**Patient Flow Model for Hospital Admission Analysis**

**IKMEE U-DAIDEE**

**642437002**

**FATONI UNIVERSITY**

**1445/2024**

**Chapter 1: Introduction**

* 1. **Introduction**

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 the 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 community.

This project proposes k-Nearest Neighbor (KNN) 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. **Problem statement**

1. 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.
2. 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.
3. 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. **Objective**
4. Employ data visualization techniques to explore and extract key features from the dataset, such as trends in hospital admissions, diagnosis patterns, and demographic impacts, which are relevant to predicting hospital overcrowding.
5. 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.
6. Explore how different primary diagnoses, patient demographics (age, gender), and types of admissions (emergency, elective) contribute to overcrowding in hospitals.
   1. **Scope of study**

* Module to be developed:
* Data use: Admitted Patient Care activity in England for the financial year 2022-23.
* Tools: Implement data visualization, prediction models, and Explainer Dashboard using Jupyter Notebook.
* Target User:
* The target users of this study are hospital executives 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. **Project Significance**

The significance of this project is that it can help solve important problems in hospital management, especially regarding overcrowding and patient movement. The project aims to improve how hospitals operate by using data to make better decisions, which will lead to better patient care and more efficient use of resources.

**Chapter 2: 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 overallpatient 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 "MachineLearning BasedForecast 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].

XAI holds great potential for advancing and improving healthcare. By incorporating explanatory techniques such as LIME and SHAP, medical decision-making can become more transparent, fostering a better understanding of AI model outcomes and enhancing overall treatment efficacy. Ongoing research and development in this area will contribute to making AI systems more reliable and beneficial in the future.

**Chapter 3: Project Methodology**

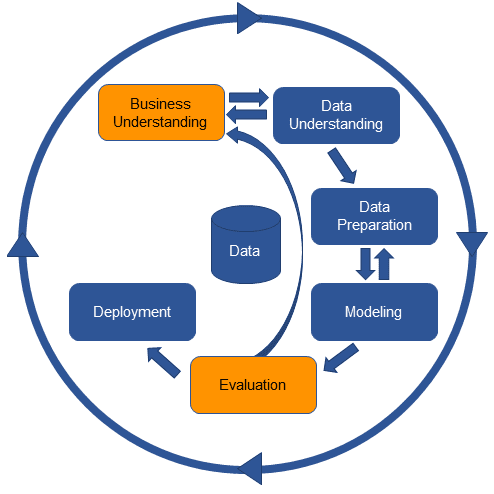
 For this project, the CRISP-DM (Cross Industry Standard Process for Data Mining) framework will be used, which is a widely recognized standard for data analysis and developing Machine Learning models. This process consists of six main stages that help organize research activities systematically and efficiently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

* 1. **Business Understanding:**

* 1. **Data Understanding:**

In this part, the study will focus on providing an overview of the data, divided into two main parts: Data Collection and Visualization. That helps to get a comprehensive understanding of the dataset and its key features.

* + 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, which provides comprehensive information about hospital admissions, diagnoses, treatments, and patient demographics.

Describe data

In this project, we collected 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. ICD-10 organizes diseases and health conditions into multiple groups [22], 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 1: Hospital Admission Data by ICD-10 Diagnosis Classification

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 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 2: Independent variables in the study

* + 1. Visualization

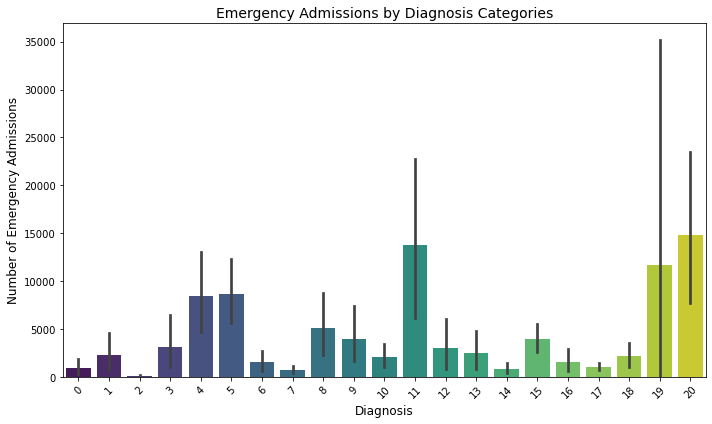


Figure 2: Number of Patients by Age Group and Diagnosis Category

In the *Figure 2*, displays the frequency of different types of neoplasms (Benign neoplasm, In situ neoplasm, Malignant neoplasm, and Neoplasms of uncertain or unknown behavior) across various age groups. We can see that Malignant neoplasms have the highest occurrence, with a sharp increase in the 55-74 age group, followed by a decline in the 75-90+ age group. Benign neoplasms show a gradual increase with age, peaking in the 55-74 group. In situ neoplasms and Neoplasms of uncertain or unknown behavior remain relatively low and stable across all age groups. This suggests that the likelihood of encountering malignant neoplasms increases significantly with age, especially in the middle-aged to senior groups.

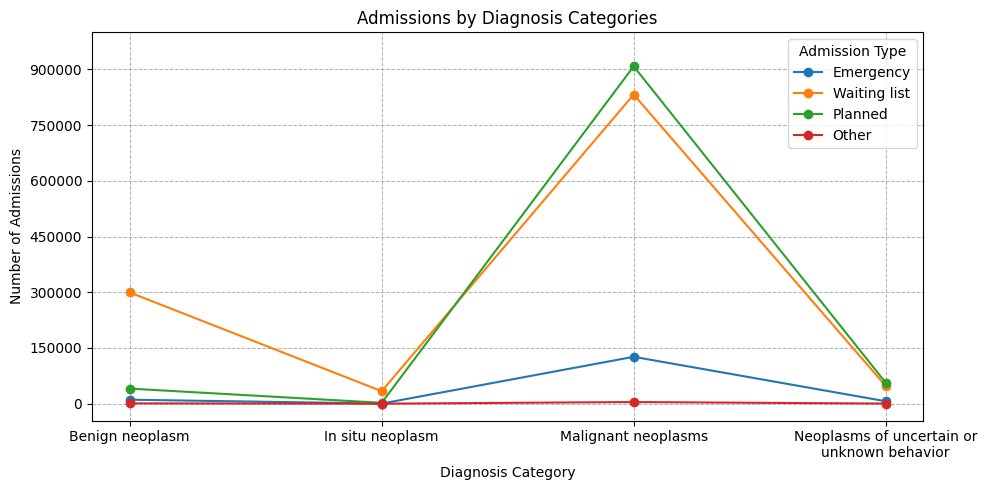


Figure 3: Number of Patients by admission types and Diagnosis Category

Forthe *Figure 3* shows the distribution of admission types (Emergency, Waiting list, Planned, and Other) for different diagnosis categories of neoplasms. We can see that Malignant neoplasms have the highest number of planned admissions, followed by the waiting list and emergency admissions. Benign neoplasms have a higher number of cases in the waiting list category compared to others. Neoplasms of uncertain or unknown behavior and In situ neoplasms show relatively low numbers across all admission types. This suggests that malignant neoplasms are primarily handled through planned admissions, highlighting their predictable treatment pattern compared to other types.

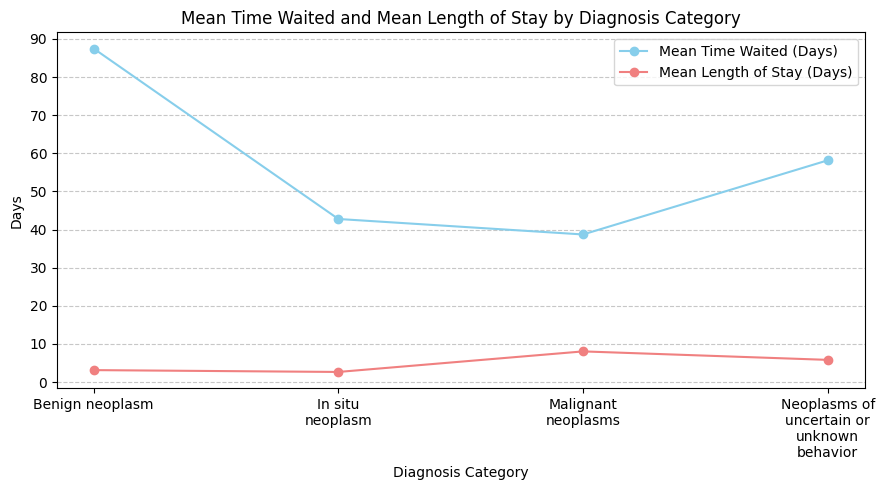


Figure 4: Mean Time Waited and Mean Length of Stay by Diagnosis Category

And the *Figure 4* illustrating the Mean Time Waited and Mean Length of Stay (LOS) categorized by different types of neoplasm diagnoses.

For the Mean Time Waited, it can be observed that patients with Benign neoplasms have the highest average waiting time, around 90 days. The waiting time steadily decreases for In situ neoplasms, dropping to about 40 days. For Malignant neoplasms, the average waiting time is around 40 days, with a slight increase for Neoplasms of uncertain or unknown behaviour, reaching approximately 50 days. This figure indicates that patients with benign neoplasms generally have the longest waiting times, while those with other types of neoplasms experience shorter waiting periods.

For the Mean LOS, it is found that the mean LOS for all types of neoplasms is relatively low compared to the mean waiting time, with averages ranging between 5 to 10 days. The line graph remains relatively steady, with a slight increase for Malignant neoplasms and Neoplasms of uncertain or unknown behaviour, but still staying below 10 days.

* 1. **Data Preparation**

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

Data Cleaning

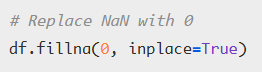
Data Transformation

Feature Engineering

Data Splitting

Figure 5: Data Preparation Process

* + 1. Data Cleaning

Figure 6: Handling Missing Value

* 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.

* + 1. Data Transformation

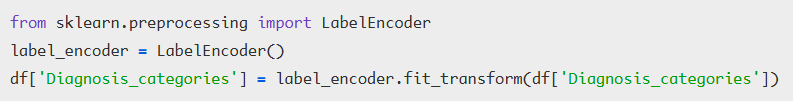


Figure 7: Label Encoding

* Label Encoding:

Use *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.

* + 1. Feature Engineering
* Create new features:

*‘Emergency Admission Ratio’*: We calculated the emergency admission ratio using the formula:

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:

1. Feature Matrix X: Independent Variables
2. Target Variables (y): Dependent Variables
   * 1. 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 80% of the data for the Training Set and 20% for the Testing Set [23].

Figure 8: Data Splitting

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.

* 1. **Modeling**

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 [24]. The distance between data points is typically calculated using the Euclidean distance formula:

where:

* is the distance between points and,
* is the number of features,
* ​ and ​ are the feature values for points and.

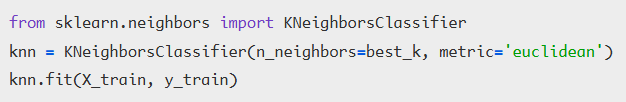


Figure 9: Example of Building a KNN model

* Target Variable: Overcrowding\_Status

This is the target variable we aim to predict, representing whether the hospital will experience overcrowding.

* Predictor Variables:

This refers to the independent variables or features used to predict the target variable.

The appropriate k value is determined using cross-validation to find the optimal setting that yields the most accurate predictions. Once tuned, the model is applied to predict overcrowding in the test dataset.

After tuning, the KNN model is used to predict overcrowding in the test set. Model performance is evaluated through metrics like Accuracy, Precision, Recall and F1 Score. These evaluation results are presented in tables and graphs within the report.

แน่นอนค่ะ! นี่คือลักษณะคำพูดอธิบายที่สามารถใช้ในการนำเสนอสไลด์ **Training Process** สำหรับทั้งสองโมเดล:

**Training Process: K-Nearest Neighbors (KNN)**

"สำหรับโมเดล KNN เราจะเริ่มต้นด้วยการเตรียมข้อมูล โดยเลือกตัวแปรที่ใช้ทำนายและตัวแปรเป้าหมายที่ต้องการทำนาย จากนั้น เราจะใช้ **GridSearchCV** เพื่อค้นหาค่าของ **k** ที่ดีที่สุด โดยการทดสอบค่าต่างๆ ของ **k** และเลือกค่าที่ให้ผลลัพธ์ดีที่สุดผ่านการ cross-validation หลังจากนั้น เราจะคำนวณระยะห่างระหว่างข้อมูลใหม่กับข้อมูลในชุดฝึกอบรม โดยใช้ **Euclidean Distance** เพื่อหาค่าที่ใกล้เคียงที่สุด จากนั้น เราจะทำการจัดกลุ่มข้อมูลใหม่ว่าเป็น “แออัด” หรือ “ปกติ” ตาม **k** เพื่อนบ้านที่ใกล้ที่สุด หลังจากโมเดลเสร็จสิ้นการฝึกแล้ว เราจะประเมินผลการทำงานของโมเดลด้วย **Accuracy**, **Precision**, **Recall**, และ **F1 Score**"

**Training Process: Gradient Boosting**

"สำหรับโมเดล Gradient Boosting เริ่มต้นด้วยการเตรียมข้อมูลเหมือนกับ KNN โดยเลือกตัวแปรที่ใช้ทำนายและตัวแปรเป้าหมาย หลังจากนั้น เราจะตั้งค่าพารามิเตอร์เริ่มต้น เช่น **n\_estimators**, **max\_depth**, และ **learning\_rate** เพื่อฝึกโมเดล โดยโมเดลนี้จะสร้างชุดของต้นไม้ตัดสินใจ ซึ่งแต่ละต้นไม้จะเรียนรู้จากข้อผิดพลาดของต้นไม้ก่อนหน้า และค่อยๆ ปรับปรุงความแม่นยำ การปรับพารามิเตอร์จะทำโดยใช้ **Gradient Descent** ซึ่งจะช่วยลดค่า **loss function** และปรับปรุงโมเดลให้ดียิ่งขึ้น หลังจากนั้น เราจะใช้ **GridSearchCV** เพื่อค้นหาค่าพารามิเตอร์ที่ดีที่สุดผ่าน cross-validation เมื่อโมเดลถูกฝึกเสร็จ เราจะประเมินผลด้วย **Accuracy**, **Precision**, **Recall**, และ **F1 Score**"

คำอธิบายนี้จะช่วยให้ผู้ฟังเข้าใจถึงกระบวนการฝึกโมเดลทั้งสองได้ชัดเจนและเข้าใจง่ายค่ะ

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.

รูปภาพประกอบด้วย ข้อความ, ภาพหน้าจอ, ตัวอักษร, ไลน์

คำอธิบายที่สร้างโดยอัตโนมัติ

*Figure 10: Code for Training the Gradient Boosting Model*

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. The model's performance is assessed using various metrics such as Accuracy, Precision, Recall, and F1-Score, the same as KNN model.

* 1. **Evaluation:**

The evaluation of the model's performance in predicting hospital overcrowding was conducted using various metrics, including Accuracy, Precision, Recall, and F1 Score, which are appropriate standards for evaluating classification models. The details of each metric are as follows:

* Accuracy: This measures the overall correctness of the model by showing the proportion of all correct predictions compared to the total test data. The formula for calculating Accuracy is:
* Precision: This metric evaluates the accuracy of predicting overcrowding by showing the proportion of true positive predictions (TP) compared to the total number of positive predictions. The formula for calculating Precision is:
* Recall: This measures the model's ability to identify actual overcrowded cases (TP) from all instances of overcrowding. The formula for calculating Recall is:
* F1 Score: This represents the harmonic mean of Precision and Recall, which assesses the balance between identifying actual overcrowded cases and minimizing false predictions. The formula for calculating F1 Score is:

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

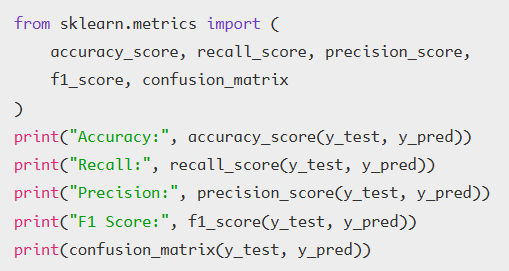


Figure 10: Example of Model Evaluation

This block of code calculates the Accuracy, Precision, Recall, and F1 Score by comparing the predictions with the actual data in the test dataset. These results help assess how well the model predicts hospital overcrowding and identify areas that may need improvement. Further details regarding the performance metrics will be presented in tables and graphs in the next chapter (Chapter 4).

* 1. **Deployment:** ขั้นตอนสุดท้ายของ CRISP-DM จะเป็นการนำโมเดลไปใช้จริง โดยในโปรเจคนี้จะมีการพัฒนา Dashboard เพื่อแสดงผลข้อมูลและการทำนายของโมเดลในรูปแบบที่เข้าใจง่ายและใช้ได้จริง โดยจะใช้เครื่องมือในการสร้างแดชบอร์ด เช่น ExplainerDashboard ที่เชื่อมต่อกับโมเดลที่พัฒนาไว้ เพื่อให้ผู้บริหารโรงพยาบาลสามารถดูข้อมูลเชิงลึกเกี่ยวกับการทำนายการแอดมิทและสถานการณ์ overcrowding และสามารถนำข้อมูลเหล่านี้ไปใช้ในการตัดสินใจและจัดการทรัพยากรได้อย่างมีประสิทธิภาพ

**Chapter 4: Proposed Solution**

**4.2 KNN Modeling**

**K-Nearest Neighbors (KNN): Chosen for its simplicity and interpretability.**

**Random Forest (RF): Selected for its robustness in handling complex, non-linear data.**

**Gradient Boosting (GB): Known for its superior performance in handling imbalanced datasets.**

จากการวัดประสิทธิภาพด้วยตัวชี้วัดเหล่านี้ เราพบว่าโมเดลมีค่า Accuracy, Precision, Recall และ F1 Score ที่แสดงถึงความสามารถในการทำนายสถานะความแออัดในโรงพยาบาลได้อย่างเหมาะสม ค่า F1 Score ที่สูงแสดงถึงความสมดุลที่ดีระหว่าง Precision และ Recall ซึ่งทำให้มั่นใจได้ว่าโมเดลสามารถจำแนกข้อมูลกลุ่มที่เกิดความแออัดจริงได้อย่างมีประสิทธิภาพ รายละเอียดและข้อมูลการประเมินเพิ่มเติมจะแสดงในตารางและกราฟในบทถัดไปเพื่อการวิเคราะห์ที่ลึกขึ้น

---

\*\*คำพูดในการพรีเซนต์\*\*:

"ในงานวิจัยนี้เรามี 3 วัตถุประสงค์หลัก:

1. การใช้เทคนิคการแสดงผลข้อมูลเพื่อสำรวจและดึงข้อมูลสำคัญจากชุดข้อมูล ซึ่งจะช่วยให้เราสามารถระบุแนวโน้มการเข้าโรงพยาบาล, รูปแบบการวินิจฉัย, และผลกระทบจากข้อมูลทางประชากร เช่น อายุ, เพศ, และประเภทของการรับเข้าโรงพยาบาล ที่จะนำมาใช้ในการพยากรณ์ความแออัดในโรงพยาบาล

2. เปรียบเทียบและประเมินโมเดลการเรียนรู้ของเครื่องหลายๆ ตัว เช่น K-Nearest Neighbors, Random Forest และ Gradient Boosting เพื่อหาวิธีที่มีประสิทธิภาพที่สุดในการทำนายช่วงเวลาที่มีความเสี่ยงต่อการเกิดความแออัด

3. นำ Explainable AI (XAI) มาใช้ผ่าน LIME เพื่อทำให้การทำนายของโมเดลสามารถเข้าใจและอธิบายได้ โดยจะช่วยให้ผู้เชี่ยวชาญในด้านการแพทย์สามารถตัดสินใจอย่างมีข้อมูลและเป็นหลักฐานมากยิ่งขึ้น"

**Chapter 5: Conclusion**

**เริ่มต้นด้วยการสรุปผลการศึกษา: “ในการศึกษาครั้งนี้ เราได้ใช้โมเดลการเรียนรู้ของเครื่องเพื่อทำนายความแออัดในโรงพยาบาลจากข้อมูลการเข้ารับการรักษาพยาบาลในอดีต ซึ่งผลการศึกษาพบว่าโมเดล Gradient Boosting มีประสิทธิภาพในการทำนายความแออัดได้ดีที่สุด เมื่อเทียบกับโมเดล K-Nearest Neighbors และ Random Forest.”**

**อธิบายความสำคัญของผลการศึกษา: “ผลการศึกษาแสดงให้เห็นว่าเราสามารถใช้ข้อมูลการวินิจฉัยโรคและการรักษาพยาบาลในอดีตเพื่อคาดการณ์ว่าช่วงเวลาใดในโรงพยาบาลจะมีความแออัด โดยไม่ต้องอิงกับช่วงเวลาหรือวันเวลาเฉพาะ ทำให้สามารถใช้การทำนายเพื่อช่วยในการวางแผนและจัดการทรัพยากรของโรงพยาบาลได้ดียิ่งขึ้น.”**

**พูดถึงข้อจำกัดของการศึกษา: “อย่างไรก็ตาม การศึกษานี้มีข้อจำกัด เช่น ข้อมูลที่บางครั้งอาจไม่สมบูรณ์ ซึ่งอาจมีผลต่อความแม่นยำของโมเดลและการทำนาย.”**

**เสนอแนวทางการศึกษาต่อในอนาคต: “ในการศึกษาครั้งต่อไป เราสามารถปรับปรุงข้อมูลที่ใช้ในการฝึกโมเดลให้ดียิ่งขึ้น และลองใช้โมเดลอื่น ๆ หรือเทคนิคต่าง ๆ เพื่อเพิ่มความแม่นยำในการทำนายการแออัด.”**

**สรุปข้อเสนอแนะหรือความสำคัญ: “การใช้ Explainable AI (เช่น LIME) ช่วยให้ผู้ใช้สามารถเข้าใจการตัดสินใจของโมเดลได้ดีขึ้น ซึ่งจะช่วยให้บุคลากรทางการแพทย์สามารถใช้ข้อมูลในการตัดสินใจได้อย่างมีประสิทธิภาพและโปร่งใสมากขึ้น.”**

Reference:

[1]

# การอ้างอิง

|  |  |
| --- | --- |
| [1] | Aminuddin, W. M. W. M., & Ismail, W. R. (2021). Improving emergency department overcrowding in Malaysian government hospitals. Journal of Quality Measurement and Analysis JQMA, 17(1), 19-39. |
| [2] | Savioli, G., Ceresa, I. F., Gri, N., Bavestrello Piccini, G., Longhitano, Y., Zanza, C., ... & Bressan, M. A. (202 2). Emergency department overcrowding: understanding the factors to find corresponding solutions. Journal of pers onalized medicine, 12(2), 279. |
| [3] | Jefferson, Guerrero., A., Alqarni., Rock, Parreno, Cordero., Imad, AlJarrah., Mohsen, Almahaid. (2024). Perceived Causes and Effects of Overcrowding Among Nurses in the Emergency Departments of Tertiary Hospitals: A Multicenter Study. Risk Management and Healthcare Policy, doi: 10.2147/rmhp. s454925 |
| [4] | Mathieu, Oberlin., Emmanuel, Andrès., Martin, Behr., Sabrina, Kepka., P., Le, Borgne., Pascal, Bilbault. (2020). La saturation de la structure des urgences et le rôle de l'organisation hospitalière: réflexions sur les causes et les solutions. Revue de Médecine Interne, doi: 10.1016/J.REVMED.2020.05.023 |
| [5] | CADTH. (2024). Artificial Intelligence for Patient Flow. Canadian journal of health technologies, doi: 10.51731/cjht.2024.877 |
| [6] | Rebecca, Woods., Rebecca, Sandoval., Gregory, Vermillion., Brenda, Bates-Jackson., Audrey, Nwankwo., Catherine, P, Canamar., Laura, Sarff. (2020). The Discharge Lounge: A Patient Flow Process Solution.. Journal of Nursing Care Quality, doi: 10.1097/NCQ.0000000000000469 |
| [7] | Maji, Ibrahim, Abubakar. (2022). 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, doi: 10.46792/fuoyejet. v7i4.922 |
| [8] | V. Domova and S. Sander-Tavallaey, "Visualization for Quality Healthcare: Patient Flow Exploration," in Proceedings of the 2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, 2019, pp. 1072-1079, doi: 10.1109/BigData47090.2019.9006351. |
| [9] | M. Islam and S. Jin, "An Overview of Data Visualization," 2019 International Conference on Information Science and Communications Technologies (ICISCT), Tashkent, Uzbekistan, 2019, pp. 1-7, doi: 10.1109/ICISCT47635.2019.9012031. |
| [10] | Addepalli, L., Gaurav, L., Sindhuja, S., Hussain, S., Mookerjee, J., Uppalapati, V., Ali, W., & Sagar, V. (2023). Assessing the Performance of Python Data Visualization Libraries: A Review. International Journal of Computer Engineering in Research Trends, 10(1), 28-39. https://doi.org/10.22362/ijcert/2023/v10/i01/v10i0104 |
| [11] | Kassie, B. A., Senay, A. B., & Tegenaw. (2024). Developing a patient flow visualization and prediction model using aggregated data for a healthcare network cluster in Southwest Ethiopia. PLOS Digital Health, 3(2), e0000455. https://doi.org/10.1371/journal.pdig.0000455 |
| [12] | Gupta, A., Singh, K. B., Singh, K., Kushwaha, P. K., Lohani, B. P., & Kumar, S. (2024). Unveiling Insights: Exploring Healthcare Data through Data Analysis. |
|  |  |
|  |  |

[7] Maji, Ibrahim, Abubakar. (2022). 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, doi: 10.46792/fuoyejet. v7i4.922

[8] V. Domova and S. Sander-Tavallaey, "Visualization for Quality Healthcare: Patient Flow Exploration," in Proceedings of the 2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, 2019, pp. 1072-1079, doi: 10.1109/BigData47090.2019.9006351.

[9] M. Islam and S. Jin, "An Overview of Data Visualization," 2019 International Conference on Information Science and Communications Technologies (ICISCT), Tashkent, Uzbekistan, 2019, pp. 1-7, doi: 10.1109/ICISCT47635.2019.9012031.

[10] Addepalli, L., Gaurav, L., Sindhuja, S., Hussain, S., Mookerjee, J., Uppalapati, V., Ali, W., & Sagar, V. (2023). Assessing the Performance of Python Data Visualization Libraries: A Review. International Journal of Computer Engineering in Research Trends, 10(1), 28-39. https://doi.org/10.22362/ijcert/2023/v10/i01/v10i0104

[11] Kassie, B. A., Senay, A. B., & Tegenaw. (2024). Developing a patient flow visualization and prediction model using aggregated data for a healthcare network cluster in Southwest Ethiopia. PLOS Digital Health, 3(2), e0000455. https://doi.org/10.1371/journal.pdig.0000455

[12] Gupta, A., Singh, K. B., Singh, K., Kushwaha, P. K., Lohani, B. P., & Kumar, S. (2024). Unveiling Insights: Exploring Healthcare Data through Data Analysis.

[13] G., Ratthna S. J., Jothikumar, K., & Priyadharshini, P. (2024). Advancing Healthcare Through Data Science Techniques for Comprehensive Analysis and Visualization of Healthcare Data. Advances in Healthcare Information Systems, IGI Global, pp. 1-15.

[14] Battineni, G., Mittal, M., & Jain, S. (2021). Data Visualization in the Transformation of Healthcare Industries. Springer, Singapore, pp. 1-23.

[15] Asuquo, D., Umoren, I., Osang, F., & Attai, K. (2023). A Machine Learning Framework for Length of Stay Minimization in Healthcare Emergency Department. Studies in Engineering and Technology, 10(1), Article 6372. https://doi.org/10.11114/set.v10i1.6372

[16] Tello, M., Reich, E., Puckey, J., Maff, R., García Arce, A., Bhattacharya, B., & Feijoo, F. N. (2022). Machine learning-based forecast for the prediction of inpatient bed demand. BMC Medical Informatics and Decision Making, 22(1).

[17] Singhal, S., Sharma, A. K., Singh, A. K., Pandey, A., & Sharma, A. K. (2024). Industry-specific AI and ML applications in healthcare focus on improving diagnostics, patient care, and medical research. Advances in Systems Analysis, Software Engineering, and High-Performance Computing. IGI Global, pp 110-124.

[18] Ponce, A. G., López-Bautista, J., & Fernandez-Beltran, R. (2024). Interpretando Modelos de IA en Cáncer de Mama con SHAP y LIME. Ideas en Ciencias de la Ingeniería, Universidad Autónoma del Estado de México, Vol. 2, Iss: 2, pp. 15-15.

[19] Salih, A., et al. (2024). A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME. Advanced Intelligent Systems.

[20] Huang, G., et al. (2024). From Explainable to Interpretable Deep Learning for Natural Language Processing in Healthcare: How Far from Reality? Computational and Structural Biotechnology Journal, Vol. 24, pp. 362-373.

[21] Kovvuri, V. R. R. (2024). Explainable Artificial Intelligence across Domains: Refinement of SHAP and Practical Applications. [No Journal Name Provided].

[22] World Health Organization. International Classification of Diseases 10th Revision (ICD-10), 2019, https://icd.who.int/browse10/2019/en.

[23] A. A. Tokuç, "Splitting a Dataset into Train and Test Sets," Tarnum Java SRL, 12 May 2023. [Online]. Available: https://www.baeldung.com/cs/train-test-datasets-ratio. [Accessed 11 November 2023].

[24] Duda, R. O., Hart, P. E., & Stork, D. G. (2001). Pattern classification. John Wiley & Sons.