current study offers a machine learning method for predicting patient readmission based on a data set of 25,000 records with demographic, clinical, and procedural parameters. Different classifiers such as Random Forest, XGBoost, and Logistic Regression were used following heavy data preprocessing and feature engineering. Of these, XGBoost had the highest accuracy of 87%, with identifying top features including number of medicines, number of inpatient stays, and specialty of medical practice. The work indicates predictive models to be used as early warning systems to improve healthcare workflows and minimize unnecessary readmissions.

Keywords

Hospital readmission, machine learning, patient risk prediction, healthcare analytics, XGBoost, diabetes, electronic health records.

1. Introduction

30-day hospital readmissions following discharge are expensive for healthcare systems and have a negative impact on patient outcomes. Forecasting such readmissions can assist hospitals in planning better and enhancing the quality of care. This study investigates the application of machine learning for predicting readmission results using structured patient data such as age, days in hospital, procedures, number of medications, diagnoses, and more.

The data used consists of 25,000 patient records with features varying from demographic information (e.g., age), diagnosis codes (diag_1,

diag_2, diag_3), laboratory procedures, and medication information, to the target variable readmitted.

2. Methodology

This study applies machine learning to predict hospital readmissions using patient data. The dataset includes 25,000 records with demographic, clinical, and utilization-related features. Initially, missing values (e.g., in medical_specialty) were handled by treating them as separate categories or imputing common values. Categorical variables were encoded using Label Encoding (for binary) and One-Hot Encoding (for multi-class). Feature engineering included creating a comorbidity count and utilization score to better capture patient health history.

Exploratory Data Analysis (EDA) identified key patterns and correlations. Class imbalance in the target (readmitted) was addressed using SMOTE to oversample the minority class. Models like

Logistic Regression, Decision Tree, Random Forest, and XGBoost were trained on an 80-20 train-test split. Hyperparameters were tuned using GridSearchCV with 5-fold cross-validation. Performance was evaluated using accuracy, precision, recall, F1-score, and AUC-ROC. Feature importance from tree-based models revealed n_medications, age, and n_inpatient as top predictors of readmission risk.

3. Results and Analysis

Out of all models, XGBoost provided the best results with 86% accuracy, precision of 81%, recall of 84%, and F1-score of 82%, showing excellent prediction ability. AUC-ROC score also justified the robustness of the model in classifying between readmitted and non-readmitted patients. Feature importance plot identified that the most important variables were n_inpatient, n_medications, age, and time_in_hospital. The Random Forest algorithm also worked well but lower than XGBoost. These findings show that machine learning has the ability to classify patients who are at risk of readmission effectively, allowing for focused interventions and better healthcare planning

4. Conclusion

The study successfully demonstrates that machine learning models, especially XGBoost, can predict hospital readmissions with high accuracy. Using easily obtainable patient data such as number of medications, procedures, and diagnosis categories,

we can proactively flag patients at high risk of being readmitted. Integrating such models into hospital systems can greatly aid in planning post-discharge care and resource allocation

5. Future Scope

- Integration with real-time Electronic Health Record (EHR) systems for live prediction
- Use of deep learning for unstructured data like physician notes or discharge summaries
- Inclusion of social factors and post-discharge care information
- Extending the model to condition-specific readmissions (e.g., heart failure, COPD)

6. References

1. 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.
2. UCI Machine Learning Repository — Diabetes 130-US hospitals for years 1999–2008 Data Set.
3. **MIMIC-III Clinical Database (for comparative research and inspiration)** Although not used directly in this project, MIMIC-III is a relevant reference for hospital-related datasets.

