Problem Selection

Q1: why?

In hospitals, resource allocation management is a multifaceted process that involves managing resources such as personnel, equipment, and supplies to meet patients' needs while maximizing operational efficiency. It is an essential aspect of healthcare administration that can influence patient outcomes, cost-effectiveness, and resource utilization. In recent years, resource allocation management has become even more important due to the rising demand for healthcare services and the COVID-19 pandemic, which has placed an unprecedented strain on healthcare systems around the world.

The COVID-19 pandemic has illuminated the significance of resource allocation management in healthcare systems. Critical resources, such as hospital beds, medical equipment, and healthcare personnel, are in short supply due to the pandemic. To ensure that patients receive the necessary care, healthcare providers and hospital administrators have been forced to make difficult decisions regarding the allocation of resources (Laventhal *et al.*, 2020). The significance of hospital resource allocation management in this context cannot be overstated. Effective strategies for resource allocation management can guarantee patients receive the necessary care while minimizing waste and optimizing the use of available resources. In addition, evidence-based strategies for resource allocation can improve patient outcomes, reduce costs, and boost the overall efficiency of healthcare systems.

The American Geriatrics Society (AGS) has published a policy statement (Farrell *et al.*, 2020) outlining ethical considerations that should be considered when developing strategies for allocating scarce resources during an emergency involving older adults. The statement emphasizes the need to avoid using age as an arbitrary criterion for excluding anyone from care, instead focusing on clinically relevant factors and the disparate impact of social determinants of health. Also recommended are the formation and staffing of triage committees charged with allocating scarce resources, the development of transparent and uniformly applied institutional resource allocation strategies, and the facilitation of appropriate advance care planning. These recommendations are essential for addressing strategies for resource allocation during the COVID-19 pandemic and future health crises. The AGS statement emphasizes the ethical implications of resource allocation decisions and the need for equitable emergency resource allocation strategies that avoid discriminatory terminology and practice. To ensure that healthcare resources are used efficiently, fairly, and with consideration for the needs of vulnerable populations, hospital resource allocation management is crucial.

Due to illness and quarantine, the shortage of healthcare workers has further complicated the situation (Kirkpatrick *et al.*, 2020). As a result, clinicians and healthcare systems must make challenging decisions regarding resource allocation and patient triage. This situation has raised important ethical questions regarding the allocation of resources and the balance between the duty to advance the interests of the individual patient and the duty to manage resources for populations. The significance of hospital resource allocation cannot be overstated, particularly in the context of a pandemic. Critical resource scarcity creates ethical dilemmas that necessitate careful consideration of competing obligations, such as the duty to protect the interests of individual patients, the duty to use resources efficiently, and the duty to ensure equity in resource allocation. Failure to address these ethical concerns can result in unwarranted harm to patients and healthcare professionals and can erode public confidence in healthcare systems. In order for healthcare providers to make informed decisions regarding resource allocation in times of scarcity, it is crucial to develop ethical guidelines and policies.

Even though the Covid-19 pandemic is in the past, it has demonstrated the critical importance of being prepared for similar public health crises in the future. One of the most important lessons learned from this pandemic is the importance of effective healthcare management and resource allocation to ensure that hospitals are equipped to handle an influx of patients without compromising the quality of care provided. In the event of a future pandemic, accurate predictions of patient length of stay can optimize resource allocation and improve hospital healthcare management efficiency. By accurately predicting the length of stay for each patient on an individual basis, hospitals can better allocate resources such as beds, staff, and medical supplies, allowing them to provide better care for patients and reduce the risk of infection among staff and visitors. Work on the problem of accurately predicting the length of hospital stays for patients is thus not only pertinent in the context of the recent Covid-19 pandemic, but also essential for preparing for potential future public health crises. By developing accurate and efficient solutions to this problem, healthcare management can be enhanced, and hospitals will be better prepared to face future challenges.

Q2: What do you expect to solve?

Solving this issue can have a significant impact on the healthcare industry by increasing the effectiveness of hospital healthcare administration. By accurately predicting each patient's length of stay, hospitals can maximize their resources and allocate them in the most effective manner. This can help reduce patient wait times and enhance the quality of care overall. The COVID-19 pandemic has highlighted the significance of effective healthcare administration. During the pandemic, hospitals and healthcare systems were placed under tremendous strain, and many struggled to accommodate the influx of patients. By predicting the length of stay for each patient, hospitals can plan and allocate their resources more efficiently, allowing them to be better prepared for future pandemics and other healthcare emergencies. Additionally, improved healthcare management can reduce healthcare costs. By optimizing the length of stay for patients, hospitals can reduce the number of costly and burdensome unnecessary hospitalizations. Hospitals can reduce the risk of hospital-acquired infections, which can lead to additional costs and even deaths, by reducing the length of stay.

Additionally, resolving this issue can improve the overall quality of patient care. By accurately predicting the length of stay for each patient, hospitals are able to identify patients who are at high risk for a longer hospital stay and optimize their treatment plan accordingly. This can help reduce the risk of complications and ensure patients receive the necessary care in a timely manner.

Q3: Who will use that solution?

The solution to this problem can be utilized by various healthcare industry stakeholders. This solution enables hospitals and healthcare management organizations to optimize resource allocation and enhance operational efficiency. By accurately predicting each patient's length of stay, hospitals can plan their logistics, including room and bed allocation, in advance. This can help prevent overcrowding and decrease the likelihood of staff and visitor infections. Moreover, insurance firms can also benefit from this solution. By accurately predicting each patient's length of stay, insurance companies can make more informed decisions regarding the coverage they provide to policyholders. This can reduce the overall costs of healthcare for both patients and insurance companies.

In addition, policymakers and public health officials can utilize this solution to better prepare for and respond to future pandemics and other healthcare emergencies. By having accurate predictions of a patient's length of stay, policymakers can more effectively allocate resources and make better decisions regarding healthcare policies and regulations.

Understanding the Nature of the Problem

Q1: root cause?

This dataset addresses a problem whose root cause is ineffective healthcare management, which can result in longer patient length of stay (LOS) in hospitals. Observing and predicting patient length of stay is crucial if one wishes to enhance the effectiveness of hospital healthcare management. At the time of admission, it can help hospitals identify patients with a high risk of LOS (i.e., patients who will stay longer). Once identified, patients at high risk for length of stay can have their treatment plan optimized to reduce length of stay and staff or visitor infection risk. In addition, prior knowledge of LOS can aid in logistics such as the allocation of rooms and beds.

The issue arises when hospitals are unable to accurately predict the length of patients' stays. This may result in hospital overcrowding, a lack of resources, a delay in the admission of new patients, and an increase in healthcare costs. Long-staying patients may also be more susceptible to hospital-acquired infections and other complications. In addition, prolonged LOS can negatively impact the mental and emotional health of patients and their families, resulting in elevated levels of stress and anxiety.

Q2: how can data mining help?

Data mining can be an effective resource for hospitals in allocating resources and optimizing patient care (Ayyoubzadeh *et al.*, 2020). By analyzing data on factors influencing hospital performance indicators such as bed occupancy rate, LOS, bed turnover, bed turnover interval, and mortality rates, hospitals can gain insight into the efficacy of their organizational units. This data can then be utilized to identify bottlenecks and make well-informed decisions regarding resource allocation and service enhancements. In the case of hospital length of stay, data mining techniques can be used to identify factors such as delayed tests and imaging, complications, and comorbidities that contribute to longer hospital stays. By analyzing this data, hospitals are able to develop targeted interventions to address these factors and reduce length of stay, thereby enhancing resource utilization and patient outcomes. Based on historical data and current trends, data mining can also assist hospitals in predicting demand, by using techniques such as Naïve Bayes (Vikramkumar, B and Trilochan, 2014) and K-nearest Neighbors (Peterson, 2009), for resources like beds and ventilators. This can allow hospitals to allocate resources more efficiently and effectively, ensuring that patients receive the care they require promptly.

It is possible for hospitals, patients, and other stakeholders to gain significant benefits from the utilization of data mining in the management of healthcare. The management of the hospital is able to make educated decisions regarding the distribution of resources, the scheduling of staff, and the development of treatment plans when they have an accurate prediction of how long patients will remain in the facility. In the long run, this could lead to improved health outcomes, increased patient satisfaction, and cost savings for the hospital. For instance, hospitals can optimize the use of hospital beds and staff, reduce wait times, and minimize the risk of staff and patient infections by making accurate predictions of the length of stay patients will have in the hospital. The predictions can also be used by hospitals to plan discharge dates and schedule follow-up

appointments, both of which can help reduce the number of patients who require readmission and improve the overall outcomes for patients. Accurate projections of patients' lengths of stays have benefits not only for the hospitals themselves, but also for the patients themselves. Patients are able to better plan their recovery and reduce the stress and uncertainty associated with extended hospital stays when they are provided with accurate information regarding their hospital stay. In addition, patients have the opportunity to improve their readiness to take control of their own health after being released from the hospital. In addition, data mining can provide insights into patterns and trends in patient data, which can assist hospitals in identifying risk factors and early warning signs of complications. This information can be mined from patient records. This can result in proactive interventions that can prevent or lessen the impact of adverse events.

Understanding the Data

Describe data

The training dataset for the aforementioned problem contains 18 columns and total instance are 318438, while test dataset contains same columns as training set except target variable which is ‘Stay’ and it contains 137057 instances. Dataset is available in comma separated values format. Each patient case is assigned a unique identifier denoted by the "case id" column. "Hospital code" is the unique identifier assigned to each hospital where the patient is receiving treatment. "Hospital type code" indicates the type of hospital where the patient is receiving treatment. The city where the hospital is located is indicated in the fourth column, labelled "City Code Hospital." "Hospital region code" refers to the region in which the hospital is located. The sixth column, titled "Available Extra Rooms in Hospital," indicates the number of available extra rooms in the hospital for admitting new patients. The seventh column is titled "Department," which refers to the patient's treating department. The eighth column, "Ward Type," indicates the type of ward in which the patient is receiving treatment. The "Ward Facility Code" column contains the code assigned to the facility where the ward is located. The tenth column is labelled "Bed Grade," which indicates the grade assigned to the patient's bed. "patientid" is the eleventh column and is a unique identifier assigned to each patient. The twelfth column, "City Code Patient," indicates the patient's city of origin. The thirteenth column is labelled "Type of Admission," which indicates whether the patient's admission was an emergency or planned.

The fourteenth column, "Severity of Illness," refers to the patient's illness severity, which may be mild, moderate, or severe. "Visitors with Patient" refers to the number of visitors who are permitted to visit the patient during his or her hospital stay. The sixteenth column is "Age," which indicates the patient's age. "Admission Deposit" is the seventeenth column and refers to the amount of deposit paid by the patient upon admission. The final column is "Stay," which is the dependent variable, i.e., the duration of the patient's hospital stay. This column is divided into eleven distinct categories ranging from 0 to 10 days to over 100 days. This dataset contains information that can be used to predict the length of a patient's hospital stay, including patient demographics, hospital information, and clinical characteristics.

Why a few columns are not useful in further analysis?

The identifiers 'case id' and 'patientid' are unique to each hospitalisation event and patient, respectively. They provide no meaningful information regarding the allocation of hospital resources. The 'case id' is merely a unique identifier for each case; it does not contain any pertinent information about the patient, hospital, or resource allocation. It is only used for monitoring and to prevent duplication. Similar to 'patientid,' which is a unique identifier assigned to every patient, 'patientid' is not useful for resource allocation analysis. Analysis of resource allocation examines how hospital beds, physicians, and medical equipment are allocated to patients based on their medical condition, severity, and other variables. The patient's identifier provides no information that can be used to comprehend how hospital resources are allocated.

Filling null values

In KNN imputation, the value of k determines the number of nearest neighbours to consider when filling in missing values. As a heuristic, the square root of the total number of missing values is frequently used to determine an acceptable value for k. A value of k that is too small may result in missing values being imputed with noisy or irrelevant data, while a value of k that is too large may result in missing values being imputed with data that is too dissimilar. The columns 'Bed Grade' and 'City Code Patient' in the given code are selected for imputation using the KNN imputer. The value k=11 is used for 'Bed Grade,' while k=67 is used for 'City_Code_Patient,' which is the square root of the total number of missing values in the respective columns. The KNNImputer object is created with the specified values of k and missing values set to NaN to represent the dataset's missing values. The imputer object is then fitted to the selected columns using the fit_transform() method, which imputes missing values based on the values of their nearest neighbors. Using the respective column names, the imputed values are assigned back to the original dataframe.

How does KNN imputer work?

K-Nearest Neighbors (KNN) imputation (Zhang, 2012) is a technique for estimating missing values from their k nearest neighbors, where k is a predetermined number of neighbors to consider. It is a non-parametric method, which means it makes no assumptions about the data's underlying distribution. The KNN imputer operates by calculating the distances between each missing value and the remaining data points. The algorithm then selects the k nearest data points and uses their known values to calculate the missing value. Common distance metrics include the Euclidean distance, the Manhattan distance, and the cosine similarity. KNN imputation is commonly regarded as a better method for filling in missing values when compared to mean, median, and mode imputation. This is because it considers the relationships between variables in the dataset and uses this knowledge to impute missing values. This can result in more precise imputations and less data bias.

What is skewness?

The skewness of a distribution is a measure of its asymmetries. It indicates the deviation of a distribution from the symmetry of a normal distribution. A distribution may be positively skewed, negatively skewed, or symmetrical. In a positively biased distribution, the right-hand side of the distribution has a lengthier tail than the left. This indicates that the distribution has a greater number of values on the left and fewer extreme values on the right. The distribution's mean will

be greater than its median. In a distribution that is negatively biased, the distribution's left-hand tail is lengthier than its right-hand tail. This indicates that the distribution has a greater number of values on the right side and a smaller number of extreme values on the left side. The distribution's mean will be less than its median. Skewness provides information about the character of the data. For instance, if a dataset has a positive skew, it indicates that the dataset contains most low values and a few extremely high values. This may indicate that the dataset contains outliers or extreme values. Similarly, a negative skew indicates that a dataset contains most higher values and a few very low values. This may suggest that the dataset has been truncated or has a floor effect. Symmetrical datasets are simpler to model and analyze.

Insights from skewness analysis:

- Hospital_code, Ward_Facility_Code, and Bed Grade have slightly negative skewness values, indicating a slightly longer or fatter left tail, but the distribution is relatively symmetrical.
- The skewness values for Hospital_region_code, Ward_Type, Age, and Type of Admission are close to zero, indicating a roughly symmetrical distribution.
- The City_Code_Hospital and Severity of Illness variables have slightly positive skewness values, indicating a slightly longer or fatter tail on the right side of the distribution, but the distribution is largely symmetrical.
- Available Extra Rooms in Hospital, Admission_Deposit, and Hospital_type code have a positive skewness value greater than 0.5, indicating a distribution with more extreme values in the positive direction.
- Visitors with Patient and City_CodePatient have a very high positive skewness value, indicating a right-skewed distribution with many extreme values in the positive direction.

Univariate Analysis

Univariate analysis is a form of statistical analysis used to examine a single variable. In this form of analysis, the distribution, central tendency, variability, and morphology of the studied variable's distribution are analyzed. Univariate analysis can be used to identify outliers and examine data normality. It can also be used to compute the fundamental summary statistics of the data, such as the mean, median, mode, range, and standard deviation, as well as to identify missing values in the data.

Insights from univariate analysis:

- The majority of admitted patients were assigned hospital codes 19, 26, 28, 27, and 23.
- Most hospitals in the dataset belong to type "a" for Hospital type code. Type "a" hospitals admit the greatest number of patients, followed by types "b" and "c".
- Most hospitals within the dataset are in cities with the codes 1, 2, 6, and 7.
- Hospital region code: "X" is the predominant region code for hospitals. The region with the highest number of admitted patients is "X," followed by "Y" and "Z."
- The number of available additional hospital rooms ranges from zero to twenty-four. Most hospitals have between two and six extra rooms.

- The majority of patients were admitted to the department of gynecology, followed by the departments of anesthesia, surgery, and radiotherapy.
- The private ward admitted the greatest number of patients, followed by the semi-private and general wards.
- Ward Facility Code: "F" is the facility code for most wards. The greatest number of patients are admitted to facility code "F," followed by facility codes "E" and "D."
- Most hospitals within the dataset have bed grades 2 and 3.
- City Code Patient: Most patients come from cities with the codes 1, 2, 7, 8, and 10.
- Most patients admitted were emergency cases, followed by urgent and elective admissions.
- Most patients experienced moderate illness severity, followed by mild and severe cases.
- The number of visitors accompanying patients ranges from zero to thirty-two. The majority of patients had no visitors, followed by those with two.
- Most patients in the dataset are between the ages of 31 and 40.
- Admission Deposit: The amount of the admission deposit ranges from $1,800 to $11,000. Most patients had admission deposits between $2,000 and $6,000.

Bivariate Analysis

simultaneously to ascertain their relationship. In other terms, it is an examination of the association between two variables. The variables may be numerical or categorical. The primary purpose of bivariate analysis is to determine whether or not a relationship exists between the two variables and to quantify the relationship's strength and direction. Statistical techniques such as correlation analysis, regression analysis, and scatter plots are utilized for this purpose.

Bivariate analysis insights:

- Compared to other cities, patients admitted to hospitals in city code 1 have the highest proportion of lengthy stays (over 20 days).
- The Trauma department has the highest number of patients with extended stays, followed by the Surgery department and the Medical department.
- Patients with severe illness are more likely to require longer hospital stays than those with moderate or minor conditions.
- Length of stay is significantly influenced by admission type (emergency, urgent, or elective). Compared to patients admitted as urgent or elective cases, patients admitted as emergencies have longer hospital stays.
- There is no significant correlation between the number of visitors and length of stay.
- Patients with larger admission deposits tend to have longer hospital stays.

Preparing Data for Machine Learning Models

Label Encoding

Label encoding (Pedregosa *et al.*, 2011) is the process of transforming categorical data into numerical data by designating a distinct integer value to each category. This method is used to convert categorical data into a format that can be utilized by machine learning algorithms.

Take the "Department" column from the hospital dataset as an example. There are five categories: radiotherapy, anesthesia, gynecology, tuberculosis and chest diseases, and surgery. To apply label encoding, we would assign a unique integer value to each category, as follows:

- Radiotherapy: 0
- Anesthesia: 1
- Gynecology: 2
- TB & Chest disease: 3
- Surgery: 4

Similarly, label encoding was applied to other categorical columns in the dataset, such as 'Hospital_type_code,' 'City_Code_Hospital,' 'Ward_Type,' 'Ward_Facility_Code,' 'Type of Admission,' and 'Severity of Illness.'

Standard Scaling

Standard scaling (Pedregosa *et al.*, 2011) is a variety of feature scaling used to convert numerical data to have a mean of 0 and a standard deviation of 1. This technique is commonly used in machine learning algorithms to normalize the data's features, thereby enhancing the algorithm's performance. The standard procedure for scaling includes the following steps:

- Calculate the mean and standard deviation for each numerical characteristic in the training data.
- Subtract the mean from each feature's value.
- Each value is multiplied by the standard deviation.

This yields a transformed characteristic with a mean of 0 and a standard deviation of 1.

References

Ayyoubzadeh, S.M. *et al.* (2020) 'A study of factors related to patients' length of stay using data mining techniques in a general hospital in southern Iran', *Health information science and systems*, 8, pp. 1–11.

Farrell, T.W. *et al.* (2020) 'AGS position statement: resource allocation strategies and age-related considerations in the COVID-19 era and beyond', *Journal of the American Geriatrics Society*, 68(6), pp. 1136–1142.

Kirkpatrick, J.N. *et al.* (2020) 'Scarce-resource allocation and patient triage during the COVID-19 pandemic: JACC review topic of the week', *Journal of the American College of Cardiology*, 76(1), pp. 85–92.

Laventhal, N. *et al.* (2020) 'The ethics of creating a resource allocation strategy during the COVID-19 pandemic', *Pediatrics*, 146(1).

Pedregosa, F. *et al.* (2011) 'Scikit-learn: Machine Learning in {P}ython', *Journal of Machine Learning Research*, 12, pp. 2825–2830.

Peterson, L.E. (2009) '{K}-nearest neighbor', *Scholarpedia*, 4(2), p. 1883. Available at: https://doi.org/10.4249/scholarpedia.1883.

Vikramkumar, B, V. and Trilochan (2014) 'Bayes and Naive Bayes Classifier', *CoRR*, abs/1404.0. Available at: http://arxiv.org/abs/1404.0933.

Zhang, S. (2012) 'Nearest neighbor selection for iteratively kNN imputation', *Journal of Systems and Software*, 85(11), pp. 2541–2552. Available at: https://doi.org/https://doi.org/10.1016/j.jss.2012.05.073.