PREDICTIVE PATIENT FLOW MODEL FOR HOSPITAL OVERCROWDING ANALYSIS USING MACHINE LEARNING AND EXPLAINABLE AI

IKMEE U-DAIDEE

642437002

FATONI UNIVERSITY

1445/2024

PREDICTIVE PATIENT FLOW MODEL FOR HOSPITAL OVERCROWDING ANALYSIS USING MACHINE LEARNING AND EXPLAINABLE AI

IKMEE U-DAIDEE

642437002

A PROJECT REPORT SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENT FOR THE DEGREE OF BACHELOR OF SCIENCE (DATA SCIENCE AND ANALYTICS)

FATONI UNIVERSITY

1445/2024

TABLE OF CONTENTS

Table of Contents

LIST OF TABLES

LIST OF FIGURES

CHAPTER I
INTRODUCTION

1.1 PROJECT OVERVIEW

The issue of hospital overcrowding has been a recurring issue, leading to long waiting hours and delayed admissions to intensive care wards. This has been identified as a major challenge facing hospitals globally [1]. Overcrowding occurs when healthcare is forced to operate beyond its capacity due to a shortage of medical staff and an excessive number of patients seeking medical treatment [2]. Hospital overcrowding is primarily caused by factors such as unnecessary patient visits, lack of inpatient beds, and prolonged waiting times for available beds in wards. Research indicates that unnecessary visits often stem from inadequate standard procedures, while a shortage of inpatient beds exacerbates delays in emergency departments (EDs) and contributes to increased mortality rates among vulnerable populations, such as chronic kidney disease patients [3]. To mitigate these effects, healthcare systems can implement several strategies. Enhancing bed management and fostering departments can streamline patient flow and reduce boarding times [4]. Additionally, optimizing staffing levels in outpatient departments and employing queuing models to manage patient arrivals can significantly decrease wait times and improve overall operational effectiveness [3]. These measures can help alleviate overcrowding and enhance patient care quality.

Patient flow plays a critical role in hospital overcrowding, as inefficient management of patient movement can lead to significant delays and negative outcomes. Research indicates that effective patient flow management, including the use of artificial intelligence (AI) tools, can enhance the forecasting and monitoring of patient admissions, transfers, and discharges, thereby alleviating overcrowding in hospitals [5]. For instance, the implementation of discharge lounges has been shown to improve patient flow by increasing discharge rates and reducing turnaround times, which directly correlates with decreased overcrowding [6]. Additionally, systematic reviews highlight that managing patient flows across various hospital departments is essential, as disruptions in one area can impact the entire system. Factors such as prolonged waiting times and inadequate staffing in emergency departments exacerbate overcrowding, underscoring the need for targeted interventions to streamline patient flow [7]. Overall, optimizing patient flow is vital for improving hospital efficiency and patient care quality. The emerging technique of Artificial Intelligence (AI) has made it possible to manage overcrowding in emergency departments hence getting more attention in the community.

This project proposes comparison between k-Nearest Neighbor (KNN) and Random Forrest model of Machine Learning to be employed and trained using hospital admission data encompassing attributes such as diagnosis, consultancy episodes, number of admission and demography. The model will identify patterns and trends to predict which diagnosis requires the patient to have longer hospital stays or readmissions to help stakeholders to prioritize resource allocation accordingly. Apart from that, this project also emphasizes data visualization as it is essential for understanding and addressing the relationship between diagnosis and overcrowding in hospitals. It can help identify patterns, bottlenecks, and trends in the data, offering actionable insights for improving patient throughput and resource management. Data visualization using Python with libraries such as Matplotlib, Seaborn, and Plotly is a powerful approach to transforming raw data into meaningful insights through graphical representation.

1.2 PROBLEM STATEMENT

a) Insufficient Understanding of Diagnosis-Specific Flow Patterns: The absence of data visualized of how specific diagnoses contribute to patient flow dynamics creates challenges in identifying which medical conditions are most closely associated with overcrowding at different times.

b) Difficulty in Integrating Historical Data for Predictions: Hospitals face challenges in integrating historical patient diagnosis data to create accurate prediction models, limiting their ability to anticipate and mitigate future overcrowding effectively.

c) Lack of Explainability in Prediction Models: Stakeholders struggle with interpreting and understanding the predictive models used for anticipating overcrowding. The absence of explainability makes it difficult for healthcare professionals to trust and act upon predictions, which limits the effectiveness of these models in decision-making and patient flow management.

1.3 OBJECTIVE

a) Employ data visualization techniques to explore and extract key features from the dataset. This includes identifying trends in hospital admissions, diagnosis patterns, and demographic impacts (such as age, gender, and types of admissions), which are relevant to predicting hospital overcrowding.

b) Build a machine learning model to predict hospital overcrowding based on historical hospital admission data and diagnosis patterns. The model will classify periods of potential overcrowding to assist in hospital resource planning and management.

c) Incorporates Explainable AI (XAI) through LIME to provide transparent, interpretable explanations of the model's predictions. By using LIME, this study seeks to enhance stakeholders' understanding of the factors driving individual predictions, enabling healthcare professionals to make informed, data-driven decisions.

1.4 SCOPE OF STUDY

- Data use: Admitted Patient Care activity in England for the financial year 2023-24.
- Required Tools:

Category	**Tools/Technologies**	**Description**

Programming Language	Python 3.8 or higher	Used for developing code for data analysis, model creation, and dashboard development.
Libraries and Frameworks	Scikit-learn	Used for building machine learning models, including KNN and Random Forest (RF), and evaluating model performance.
	XGBoost	For Gradient Boosting to improve prediction performance.
	Matplotlib, Seaborn, Plotly	Used for creating graphs and visualizing data insights.
	Pandas, Numpy	For handling and processing data in tabular format and performing numerical computations.
	LIME	Helps explain AI model results (Explainable AI).
	Streamlit	Used for building web applications to showcase model predictions and results.
Development Tools	Jupyter Notebook	Used for writing code and conducting data analysis during the development phase.
	Visual Studio Code	Used for writing and managing code for model development and dashboard creation.
Hardware Requirements	Operating System: Windows 10	Supports the installation and execution of all necessary tools and software.
	RAM: At least 8GB	Ensures smooth processing of large datasets.
	Processor: Intel i5 or equivalent	Needed to run the program and perform high-performance calculations.

Version Control	Git and GitHub	- Git: Used to track changes in the code and manage project versions.
		- GitHub: Used for cloud storage of code, collaboration, and version control.

Table 1: Required Tools

- Target User:

  The target users of this study are hospital stakeholders who aim to manage resources effectively to reduce overcrowding in hospitals and improve the quality of healthcare services. They can use data and analysis to plan strategies for resource management and enhance operational efficiency. Additionally, it includes healthcare providers (doctors and nurses) who need in-depth information about patient flow, which will help them make better decisions in patient care.

1.5 PROJECT SIGNIFICANCE

This project holds significant potential for enhancing hospital resource management and improving patient care by predicting patient flow patterns and identifying factors that contribute to longer hospital stays or readmissions. With limited resources and frequent issues of overcrowding, hospitals often face challenges that impact both operational efficiency and patient safety. By developing a predictive model based on hospital admission data, this project enables administrators to allocate resources proactively, reducing bottlenecks and optimizing patient throughput. This data-driven approach supports strategic decision-making for staffing, facility management, and scheduling, ultimately contributing to a more responsive and well-organized healthcare environment.

Additionally, this project integrates explainable AI using LIME, allowing healthcare providers to understand the reasons behind each prediction. This transparency builds trust in the model’s recommendations and helps stakeholders make informed decisions aligned with patient needs. With insights into which factors most influence patient outcomes, hospitals can prioritize high-risk cases, implement preventive measures, and address potential issues before they escalate. The project not only enhances hospital efficiency but also sets a foundation for broader healthcare applications, offering scalable solutions that can improve patient care and operational efficiency across various healthcare settings.

1.6 CHAPTER ORGANIZE

Chapter I: Introduction

This chapter introduces the main problem that this research aims to address, specifically the issue of overcrowding in hospitals, and highlights the importance of developing predictive models to manage hospital resources more effectively. The chapter also outlines the objectives and scope of the research, discusses the limitations that may arise, and explains the methodology employed throughout the study.

Chapter II: Literature Review

This chapter reviews existing literature related to the topic of hospital overcrowding prediction. It covers various methods currently used in the field, such as machine learning applications for predicting hospital-related problems, including emergency room data and patient information for model development. This chapter also examines the use of models like Random Forest, K-Nearest Neighbors (KNN), and Gradient Boosting, as well as the integration of Explainable AI techniques, such as LIME (Local Interpretable Model-agnostic Explanations), to enhance model transparency and understanding.

Chapter III: Methodology

This chapter provides a detailed explanation of the research methodology employed in the study. It describes the data preprocessing steps, including the selection of relevant features, handling missing data, and the application of techniques. The chapter then outlines the development and training of the predictive models using KNN, Random Forest, and Gradient Boosting. Furthermore, it discusses the evaluation metrics used to assess model performance, such as accuracy, precision, recall, and F1-score.

Chapter IV: Propose Solution and Result

This chapter presents the proposed solution to address the issue of overcrowding in hospitals by incorporating new features. These features are designed to enhance the model’s predictive power and interpretability. The chapter also provides the results of model testing, comparing the performance of KNN, Random Forest, and Gradient Boosting. It includes an analysis of the model’s accuracy, precision, recall, and F1-score, as well as the use of LIME to explain the decision-making process of the Gradient Boosting model.

Chapter V: Conclusion and Discussion

This chapter summarizes the key findings of the project, discusses the implications of the results, and highlights the significance of using machine learning and Explainable AI in addressing hospital overcrowding. It also acknowledges the limitations of the study and provides recommendations for future research to further enhance the model's applicability and functionality.

CHAPTER II
LITERATURE REVIEW

2.1 DATA ANALYSIS AND VISUALIZATION IN PATIENT FLOW MANAGEMENT

Data analysis and visualization are crucial tools for managing and interpreting healthcare data, significantly enhancing the management of medical resources and improving patient service delivery [8]. Data visualization enables healthcare executives and staff to view data in an accessible format, making it easier to identify patterns, trends, and bottlenecks in patient admissions and movements [9]. Python libraries such as Matplotlib, Seaborn, Plotly, Bokeh, Altair, and ggplot are used to create detailed visualizations that support better decision-making [10]. Utilizing these visualization tools allows hospitals to optimize resource allocation, manage bed occupancy more effectively, and reduce patient waiting times clearly.

- Matplotlib is an established and popular library used for creating various types of graphs, such as line charts and histograms. It provides flexibility and detail for data visualization, making it a fundamental tool for healthcare data analysis and visualization [10].
- Seaborn builds on Matplotlib and is designed for statistical data visualization, facilitating the creation of complex plots like Heatmaps and Pair plots, which help in exploring data relationships and trends [10].
- Plotly supports interactive and 3D graph creation, enhancing the effectiveness of detailed dashboards and making it suitable for visualizing patient flow data [10].
- Bokeh focuses on creating interactive and web-based visualizations, ideal for detailed and specific data representation [10].
- Altair is known for its simple syntax for creating statistical graphs and interactive visualizations, making it suitable for in-depth data analysis and presentation [10].

- ggplot adapted from R, uses a grammar of graphics approach, allowing for straightforward and clear graph creation, which is beneficial for detailed data interpretation [10].

An example of data visualization application is a study in Southwest Ethiopia, where a health information system was developed to aggregate data from 21 healthcare facilities over 41 months. Using Python Sankey diagrams, the researchers visualized patient flow and employed machine learning algorithms to achieve high prediction accuracy for outpatient flows [11]. The study found that Sankey diagrams effectively visualized patient flow across healthcare facilities, enabling stakeholders to monitor and predict patient movements with high accuracy (up to 85%) [11].

Additionally, Exploratory Data Analysis (EDA) using Python libraries such as Pandas and Matplotlib plays a crucial role in cleaning and visualizing healthcare data. This aids in discovering trends and relationships that inform patient care strategies [12]. Data visualization enhances understanding of complex datasets, allowing healthcare professionals to identify patterns and relationships crucial for evidence-based decision-making [12], [13]. Interactive dashboards also enable rapid data analysis, significantly improving response times in clinical settings and potentially saving lives [14].

2.2 MACHINE LEARNING TECHNIQUES FOR MANAGING OVERCROWDING

Managing hospital overcrowding is a critical challenge that directly affects the quality of patient care and resource management within healthcare facilities. Machine learning (ML) techniques play an essential role in forecasting patient flow, optimizing resource allocation, and enhancing existing services to reduce congestion.

Predicting patient flow and hospital admissions can be achieved through predictive modeling techniques such as Random Forest, K-Nearest Neighbors (KNN), and Support Vector Machines (SVM). These models are instrumental in analyzing patient data and forecasting future admission volumes. For instance, a study conducted in the southwestern region of Ethiopia utilized ML models to predict outpatient and inpatient flow, achieving an accuracy of up to 85% for outpatient admissions and 83% for predicting overall patient flow. Techniques like NearMiss, SMOTE, and SMOTE-Tomek were employed to address data imbalance issues commonly found in patient data, significantly enhancing model performance and reliability [11]. These models are invaluable for anticipating patient demand and effectively planning hospital resource allocation, helping reduce overcrowding and improving service delivery efficiency.

In the context of managing patient flow in emergency departments, classification algorithms have been applied to predict and manage patient length of stay (LOS). A study in Nigeria explored various classification techniques, including SVM, Classification and Regression Trees (CART), and Random Forest, to forecast LOS in emergency rooms. The study found that the SVM algorithm performed the best, with an accuracy of 0.986984 and a Mean Squared Error (MSE) of 0.358594, demonstrating its effectiveness in predicting LOS and managing patient flow [15]. This high accuracy allows hospitals to better manage resources and patient treatment times, thereby reducing congestion and enhancing service efficiency.

Additionally, clustering techniques are pivotal in analyzing and managing inpatient bed demand by identifying patterns and trends within the data, which facilitates accurate predictions of bed requirements. The study "Machine Learning Based Forecast for the Prediction of Inpatient Bed Demand" employed K-means clustering combined with Support Vector Machine Regression (K-SVR) to predict inpatient bed demand. The study achieved a Mean Absolute Percentage Error (MAPE) ranging between 0.49% and 4.10%, highlighting the effectiveness of clustering and regression techniques in improving bed management and alleviating hospital congestion [16]. These techniques enable hospitals to better plan admissions, reduce waiting times, and optimize the allocation of limited bed resources.

Overall, the application of machine learning techniques in hospital overcrowding management demonstrates significant potential in forecasting patient flow, optimizing resource allocation, and enhancing medical services. These approaches contribute to reducing congestion and improving the overall patient care experience.

2.3 THE ROLE AND CHALLENGES OF EXPLAINABLE AI (XAI) IN HEALTHCARE.

In recent years, artificial intelligence (AI) has become increasingly significant in healthcare, particularly in diagnostics and treatment recommendations. However, a crucial challenge is enabling users to understand and trust AI model outcomes. To address this issue, Explainable Artificial Intelligence (XAI) has emerged as a key concept, providing in-depth explanations of AI model operations. This helps medical professionals understand the rationale behind AI decisions, enhancing transparency and fostering trust between AI systems and healthcare providers [17]. Additionally, XAI techniques contribute to improving decision-making processes, ensuring that AI systems operate effectively and are understandable [17].

Among the important XAI techniques are Local Interpretable Model-Agnostic Explanations (LIME) and Shapley Additive explanations (SHAP), which have been applied in various healthcare contexts. LIME is a tool that provides pixel-level explanations of model outcomes, which is particularly useful in medical imaging tasks such as breast cancer diagnosis. LIME allows physicians to visualize model operations and understand decision-making on a granular level [18]. SHAP, on the other hand, offers a robust framework for understanding the contributions of individual features to model predictions. SHAP provides clear attribution scores, enabling detailed analysis of model outcomes. However, its performance can be influenced by model choice and feature relationships [19].

Despite their significant benefits, XAI techniques like LIME and SHAP face several challenges. One key challenge is the need for systematic evaluation and improvement to ensure these methods are effective in diverse healthcare scenarios. Additionally, there is a need for developing mechanisms that can adapt to complex feature relationships, which remains a major limitation of current XAI applications [20]. Integrating XAI techniques with medical models requires systematic assessment to ensure reliability and practical applicability. Future research should focus on refining these methods to overcome existing limitations and enhance the capabilities of XAI in personalized medicine [21].

Incorporating XAI fulfills the essential requirement for trust and accountability in AI-driven decision-making. For healthcare professionals, the ability to interpret model outputs is critical for making confident, data-informed decisions. Explainable AI, particularly through LIME, offers the clarity needed to understand and rely on these predictions. With LIME, the model’s outputs move beyond "black-box" predictions to become transparent and interpretable, enabling stakeholders to grasp the factors driving each result. This transparency empowers them to verify, understand, and act effectively on the insights provided by the patient flow model. Consequently, explainability enhances the model's practical relevance in addressing hospital overcrowding challenges while encouraging AI's responsible and ethical use within healthcare. Ongoing research and development in this area will contribute to making AI systems more reliable and beneficial in the future.

CHAPTER III
METHODOLOGY

For this project, the CRISP-DM (Cross Industry Standard Process for Data Mining) framework will be used, which is a widely recognized standard for data analysis and developing Machine Learning models. This process consists of six main stages that help organize research activities systematically and efficiently manage large datasets.

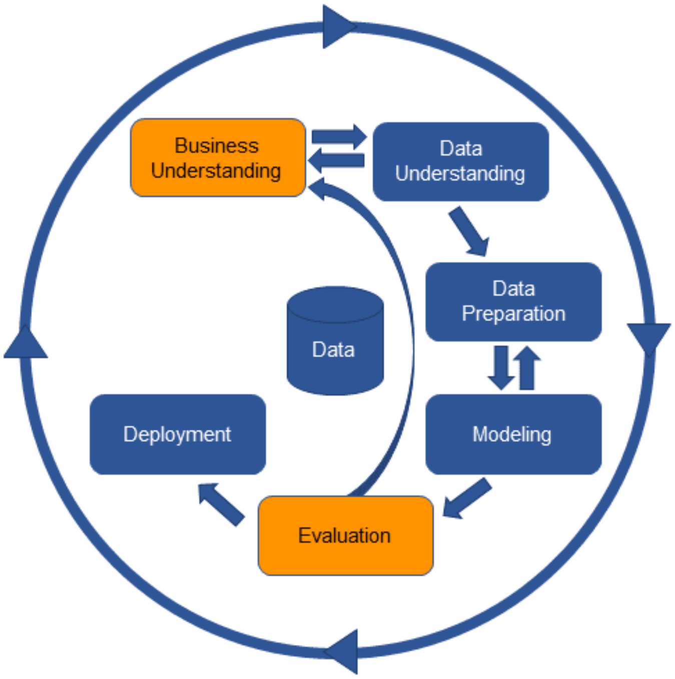


Figure 1: CRISP-DM diagram

CRISP-DM is flexible and can be adapted to different data and situations. In this project, CRISP-DM will guide the following steps:

3.1 BUSINESS UNDERSTANDING

In this project, NHS UK data in this project is utilized, structuring the business understanding phase to align insights from the dataset with the operational needs of hospitals. To support this alignment, an interview with a local Thailand hospital is conducted where study goals, such as predicting hospital stays, understanding readmission risks, and managing patient flow, in collaboration with the hospital staff are defined. Although the data is specific to the UK healthcare system, it provided a foundation for building predictive models that address universal hospital management challenges, applicable to Thailand hospitals as well. By framing the interview objectives around common issues like resource allocation and discharge planning, broader relevance and applicability across diverse healthcare contexts are achieved.

The interview also identified the needs of key stakeholders—hospital administrators, clinicians, and resource managers, as they are the primary beneficiaries of such predictive insights. Despite the use of NHS data, predictions related to patient flow and resource optimization may prove valuable in settings similar to the Thailand hospital referenced in the interview. Attributes such as demographics, diagnoses, and admission histories in the NHS dataset were emphasized, as these factors often impact patient flow. From the interview, potential data limitations, especially concerning differences between the NHS data structure and the intended deployment environment are anticipated.

Finally, concrete data science tasks from these insights to address hospital management priorities are outlined. Below are findings gain from the interview:

- Models from NHS data to predict long hospital stays or frequent readmissions, offering hospital staff actionable insights into patient throughput and bed occupancy.
- Model would be deployed, whether for real-time resource planning or for periodic reporting is important for strategic oversight.
- By incorporating explainable AI (e.g., LIME), model's predictions are interpretable, allowing healthcare providers to act confidently on its insights.

- By aligning NHS data with practical applications in hospital management, resource optimization and improved patient care are able to establish the model as a valuable decision-support tool.

3.2 DATA UNDERSTANDING

In this section, the study will provide a structured overview of the dataset by focusing on two critical components: Data Collection and Visualization. The Data Collection part will outline the sources, methods, and processes used to gather the dataset, ensuring a clear understanding of its origins, scope, and any potential biases or limitations. The Visualization part will then employ various graphical representations to illustrate the dataset's core attributes, trends, and distributions, offering an intuitive grasp of its structure and patterns. Together, these components will offer a comprehensive perspective on the dataset, enabling deeper insights into its characteristics and underlying features, which will serve as a foundation for subsequent analysis.

3.2.1 Data Collection

The data used for this project was obtained from NHS Digital, specifically from the Hospital Episode Statistics (HES) dataset for admitted patient care during the 2023-24 financial year. The dataset was downloaded directly from the official NHS website [22], which provides comprehensive information about hospital admissions, diagnoses, treatments, and patient demographics.

2.2.3 Data Description

In this project, we gather hospital admission data that encompasses various details about patient diagnoses, demographic information, and admission types. The primary focus is on a comprehensive range of diagnoses, classified according to the ICD-10 system [23]. ICD-10 organizes diseases and health conditions into multiple groups [24], covering a broad spectrum of medical conditions:

Code Range	Description
A00-B99	Certain infectious and parasitic diseases
C00-D49	Neoplasms
D50-D89	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
E00-E89	Endocrine, nutritional and metabolic diseases
F01-F99	Mental, Behavioral and Neurodevelopmental disorders
G00-G99	Diseases of the nervous system
H00-H59	Diseases of the eye and adnexa
H60-H95	Diseases of the ear and mastoid process
I00-I99	Diseases of the circulatory system
J00-J99	Diseases of the respiratory system
K00-K95	Diseases of the digestive system
L00-L99	Diseases of the skin and subcutaneous tissue
M00-M99	Diseases of the musculoskeletal system and connective tissue
N00-N99	Diseases of the genitourinary system
O00-O9A	Pregnancy, childbirth and the puerperium
P00-P96	Certain conditions originating in the perinatal period
Q00-Q99	Congenital malformations, deformations, and chromosomal abnormalities
R00-R99	Symptoms, signs, and abnormal clinical and laboratory findings, not elsewhere classified

S00-T88	Injury, poisoning, and certain other consequences of external causes
U04-U82	Codes for special purposes
Z00-Z99	Factors influencing health status and contact with health services

Table 2: Hospital Admission Data by ICD-10 Diagnosis Classification

3.2.4 About Dataset

The dataset consists of hospital admission statistics for various diagnoses, detailing the outcomes of finished admission episodes (FAEs) across different categories. represented by ICD-10 codes. Table 1 lists all the variables used in the analysis.

Variable	Description
Code	ICD-10 code representing the diagnosis
Diagnosis	Description of the diagnosed condition
Diagnosis Category	Category under which the diagnosis falls, as classified by ICD-10
Finished Admission Episodes (FAEs)	Total number of completed admissions for each diagnosis
Admission Types	Classification of admissions as Emergency, Waiting List, Planned, and Other
Mean Time Waited (Days)	Average time waited for treatment for each diagnosis
Mean Length of Stay (Days)	Average duration of stay in the hospital for each diagnosis
Age Distribution	Number of patients in various age categories

Table 3: Independent variables in the study

3.2.5 Visualization

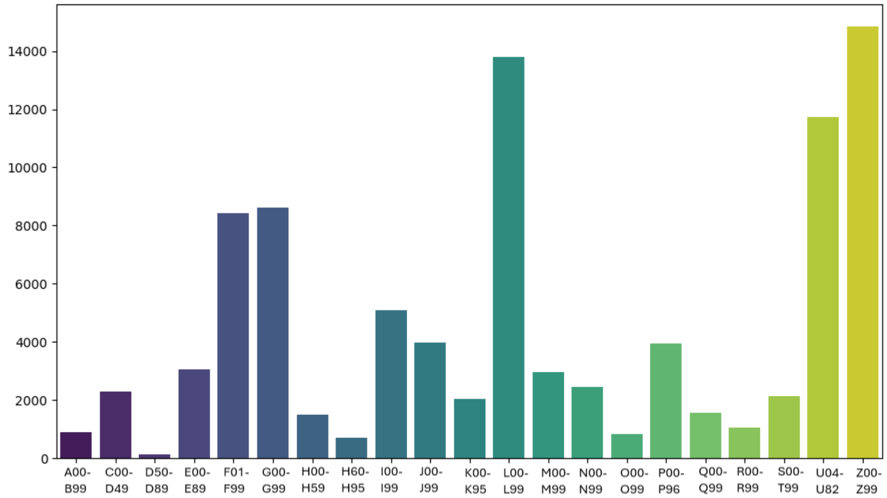


Figure 2: Shows the distribution of emergency admissions across different diagnosis codes.

Figure 2 shows the distribution of emergency admissions across different diagnosis codes. From the bar chart, the diagnosis codes Z00-Z99, K00-K95, and U04-U82 have the highest number of emergency admissions, respectively. This may indicate the severity and urgency of treatment required for diseases in these categories. On the other hand, the diagnosis codes D50-D89, H60-H95, O00-O9A, and R00-R99 show very low numbers of emergency admissions, which may reflect the ability to manage these conditions without the need for urgent care, such as planned treatments or effective symptom management.

This analysis highlights the importance of prioritizing resource allocation, such as medical personnel and equipment, for diagnosis categories with high emergency admission numbers to ensure that treatment demands can be met efficiently and promptly.

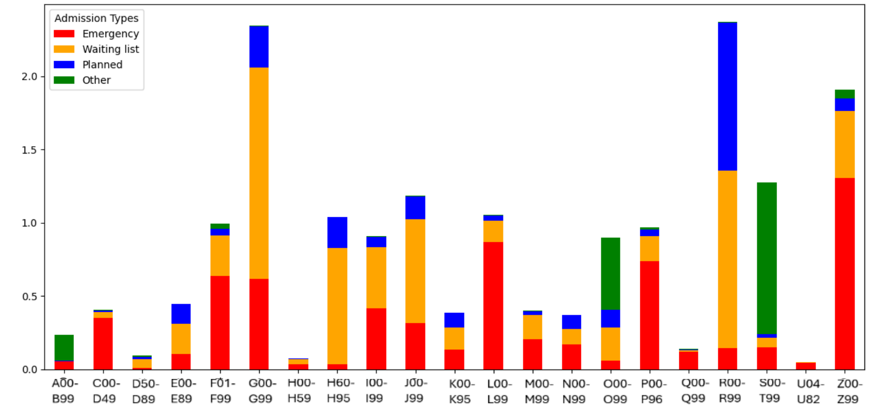


Figure 3: Distribution of Admission Types Across Different diagnosis codes

Figure 3 shows the distribution of admission types (Emergency, Waiting list, Planned, Other) across different diagnosis codes, providing a clear overview of how patient care is managed and distributed within each diagnosis group. Diagnosis codes G00-G99 shows a high number of patients requiring emergency care, with a notable proportion of emergency admissions and waiting list cases compared to planned or other types of care.

This suggests that the conditions in this category are severe and require urgent treatment. In contrast, diagnosis codes R00-R99 and Z00-Z99 display a high number of emergency admissions, yet also show significant numbers of planned admissions, indicating that care for these conditions may involve both urgent and planned treatments. Additionally, some diagnosis codes, such as A00-B99, C00-D49, E00-E89, I00-I99, and J00-J99, demonstrate low emergency admissions but higher numbers in planned or waiting list categories, suggesting that these conditions are more manageable in advance without the need for emergency care. The information from this chart can be used for planning and allocating medical resources to respond effectively and promptly to patient needs in each category.

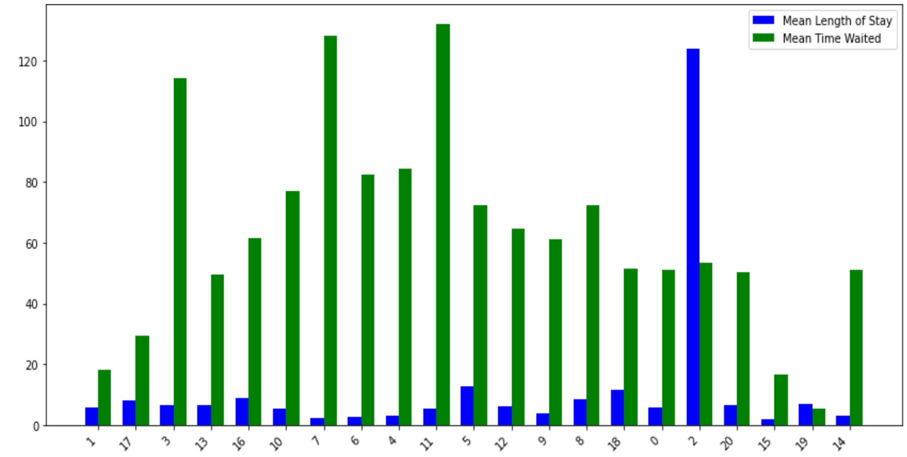


Figure 4: Mean Time Waited and Mean Length of Stay by Diagnosis Code

Figure 4 presents a graph the Mean Time Waited and Mean Length of Stay (LOS), categorized by different diagnosis codes. From the graph, it is evident that the Mean Time Waited (represented by the blue bars) varies significantly across different diagnosis codes, with some diagnosis codes showing a longer waiting time, such as those that require patients to wait over 90 days, while others have much shorter waiting times. Regarding the Mean Length of Stay (represented by the green bars), the length of stay for many diagnoses is relatively short, averaging no more than 10 days. However, there are some diagnosis codes that require longer stays, particularly those that involve special care needs. Despite the differences in the length of stay, waiting times are generally longer than the length of stay, highlighting challenges in bed management and resource allocation within hospitals. The data from this graph can be used to improve hospital bed management, particularly in predicting bed requirements and developing more efficient care plans.

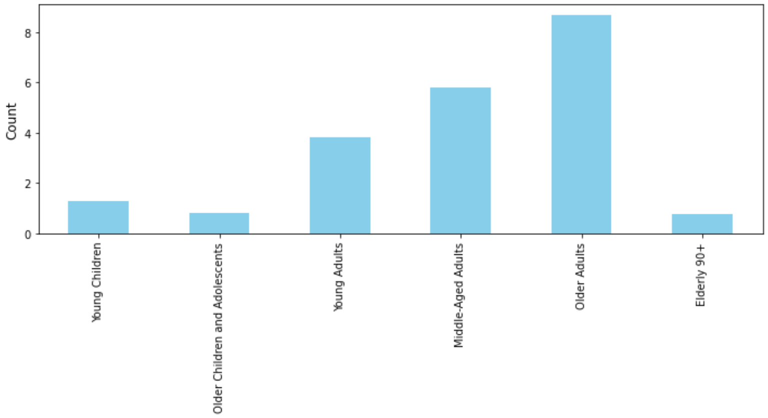


Figure 5: Distribution of Patients by Age Group

The bar chart illustrates the distribution of patients across various age groups, showing a clear variation in the number of patients within each category. On the left side of the chart, we observe age groups with relatively few patients, which are likely to represent younger populations, such as children or young adults, where the incidence of medical conditions requiring hospitalization is lower. As we move towards the right side, there is a noticeable increase in the number of patients in older age groups, peaking at a specific age range. This suggests a higher prevalence of health conditions or hospitalizations in those age categories, potentially due to chronic diseases or more frequent medical interventions required at these ages. The age group with the tallest bar represents the category with the highest number of patients, which may indicate that this age group has the greatest healthcare needs, possibly related to common health issues among middle-aged adults or elderly individuals. Overall, the graph highlights how hospitalization trends differ by age, with older age groups showing a higher demand for healthcare services.

3.3 DATA PREPARATION

In this project, the data was prepared and cleaned using a comprehensive set of steps to ensure its readiness for analysis.

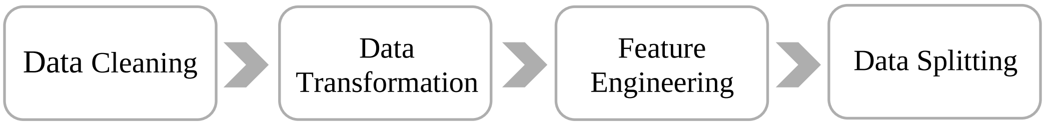


Figure 6: Data Preparation Process

3.3.1 Data Cleaning

- Handling Missing Data:

Use fillna(0) function**,** causing all missing values to be replaced with 0 so that no NaN is left in the data.

3.3.2 Data Transformation

- Label Encoding:

The *Diagnosis_categories* column, which is a categorical variable, was transformed into numeric values using LabelEncoder to ensure compatibility with machine learning models. This transformation assigns each category with a unique numeric value.

3.3.3 Feature Engineering

- Create new features:

‘Age Binning’

In this step, Age Binning was applied to consolidate highly detailed age categories into broader, more meaningful groups to simplify analysis and improve model performance. The original dataset included granular age ranges such as 'Age 0,' 'Age 1-4,' and so on.

These were grouped into six major categories:

- Young Children (0-4 years): Combined data from 'Age 0' and 'Age 1-4'
- Older Children and Adolescents (5-17 years): Combined data from 'Age 5-9' and 'Age 10-14'
- Young Adults (18-39 years): Spanned data from 'Age 18' to 'Age 35-39'

- Middle-Aged Adults (40-64 years): Included ages from 'Age 40-44' to 'Age 60-64'
- Older Adults (65-89 years): Grouped ages from 'Age 65-69' to 'Age 85-89'
- Elderly (90 years and above): Data from 'Age 90+'

This process reduced the number of age-related features in the dataset, simplifying it significantly. By grouping ages into broader categories, the analysis became more interpretable, and the risk of overfitting due to excessive granularity was minimized. Furthermore, it enhanced the ability to identify trends and actionable insights related to specific age groups, such as hospital overcrowding trends. The benefits include simplifying the data set, improving model interpretability, and facilitating clearer communication of results to stakeholders by presenting insights in a more digestible format.

Age Binning not only improved the efficiency of the data preparation process but also contributed to building more robust and interpretable models, making this an essential step in the overall methodology.

‘Emergency Admission Ratio’

We calculated the emergency admission ratio using the formula [25], [26]:

$$Emergency\ Admission\ Ratio\ (\%) = \left(\frac{Emergency}{Finished\ Admission\ Episodes}\right) \times 100$$

This feature helps to understand the level of emergency admissions for each type of patient and can be used to identify hospital overcrowding status. If the number of emergency admissions exceeds 85% of the total admissions, it is also identified as Overcrowding.

- Variable Selection:

Variable selection is a crucial step in the modeling process because it directly impacts on the accuracy and effectiveness of the developed model. In the context of this project, as mentioned in the data description above, we split the data into features (X) and target variables (y), with the following details:

i. Feature Matrix X: Independent Variables

ii. Target Variables (y): Dependent Variables

3.3.4 Data Splitting

In this part, we performed data splitting to create training and testing sets, which is an essential step for assessing the model's performance. By dividing the dataset, we can evaluate how well the model generalizes to unseen data, allowing us to reduce overfitting and improve the reliability of predictions. Typically, we allocate around 70% of the data for the Training Set and 30% for the Testing Set [27].

We utilized the *train_test_split* function from the *sklearn.model_selection* module, which offers advantages such as automatic random data splitting and customizable proportions. This approach helps ensure that both the training and testing sets share similar distributions of the target variable, enhancing the objectivity and reliability of model evaluation.

3.4 MODELING

In this project, we compared three machine learning models: K-Nearest Neighbors (KNN), Random Forest (RF), and Gradient Boosting (GB). The selection of these three models was based on their advantages and limitations, making them suitable for predicting hospital overcrowding, which involves complex and imbalanced data. This comparison helped us identify which model would be the most appropriate for making predictions based on such intricate datasets, considering factors like computational speed, accuracy, and the ability to handle imbalanced data.

The reasons for selecting these three models are as follows:

K-Nearest Neighbors (KNN):

KNN was chosen for its simplicity and suitability for less complex datasets. Although KNN does not require a complicated training process, it has limitations in terms of performance when working with large datasets or high-dimensional data [28]. Thus, KNN is ideal for smaller, less complex prediction tasks, helping us understand how a basic model works without requiring numerous parameters. It also serves as a benchmark for comparing more complex models.

Random Forest (RF):

Random Forest was selected for its ability to handle high-dimensional data and imbalanced classes effectively. It uses the bagging process, which helps reduce the risk of overfitting. This model tends to provide accurate results when dealing with complex data and uneven data distributions [29]. The use of Random Forest allows us to leverage ensemble learning, combining decisions from multiple trees to strengthen the model and reduce overall error.

Gradient Boosting (GB):

Gradient Boosting was chosen for its ability to capture non-linear relationships and improve model performance sequentially. This is particularly useful for complex, imbalanced datasets [28], [29]. Gradient Boosting often yields excellent results when predicting complex data with a lot of noise. By selecting Gradient Boosting, we aim to improve the accuracy of our predictions efficiently.

3.4.1 K-Nearest Neighbors (KNN)

The K-Nearest Neighbors (KNN) model is applied to predict hospital overcrowding by considering factors that may influence this condition. This model examines the nearest neighbors based on the specified number of neighbors (k value) to determine whether a new data point should be classified as "overcrowded" or "normal," relying on most neighboring points in either group. The distance between data points is used as the criterion for calculating proximity [30]. The distance between data points is typically calculated using the Euclidean distance formula:

$$d(p,q) = \sqrt{\sum_{i=1}^{n} \quad (p_{i} - q_{i})^{2}}$$

where:

- $d(p,q)$ is the distance between points p and q,
- n is the number of features,
- p_i and q_i are the feature values for points p and q.

In this project, we used **GridSearchCV** to find the optimal value of k the number of neighbors to consider, and selected the value that provided the best prediction results. GridSearchCV tests multiple values of k and evaluates performance using cross-validation to avoid overfitting. Once the optimal k was determined, the KNN model was used to predict hospital overcrowding in the test set.

3.4.2 Random Forest

The Random Forest (RF) model is used to predict hospital overcrowding by constructing a collection of decision trees during the training phase and aggregating their predictions to improve accuracy and robustness. As an ensemble learning method, Random Forest utilizes a combination of tree predictors to classify new instances, making it a highly effective model for this type of classification problem. The decision-making process in Random Forest relies on "bagging" (Bootstrap Aggregating), where multiple subsets of data are randomly sampled with replacement to train several decision trees. Each tree independently predicts whether the hospital will be "overcrowded" or "normal," and the final output is determined by majority voting among the trees.

For Model Tuning and Optimization, in this project, we applied the **RandomizedSearchCV** process to tune hyperparameters and identify the most optimal values, such as the number of trees *(n_estimators)*, maximum tree depth *(max_depth)*, and the number of features used at each node *(max_features)*. Tuning these hyperparameters helps improve model performance and efficiency

After tuning the parameters, the best model is trained on the training dataset and tested on the test dataset.

3.4.3 Gradient Boosting

Gradient Boosting is an ensemble learning technique that combines multiple weak models, typically decision trees, to create a stronger model through sequential training. In each step, a new model is trained to correct the errors (residual errors) made by the previous model. This makes the process additive, where each new model builds upon the errors of the previous one, leading to incremental improvements.

The model minimizes the loss function by training a new model to fit the residual errors of the previous model. Parameters of the model are adjusted based on the gradient of the loss function using gradient descent. The gradient indicates the direction in which the model should adjust its parameters to reduce the loss value.

During the initial training phase, parameters such as n_estimators (the number of trees), max_depth (the depth of the trees), and learning_rate (the learning rate) are defined to train the model on the training set.

Afterward, parameters are fine-tuned using **GridSearchCV** to find the best parameter values from a predefined grid. GridSearchCV optimizes the parameters through cross-validation, using performance statistics derived from the training data. Once the best parameters are identified, the model is trained and evaluated.

3.4.4 Evaluation

The performance of each model was evaluated using several classification metrics: Accuracy, Precision, Recall, and F1-Score. These metrics provide insights into the model’s ability to predict hospital overcrowding accurately and identify areas for improvement.

- Accuracy: Measures the overall correctness of the model, calculated as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- Precision: Indicates the proportion of true positive predictions among all positive predictions:

$$Precision = \frac{TP}{TP + FP}$$

- Recall: Measures of the model's ability to identify true overcrowding instances:

$$Recall = \frac{TP}{TP + FN}$$

- F1-Score: The harmonic means of Precision and Recall, providing a balance between both metrics:

$$F1\,Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

***** TP: True Positive, TN: True Negative, FN: False Negative, FP: False Positive***

These results will be further analyzed in Chapter 4, including tables and graphs that visualize the performance comparison across different models.

3.5 EXPLAINABLE AI (XAI) USING LIME

In this section, we implement Local Interpretable Model-agnostic Explanations (LIME) to interpret the predictions made by our machine learning models. The primary objective of using LIME is to explain why a model predicts a certain outcome, focusing on specific predictions rather than global behavior. This step is crucial for understanding how input features influence predictions, particularly when dealing with black-box models [31].

3.5.1 What is LIME?

LIME is a model-agnostic explainability technique that approximates the behavior of a complex model locally around a single prediction. It creates a simpler, interpretable model (e.g., a linear regression or decision tree) to explain the black-box model's decision for a specific instance. This is achieved by perturbing the input data and observing how the black-box model's predictions change [31].

3.5.2 Steps for Using LIME

The process of using LIME involves selecting the model to be explained and a specific instance from the test dataset to analyze. LIME perturbs the features of the selected instance to generate new data points and studies the relationship between features and predictions. A local surrogate model, such as Linear Regression, is then constructed to approximate the black-box model's behavior in the vicinity of the selected instance. This surrogate model assigns feature importance scores and generates explanations in text or visual formats, illustrating how individual features impact the prediction [31]. This approach enables users to understand the model’s reasoning and make decisions with enhanced confidence.

3.6 DEPLOYMENT

In this project, Streamlit was chosen as the platform for presenting the dashboard and making predictions. Streamlit is a tool well-suited for developing web applications that showcase data and machine learning model results quickly and easily.

3.6.1 Model Import

The model used in this project is Gradient Boosting Machine (GBM), which was trained and tested to predict hospital overcrowding based on various factors such as disease type, age, gender, and diagnosis. Once the model was trained, it was saved as a best_gbm_model.pkl file, which is imported into the application for making predictions based on user inputs.

3.6.2 Using LIME for Explanation

In this application, LIME (Local Interpretable Model-agnostic Explanations) is used to provide explanations for the model's predictions. LIME helps users understand which features have the most impact on the prediction by displaying results in an easy-to-understand format. Users can select a sample from the test dataset to view the model's explanations for each case.

3.6.3 What-If Functionality

The What-If tool allows users to input new data and see the prediction results under different scenarios. Users can input various feature values such as disease type, age, gender, and other information. The system then predicts the hospital overcrowding status and provides an explanation of the model's decision using LIME, showing how each feature influences the prediction.

3.6.4 Streamlit Integration

Streamlit was used to develop the web application because it is flexible and allows for quick visualization of results. Python is used to develop the model and the various functionalities, such as model loading, result display, and user input handling through Streamlit's interface. Streamlit's rapid development environment makes it efficient for prototyping and testing the application.

3.6.5 Deployment Steps

- Install Streamlit: Streamlit is installed via the command pip install streamlit.
- Model Import: The trained model is saved as a .pkl file and loaded into the application when the user accesses it.

- Developing the UI: The user interface is created using Streamlit, allowing users to input data, select options, and view prediction results from the model.

Using Streamlit to develop the dashboard enables quick display of model results and data. Additionally, LIME provides clear explanations of the model's outcomes, and the What-If tool allows users to experiment with different inputs to see prediction results. Developing and deploying this application enables users to easily use the prediction and explanation tools effectively and efficiently.

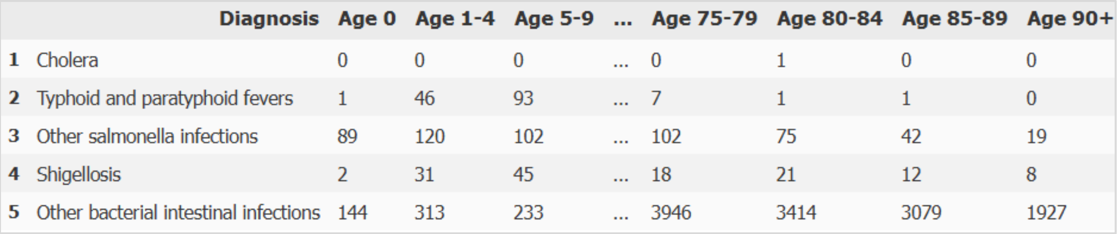
CHAPTER IV
PROPOSED SOLUTION AND RESULT

4.1 NEW FEATURE

4.1.1 Age Binning

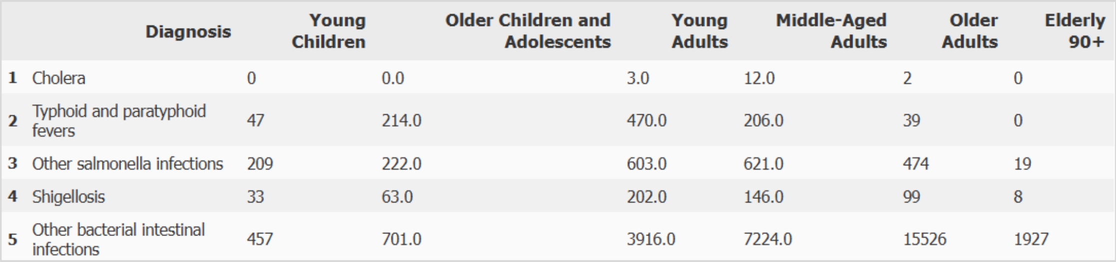
A new feature was created by applying Age Binning to group age categories into broader, more meaningful bins. The original dataset contained highly granular age categories, such as ‘Age 0,’ ‘Age 1-4,’ and so on, which were consolidated into six broader age groups. Young Children (0-4 years), Older Children and Adolescents (5-17 years), Young Adults (18-39 years), Middle-Aged Adults (40-64 years), Older Adults (65-89 years), and Elderly (90 years and above)

Before binning, the dataset included 24 columns related to age. After the binning process, these were reduced to 6 columns, significantly simplifying the dataset while preserving critical age-related information.



	Diagnosis	Age 0	Age 1-4	Age 5-9	...	Age 75-79	Age 80-84	Age 85-89	Age 90+
1	Cholera	0	0	0	...	0	1	0	0
2	Typhoid and paratyphoid fevers	1	46	93	...	7	1	1	0
3	Other salmonella infections	89	120	102	...	102	75	42	19
4	Shigellosis	2	31	45	...	18	21	12	8
5	Other bacterial intestinal infections	144	313	233	...	3946	3414	3079	1927

Figure 7: Example data before the binning process



	Diagnosis	Young Children	Older Children and Adolescents	Young Adults	Middle-Aged Adults	Older Adults	Elderly 90+
1	Cholera	0	0.0	3.0	12.0	2	0
2	Typhoid and paratyphoid fevers	47	214.0	470.0	206.0	39	0
3	Other salmonella infections	209	222.0	603.0	621.0	474	19
4	Shigellosis	33	63.0	202.0	146.0	99	8
5	Other bacterial intestinal infections	457	701.0	3916.0	7224.0	15526	1927

Figure 8: Example data after the binning process

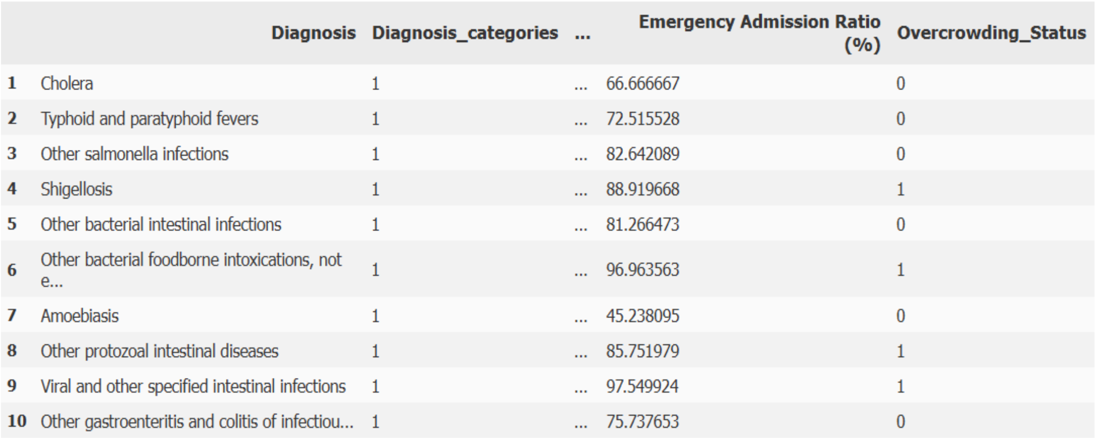
This process enhanced data interpretability and usability by simplifying age-related patterns, enabling more intuitive analysis and visualization. Grouping age categories into broader bins allowed for clearer identification of trends and actionable insights in downstream analysis, such as understanding hospital overcrowding trends among specific age groups.

4.1.2 Overcrowding status

A new ‘Overcrowding status’ feature has been added to monitor the proportion of emergency admissions to assess hospital overcrowding. This feature is created by having a threshold of 85% of the Emergency Admission Ratio.

The benchmark refers to hospitals operating at average occupancy levels above 85% that are likely to experience regular bed shortages and periodic crises that cause overcrowding. This significant threshold indicates that the hospital is operating near or at full capacity, leaving little room to accommodate unexpected surges in patient admissions, particularly emergencies [32].

The new feature was created to address overcrowding which was not mentioned in the original data. The new feature also aims to improve model explainability and interpretability as one of the objectives of this study is for model explainability for stakeholders to make informed data-driven decisions.



	Diagnosis	Diagnosis_categories	...	Emergency Admission Ratio (%)	Overcrowding_Status
1	Cholera	1	...	66.666667	0
2	Typhoid and paratyphoid fevers	1	...	72.515528	0
3	Other salmonella infections	1	...	82.642089	0
4	Shigellosis	1	...	88.919668	1
5	Other bacterial intestinal infections	1	...	81.266473	0
6	Other bacterial foodborne intoxications, not e...	1	...	96.963563	1
7	Amoebiasis	1	...	45.238095	0
8	Other protozoal intestinal diseases	1	...	85.751979	1
9	Viral and other specified intestinal infections	1	...	97.549924	1
10	Other gastroenteritis and colitis of infectiou...	1	...	75.737653	0

Figure 9: Sample Data Highlighting the New 'Overcrowding status' Feature.

4.2 COMPARISON BETWEEN KNN, RANDOM FORREST AND GRADIENT BOOSTING FOR MODEL ACCURACY

This project compares the performance of K-Nearest Neighbors (KNN) and Gradient Boosting models in predicting hospital overcrowding. Various metrics, including Accuracy, Precision, Recall and F1 Score, are used to evaluate each model's performance.

METRIC	K-NEAREST NEIGHBORS	RANDOM FOREST	GRADIENT BOOSTING
ACCURACY	0.8763	0.9629 0.9567	0.967
PRECISION	0.725	1.0	0.9444
RECALL	0.6042	0.8125 0.7812	0.8854
F1-SCORE	0.6591	0.8966 0.8772	0.914

Table 4: Model Performance Comparison: KNN, Random Forest and Gradient Boosting

The KNN model achieves an accuracy of 87.63%, but its Precision (72.50%) and Recall (60.42%) suggest a moderate capability to manage false positives and false negatives. The F1 Score of 65.91% reflects an acceptable balance between Precision and Recall, though it underperforms compared to other models.

The Random Forest model shows significant improvements, with an accuracy of 96.29%. Its Precision is perfect at 100%, indicating no false positives in the test data. However, Recall (81.25%) suggests some room for improvement in identifying all true positives. The F1 Score of 89.66% highlights the model’s strong balance between Precision and Recall, making it highly reliable in practice.

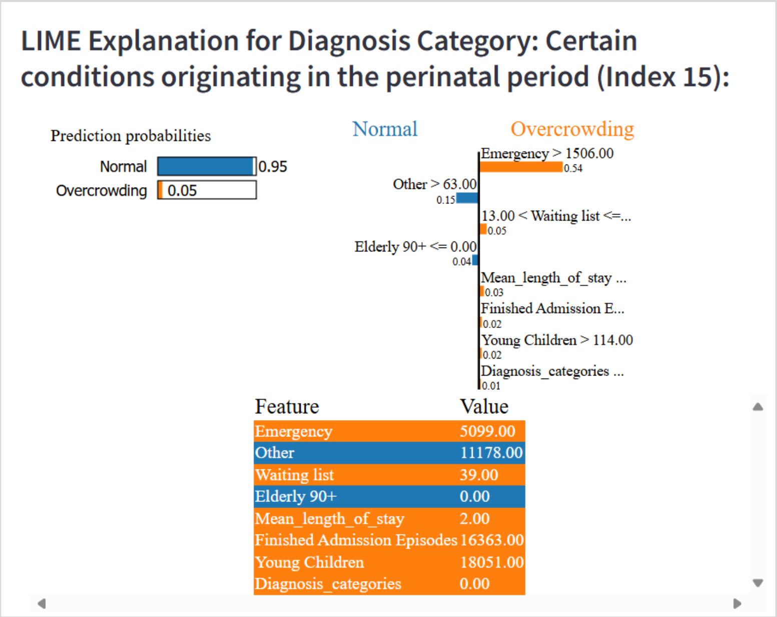
Gradient Boosting demonstrates the best overall performance, with an accuracy of 96.70%, Precision of 94.44%, and Recall of 88.54%, showing its robust ability to manage imbalanced data. Its F1 Score of 91.40% indicates the highest reliability and suitability for real-world applications, especially where errors carry significant consequences.

In summary, Gradient Boosting outperforms both KNN and Random Forest across all evaluation metrics, offering superior Precision and Recall, along with fewer false positives and negatives. The high F1 Score and model accuracy make Gradient Boosting the most suitable choice for this project. Moreover, its compatibility with Explainable AI (XAI) techniques like LIME ensures transparency and interpretability, enabling hospital administrators to understand the model's decision-making process effectively. This combination of high performance and explainability positions Gradient Boosting as the optimal model for hospital management systems focused on reducing overcrowding risks.

4.3 LIME FOR EXPLAINABLE AI

In this project, we incorporate LIME (Local Interpretable Model-agnostic Explanations) to enhance the transparency and interpretability of our Gradient Boosting Model (GBM). LIME helps explain individual predictions by approximating complex model behavior with simple, interpretable models for a specific instance. This enables us to identify key features influencing predictions and allows healthcare professionals to understand and act on model outputs effectively.

The primary objective of applying LIME is to gain insights into why certain hospital scenarios, such as "Normal" or "Overcrowding," are predicted based on features like admission types, patient demographics, and hospital resource utilization. Below are the interpretations of the LIME explanations for three different diagnosis categories:



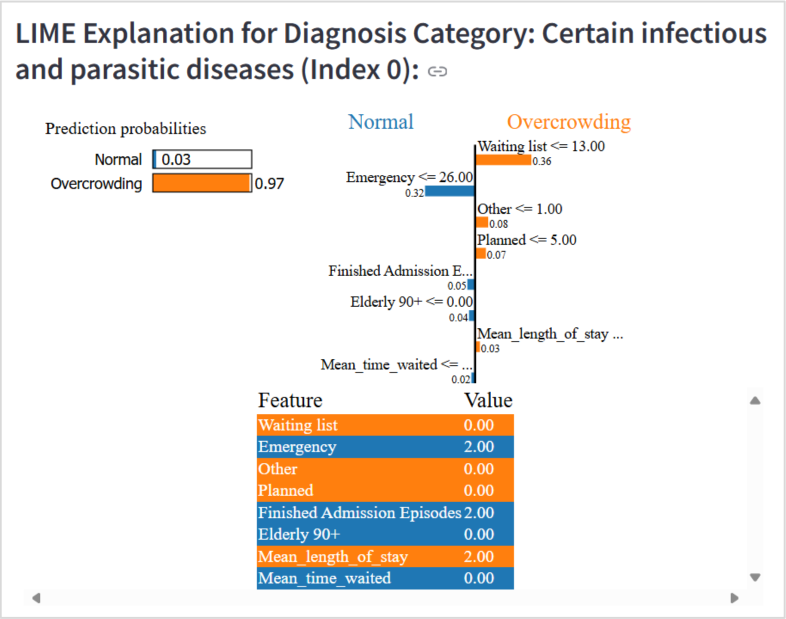
LIME Explanation for Diagnosis Category: Certain conditions originating in the perinatal period (Index 15):

Feature	Value
Emergency	5099.00
Other	11178.00
Waiting list	39.00
Elderly 90+	0.00
Mean_length_of_stay	2.00
Finished Admission Episodes	16363.00
Young Children	18051.00
Diagnosis_categories	0.00

Figure 10: LIME Explanation for Overcrowding Risk: Perinatal Conditions

4.3.1 Certain conditions originating in the perinatal period (Index 15):

- The model predicted a 95% probability for "Normal" and a 5% probability for "Overcrowding."
- Key contributors:
  - Emergency admissions (>1506): The most significant factor for overcrowding, contributing 0.54 to the prediction, indicating that a high number of emergency cases can increase overcrowding risk.
  - Other admissions (>63) and Elderly 90+ (0): These features contributed to the "Normal" prediction, suggesting that a balance in these categories helps maintain stability.
  - Other features such as Mean length of stay (≤2.0) had minor impacts on the overall prediction.
- Insight: Emergency admission volumes are critical to monitor, as they significantly affect overcrowding probabilities in this scenario.



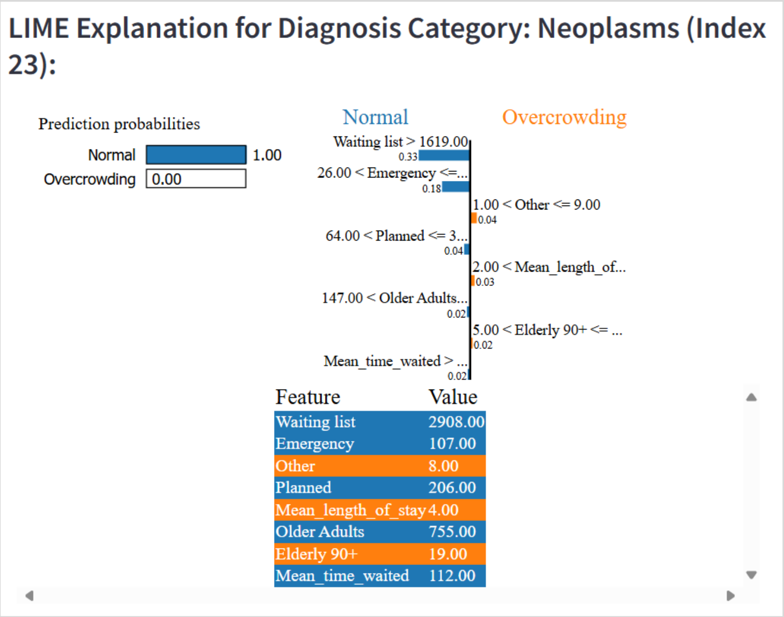
LIME Explanation for Diagnosis Category: Certain infectious and parasitic diseases (Index 0):

Feature	Value
Waiting list	0.00
Emergency	2.00
Other	0.00
Planned	0.00
Finished Admission Episodes	2.00
Elderly 90+	0.00
Mean_length_of_stay	2.00
Mean_time_waited	0.00

Figure 11: LIME Explanation for Overcrowding Risk: Infectious and Parasitic Diseases

4.3.2 Certain infectious and parasitic diseases (Index 0):

- The model predicted a 97% probability for "Overcrowding" and a 3% probability for "Normal."
- Key contributors:
  - Waiting list admissions (≤13): The most influential factor for overcrowding, contributing 0.36 to the prediction. This indicates that a backlog in the waiting list can quickly escalate overcrowding risks.
  - Emergency admissions (≤26): Contributed to the "Normal" condition, indicating that controlled emergency admissions support stability.
  - Planned admissions (≤5) and Mean length of stay (≤2.0) had smaller impacts but still leaned toward overcrowding.
- Insight: Managing waiting lists and minimizing patient delays can effectively reduce overcrowding risks for infectious disease cases.



LIME Explanation for Diagnosis Category: Neoplasms (Index 23):

Feature	Value
Waiting list	2908.00
Emergency	107.00
Other	8.00
Planned	206.00
Mean_length_of_stay	4.00
Older Adults	755.00
Elderly 90+	19.00
Mean_time_waited	112.00

Figure 12:LIME Explanation for Stable Conditions: Neoplasms

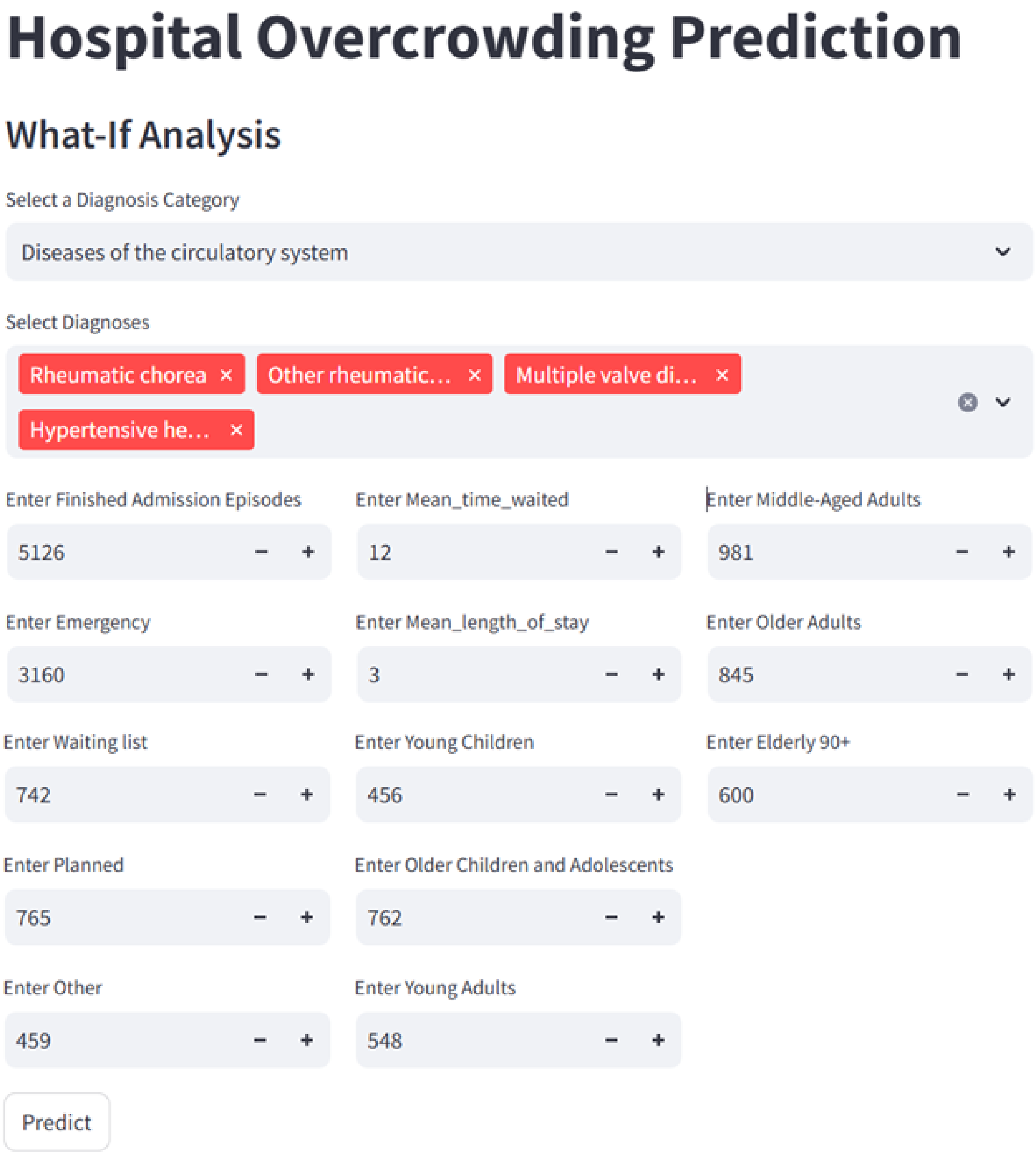
4.3.3 Neoplasms (Index 23):

- The model predicted a 100% probability for "Normal" and a 0% probability for "Overcrowding."
- Key contributors:
  - Waiting list (>1619): A strong factor supporting the "Normal" prediction, with a weight of 0.33. High waiting list numbers are not immediately indicative of overcrowding for neoplasm cases.
  - Emergency admissions (≤26): Also supported "Normal" conditions, showing that a low emergency admission count ensures stability.
  - Features like Planned admissions (≤64) and Older Adults admissions (≤147) had smaller positive impacts on the "Normal" prediction.
  - Other admissions (≤9) contributed minimally to the "Overcrowding" prediction.
- Insight: For neoplasm cases, high waiting list volumes and low emergency admissions maintain hospital capacity, allowing efficient operations.

Overall, LIME provides valuable insights into how each feature influences hospital status predictions. The results guide hospital administrators in optimizing resource allocation and mitigating overcrowding effectively and strategically.

4.4 DASHBOARD FOR HOSPITAL OVERCROWDING PREDICTION

The Hospital Overcrowding Prediction Dashboard was developed as an interactive web-based tool to facilitate predictions of hospital overcrowding and provide actionable insights. Built using Streamlit, the dashboard integrates machine learning models, explainable AI techniques, and user-friendly functionalities to enhance its usability.



Hospital Overcrowding Prediction

What-If Analysis

Select a Diagnosis Category: Diseases of the circulatory system

Select Diagnoses: Rheumatic chorea, Other rheumatic…, Multiple valve di…, Hypertensive he…

Enter Finished Admission Episodes: 5126

Enter Mean_time_waited: 12

Enter Middle-Aged Adults: 981

Enter Emergency: 3160

Enter Mean_length_of_stay: 3

Enter Older Adults: 845

Enter Waiting list: 742

Enter Young Children: 456

Enter Elderly 90+: 600

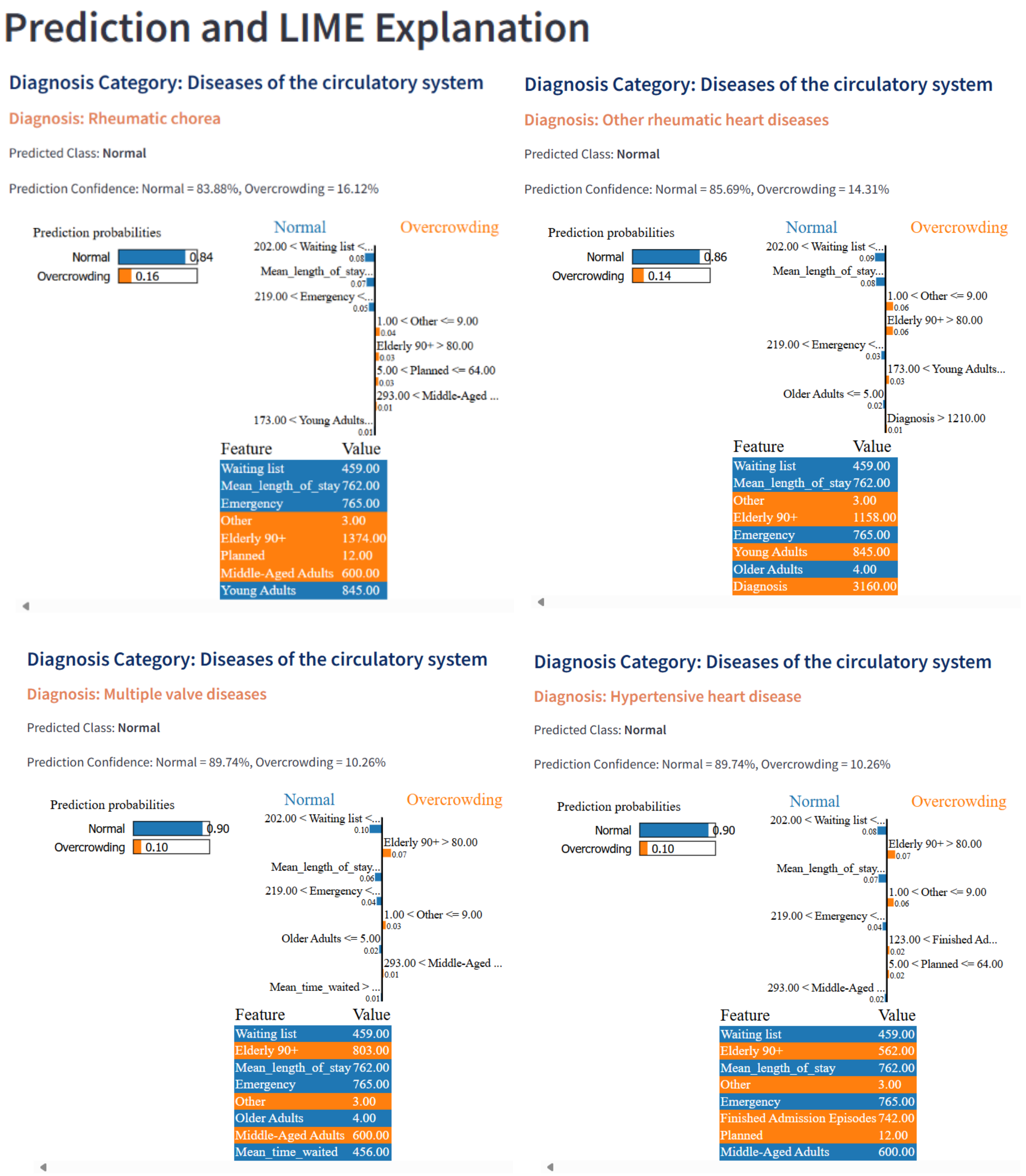
Enter Planned: 765

Enter Older Children and Adolescents: 762

Enter Other: 459

Enter Young Adults: 548

Figure 13: Hospital Overcrowding Prediction Dashboard - User Input Section



Prediction and LIME Explanation

Diagnosis Category: Diseases of the circulatory system

Diagnosis: Rheumatic chorea

Predicted Class: **Normal**

Prediction Confidence: Normal = 83.88%, Overcrowding = 16.12%

Feature	Value
Waiting list	459.00
Mean_length_of_stay	762.00
Emergency	765.00
Other	3.00
Elderly 90+	1374.00
Planned	12.00
Middle-Aged Adults	600.00
Young Adults	845.00

Diagnosis Category: Diseases of the circulatory system

Diagnosis: Other rheumatic heart diseases

Predicted Class: **Normal**

Prediction Confidence: Normal = 85.69%, Overcrowding = 14.31%

Feature	Value
Waiting list	459.00
Mean_length_of_stay	762.00
Other	3.00
Elderly 90+	1158.00
Emergency	765.00
Young Adults	845.00
Older Adults	4.00
Diagnosis	3160.00

Diagnosis Category: Diseases of the circulatory system

Diagnosis: Multiple valve diseases

Predicted Class: **Normal**

Prediction Confidence: Normal = 89.74%, Overcrowding = 10.26%

Feature	Value
Waiting list	459.00
Elderly 90+	803.00
Mean_length_of_stay	762.00
Emergency	765.00
Other	3.00
Older Adults	4.00
Middle-Aged Adults	600.00
Mean_time_waited	456.00

Diagnosis Category: Diseases of the circulatory system

Diagnosis: Hypertensive heart disease

Predicted Class: **Normal**

Prediction Confidence: Normal = 89.74%, Overcrowding = 10.26%

Feature	Value
Waiting list	459.00
Elderly 90+	562.00
Mean_length_of_stay	762.00
Other	3.00
Emergency	765.00
Finished Admission Episodes	742.00
Planned	12.00
Middle-Aged Adults	600.00

Figure 14: Hospital Overcrowding Prediction Results

Key Features of the Dashboard:

1. Prediction Functionality
   - The dashboard leverages a trained Gradient Boosting Machine (GBM) model to predict hospital overcrowding risk based on inputs like diagnosis category, patient demographics (age groups, etc.), and admission details (emergency status, waiting lists, length of stay). Once the user provides this data, the model outputs a prediction indicating whether the hospital is at risk of overcrowding, along with the confidence level of the prediction.

2. What-If Analysis
   - The "What-If" feature allows users to experiment with different scenarios by adjusting key input parameters, such as waiting list sizes, emergency admissions, and average length of stay. The dashboard dynamically updates the overcrowding predictions based on these changes, giving users the ability to test how variations in hospital conditions can impact overcrowding risks and plan accordingly.

3. Explainable AI with LIME
   - To provide transparency, the dashboard integrates Local Interpretable Model-Agnostic Explanations (LIME). For every prediction, LIME explains the most influential features that contributed to the model's decision. This is displayed through visual bar charts, helping users understand why certain factors like emergency admissions or patient demographics have the most impact on overcrowding risk, thereby fostering more informed decision-making.

4. Streamlined Interface
   - The dashboard is designed with a user-friendly interface, making it easy to interact with. It includes dropdown menus for diagnosis category selection, sliders for adjusting parameters such as waiting times, and

buttons to generate predictions. Results are displayed clearly, with both the overcrowding risk prediction and the associated LIME explanation provided in an intuitive, visually engaging format that enhances the user experience.

CHAPTER V
CONCLUSION AND DISCUSSION

This project tackled the critical issue of hospital overcrowding by developing a predictive patient flow model using machine learning and Explainable AI (XAI). The study demonstrated how models like Gradient Boosting, Random Forest, and K-Nearest Neighbors (KNN) can predict overcrowding scenarios, with Gradient Boosting proving to be the most effective in terms of accuracy and balanced performance. The integration of LIME provided transparency to the model’s decision-making, helping stakeholders interpret predictions and make informed decisions. This transparency is crucial for building trust among healthcare professionals and administrators.

The predictive model aids hospital administrators in identifying overcrowding patterns and planning resource allocation. The use of feature engineering, such as age binning and overcrowding status, improved data interpretability and model performance. The interactive dashboard developed offers a practical tool that combines predictive insights with explainable outputs, supporting data-driven decision-making.

However, the study faced limitations, including reliance on NHS data, which may not apply universally to other healthcare systems or regions. While LIME provided useful local explanations, global interpretability remains an area for improvement.

Future work could involve incorporating data from different healthcare systems for broader generalizability, exploring additional XAI techniques like SHAP, and enhancing the dashboard with real-time data for dynamic decision-making. In conclusion, this project shows that machine learning and Explainable AI can significantly improve hospital overcrowding management by providing actionable insights, fostering transparency, and enhancing operational efficiency and patient care quality.

REFERENCES

[1] W. M. W. M. A. และ W. R. I. , “Improving emergency department overcrowding in Malaysian government hospital.,” *Journal of Quality Measurement and Analysis,* เล่มที่ 17, %11, pp. 19-39, 2021.

[2] G. Savioli, I. F. Ceresa, N. Gri, G. B. Piccini, Y. Longhitano, C. Zanza, A. Piccioni, C. Esposito, G. Ricevuti และ M. A. Bressan, “Emergency Department Overcrowding: Understanding the Factors to Find Corresponding Solutions,” *Journal of Personalized Medicine,* เล่มที่ 12, %12, p. 279, 14 Feb 2022.

[3] J. G. Guerrero, A. S. Alqarni, R. P. Cordero, I. Aljarrah และ M. A. Almahaid, “Perceived Causes and Effects of Overcrowding Among Nurses in the Emergency Departments of Tertiary Hospitals: A Multicenter Study,” *Risk Management and Healthcare Policy,* เล่มที่ 17, pp. 973-982, 20 Apr 2024.

[4] M. Oberlin, E. Andrès, M. Behr, S. Kepka, P. Le Borgne และ P. Bilbault, “Emergency overcrowding and hospital organization: Causes and solutions,” *La Revue de Médecine Interne,* เล่มที่ 41, %110, October 2020.

[5] CADTH, “Artificial Intelligence for Patient Flow,” *Canadian Journal of Health Technologies,* เล่มที่ 4, %15, 2024.

[6] R. Woods, R. Sandoval, G. Vermillion, B. Bates-Jackson, A. Nwankwo, C. P. Canamar และ L. Sarff, “The Discharge Lounge: A Patient Flow Process Solution,” *Journal of Nursing Care Quality,* เล่มที่ 35, %13, pp. 240-244, Jul/Sep 2020.

[7] M. I. Abubakar, “Stochastic Simulation Modelling and Analysis of Out-Patient flow: A Case Study of General Hospital Hadejia, North-western Nigeria,” *FUOYE Journal of Engineering and Technology,* เล่มที่ 7, %14, November 2022.

[8] V. Domova และ S. Sander-Tavallaey, “Visualization for Quality Healthcare: Patient Flow Exploration,” ใน *2019 IEEE International Conference on Big Data (Big Data)*, Los Angeles, CA, USA, 2019.

[9] M. Islam และ S. Jin, “An Overview of Data Visualization,” ใน *2019 International Conference on Information Science and Communications Technologies (ICISCT)*, Tashkent, Uzbekistan, 2019.

[10] L. Addepalli, L. G. S. S. S. Hussain, M. J. V. Uppalapati, W. Ali และ V. S. , “Assessing the Performance of Python Data Visualization Libraries: A Review,” *International Journal of Computer Engineering in Research Trends,* เล่มที่ 10, %11, pp. 28-39, January 2023.

[11] B. A. Kassie, A. B. Senay และ G. S. Tegenaw, “Developing a patient flow visualization and prediction model using aggregated data for a healthcare network cluster in Southwest Ethiopia,” *PLOS Digital Health*, เล่มที่ 3, %12, 7 February 2024.

[12] A. Gupta, K. B. Singh, K. Singh, P. K. Kushwaha, B. P. Lohani และ S. Kumar, “Unveiling Insights: Exploring Healthcare Data through Data Analysis”.

[13] G. R. S. J., K. Jothikumar และ P. Priyadharshini, “Advancing Healthcare Through Data Science Techniques for Comprehensive Analysis and Visualization of Healthcare Data,” ใน *Cybersecurity and Data Management Innovations for Revolutionizing Healthcare*, IGI Global, 2024, pp. 1-15.

[14] G. Battineni, M. Mittal และ S. Jain, “Advanced Prognostic Predictive Modelling in Healthcare Data Analytics,” ใน *Data Visualization in the Transformation of Healthcare Industries*, เล่มที่ Lecture Notes on Data Engineering and Communications Technologies, 2021, pp. 1-23.

[15] D. Asuquo, I. Umoren, F. Osang และ K. Attai, “A Machine Learning Framework for Length of Stay Minimization in Healthcare Emergency Department,” *Studies in Engineering and Technology,* เล่มที่ 10, %11, 2023.

[16] M. Tello, E. Reich, J. Puckey, R. Maff, A. García Arce, B. Bhattacharya และ F. Feijoo, “Machine learning-based forecast for the prediction of inpatient bed demand,” *BMC Medical Informatics and Decision Making,* เล่มที่ 22, %11, 2022.

[17] S. Singhal, A. Sharma, A. Singh และ A. Pandey, “Industry-specific AI and ML applications in healthcare focus on improving diagnostics, patient care, and medical research,” ใน *Advances in Systems Analysis, Software Engineering, and High-Performance Computing*, IGI Global, 2024, p. 110–124.

[18] A. Ponce, J. López-Bautista และ R. Fernandez-Beltran, “Interpretando Modelos de IA en Cáncer de Mama con SHAP y LIME,” *Ideas en Ciencias de la Ingeniería,* เล่มที่ 2, %12, pp. 15-15, 2024.

[19] A. Salih และ e. al., “A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME,” *Advanced Intelligent Systems,* 2024.

[20] G. Huang และ e. al., “From Explainable to Interpretable Deep Learning for Natural Language Processing in Healthcare: How Far from Reality?,” *Computational and Structural Biotechnology Journal,* เล่มที่ 24, pp. 362-373, 2024.

[21] V. R. R. Kovvuri, “Explainable Artificial Intelligence across Domains: Refinement of SHAP and Practical Applications,” 2024.

[22] D. NHS, “Hospital admitted patient care activity, 2023-24,” 2023. [ออนไลน์]. Available: https://digital.nhs.uk/data-and-information/publications/statistical/hospital-admitted-patient-care-activity/2023-24. [%1 ที่เข้าถึง11 Nov. 2024].

[23] J. Hirsch, G. Nicola, G. McGinty, R. Liu, R. Barr, M. Chittle และ L. Manchikanti, “ICD-10: history and context,” *American Journal of Neuroradiology,* เล่มที่ 34, %14, pp. 596-599, 2016.

[24] “International Classification of Diseases 10th Revision (ICD-10),” 2019. [ออนไลน์]. Available: https://icd.who.int/browse10/2019/en.

[25] J. Hippisley-Cox และ C. Coupland, “Predicting risk of emergency admission to hospital using primary care data: derivation and validation of QAdmissions score,” *BMJ Open,* เล่มที่ 3, %18, 2013.

[26] N. Kraaijvanger, D. Rijpsma, L. Roovers, H. van Leeuwen, K. A. H. Kaasjager, L. van den Brand, L. Horstink และ M. J. R. Edwards, “Development and validation of an admission prediction tool for emergency departments in the Netherlands,” *Emergency Medicine Journal*, เล่มที่ 35, %18, pp. 464-470, 2018.

[27] A. A. Tokuç, “Splitting a Dataset into Train and Test Sets,” Tarnum Java SRL, 12 May 2023. [ออนไลน์]. Available: https://www.baeldung.com/cs/train-test-datasets-ratio. [%1 ที่เข้าถึง11 November 2023].

[28] D. N. Cosenza, L. Korhonen, M. Maltamo, P. Packalen, J. L. Strunk, E. Næsset, T. Gobakken, P. Soares และ M. Tomé, “Comparison of linear regression, k-nearest neighbour and random forest methods in airborne laser-scanning-based prediction of growing stock,” *Forestry: An International Journal of Forest Research,* เล่มที่ 94, %12, pp. 311-323, April 2021.

[29] L. Dube และ T. Verster, “Enhancing classification performance in imbalanced datasets: A comparative analysis of machine learning models,” *AIMS Data Science,* 13 Oct. 2023.

[30] R. Duda, P. Hart และ D. Stork, Pattern classification, John Wiley & Sons, 2001.

[31] J. Dieber และ S. Kirrane, “Why model why? Assessing the strengths and limitations of LIME,” 2020.

[32] C. Barchielli, M. Vainieri, C. Seghieri, E. Salutini และ P. Zoppi, “The Function of Bed Management in Pandemic Times—A Case Study of Reaction Time and Bed Reconversion,” *International Journal of Environmental Research and Public Health,* เล่มที่ 20, %112, p. 6179, 2023.