Predictive Modelling of Hospital Stay Duration: A Machine Learning Comparative Study

Ollscoil na Gaillimhe
University of Galway

Scoil Ghnó agus Eacnamaíochta J.E. Cairnes
J.E. Cairnes School of Business and Economics

MSc Postgraduate Project

Business Analytics

University of Galway

Project Members:

1. **Adithya Muralidharan K V - 23104697, a.kv1@universityofgalway.ie**
2. **Ruben Mathew - 23105480, r.mathew8@universityofgalway.ie**
3. **Vinod Rajan - 23101940, v.rajan2@universityofgalway.ie**

Integrity Statement

In submitting this Major Project, we are aware that it our responsibility to adhere to the submission guidelines. Please tick (double click… or Yes/No) for the following:

	Yes	No
We aware of what the University of Galway plagiarism policy entails.	**Y**	

Declaration for this Assignment Submission*:*

- *In submitting this work, we confirm that it is entirely my own. We acknowledge that We may be invited to interview if there is any concern in relation to the integrity, and we are aware that any breach will be subject to the University's Procedures for dealing with plagiarism (**QA220 Academic Integrity Policy and Appendix**:* *https://www.universityofgalway.ie/media/registrar/policiesmay2023/QA220-Academic-Integrity-Policy-v2.0-Sept-2023.pdf* *https://www.universityofgalway.ie/media/registrar/policiesmay2023/Appendix-1-QA220-Academic-Integrity-Policy-v2.0-Sept-2023.pdf).*

Predictive Modelling of Hospital Stay Duration: A Machine Learning Comparative Study

ABSTRACT

This study explores the field of healthcare analytics with a particular focus on length of stay (LOS) prediction for hospitalised patients. Using Kaggle's AV Healthcare Analytics II dataset, which contains a wide range of patient characteristics, we investigate the use of different machine learning (ML) models for predictive analytics. We want to evaluate the prediction of patient LOS using Decision Trees, Random Forest, Logistic Regression, Gradient Boosting, Extreme Gradient Boosting (XGBoost), k-Nearest Neighbours (kNN), CatBoost, and Neural Networks.

The dataset provides a wealth of patient data, including demographics, medical history, kind of admission, and degree of illness, among other things. Through rigorous experimentation and evaluation, we deploy each ML model to discern patterns and relationships within the data, aiming to accurately predict the duration of hospital stays. Performance metrics such as accuracy, precision, recall, and F1-score are employed to gauge the predictive capabilities of the models.

We use a comparison approach in our analysis to compare the computational efficiency and predictive accuracy of several machine learning methods. We assess how well the models handle the intrinsic complexity of healthcare data and how well they generalise across various patient groups.

Keywords*:* Length of Stay, Hospital, Healthcare Analytics, Machine Learning, Prediction, Comparative Analysis.

CONTENTS

1. INTRODUCTION

1.1 Background and Motivation

LOS prediction in hospital patients is one of the most important applications in healthcare management and resource allocation. With accurate forecasting in LOS, the healthcare providers have the means to maximize their patient flux, optimally utilize resources, and enhance the quality of care. The planning aspects can be further made easy by tracking timely discharges that can help reduce healthcare expenditure by minimizing unwarranted hospitalization.

The stay length in hospital results from both clinical and nonclinical factors. Those which deal with clinical factors include the severity of illness, comorbidity, and treatment procedures. For instance, in a study conducted by Jovanovic et al., (2019), it came to light that LOS has been affected significantly by many clinical factors; therefore, it is headquartered to identify and provide good treatment to high-risk patients within the earliest stages of being hospitalized to ensure that they do not end up staying longer in the wards than awaited.

Non-clinical factors, such as admission type, socioeconomic level, and demographic characteristics, also play a major role in influencing the prediction of LOS. In this regard, Blecker et al. (2017) associate socioeconomic characteristics with LOS because of the differences that this factor causes in accessing the services of healthcare and health outcomes.

In predictive analytics, rich and valid information is needed to train accurate forecasting models. However, healthcare real data usually suffers from some intrinsic challenges such as missing values, noisy features, and class imbalance. Feature scaling, normalization, and missing value management are some of the very important preprocessing steps in ensuring that the dataset is accurate and consistent. Notably, SMOTE is one of the techniques that have been greatly applied to address class imbalance and improve the predictive performance of models.

1.2 Research Objective

Now, ML techniques have come in with very new approaches to predictive modelling of healthcare. Some robust methodologies are defined for complex datasets to obtain insight into action manners. The set of ML algorithms used for predicting LOS included Decision Trees, Random Forest, Logistic Regression, Gradient Boosting, Extreme Gradient Boosting, or XGBoost, k-NN, CatBoost, and Neural Networks. Each of them has strengths and limitations.

In this paper, we try to compare the performance of different ML models used for LOS prediction of patients in a hospital using the AV Healthcare Analytics II dataset. We perform an exhaustive comparison of different ML models in terms of performance for understanding the characteristics of different ML algorithms and also to find out the best strategy of LOS prediction applicable in real healthcare environments.

2. LITERATURE REVIEW

2.1 Length of Stay Prediction Methods and Algorithm Review

As part of growing interest in data-driven delivery of healthcare, many studies have employed machine learning models in this study for the prediction of LOS. Jain et al. (2021) performed a comparative study of various ML algorithms for LOS prediction and derived a conclusion about the variety of models that shall be considered based on dataset characteristics for better predictive accuracy.

In the field of feature selection, Feng et al. (2021) provided a data-driven approach to hospital personnel scheduling optimization based on patient prediction and added that feature selection methods in nonlinear fashions are required to increase substantially the predictive power of the study. Secondly, Chan et al. (2022) provided the description of, from an implementation science standpoint, the implementation of prediction models in the emergency department; it describes determinants, outcomes, and real-world impact of prediction analytics in clinical environments.

Mani (2020) has presented a machine learning-based hospital recommendation system. It gave an understanding of how ML techniques could play a massive role in healthcare decision-making and patient outcome improvement. Carter and Potts (2014) have presented electronic patient record systems that could be used for LOS predictions related to primary total knee replacement surgeries. It showed how data-driven approaches in health management could become effective in the practical domain.

Turgeman et al. (2017) developed a machine learning model to predict hospital LOS at admission with valuable insights into factors affecting LOS, hence helping hospitals proactively work on resources. Shea et al. (1995) examined the impacts of computer-generated informational messages put before physicians and their impact on decision-making and LOS, showing the value of technology driven interventions in making clinical workflows and patient outcomes more optimal.

Aghajani and Kargari (2016), developed a data mining-based system that extracts parameters affecting the LOS outcome, but especially in general surgery department. From their findings, they raised a concern that there is a high need for Healthcare predictive analytics that would allow strategies to be successfully carried out, operations streamlined, and proper and effective resource allocation to these institutions is achieved. Again, Haas et al. (2011) addressed privacy concerns associated with electronic health records, emphasizing the need for robust privacy protection mechanisms to safeguard patient confidentiality and data integrity. In the context of ICU length of stay prediction following cardiac surgery, Lafaro et al. (2015) proposed a neural network-based approach leveraging pre-incision variables, underscoring the potential of advanced ML techniques in clinical decision support and patient management

Recent studies have further advanced the field of LOS prediction using machine learning. Zeleke et al. (2023) focused on predicting prolonged length of stay (PLoS) in emergency department (ED) settings using a variety of ML algorithms. This study utilized a combination of structured and unstructured data, including clinical notes and electronic health records (EHRs), to enhance predictive accuracy.

Jaotombo et al. (2023) highlighted the importance of balancing performance and interpretability in ML models for LOS prediction. They incorporated both structured tabular data and unstructured clinical text data, demonstrating that the inclusion of free-text clinical notes alongside structured data can improve model accuracy. Their study also recommended transparent reporting and the use of explainable AI (XAI) techniques to improve model acceptance among clinicians.

An ensemble learning approach was employed in another significant study to predict prolonged LOS after spine correction surgery. This study showed the effectiveness of combining multiple ML models to achieve better predictive performance. The researchers used the Boruta algorithm for feature selection and Bayesian optimization for hyperparameter tuning, achieving robust model performance with high area under the receiver operating characteristic (AUROC) values, Li et al., (2024).

A systematic review and meta-analysis compiled data from multiple studies on LOS prediction, identifying key variables used across different models and assessing their performance using metrics like AUROC. The review highlighted those models trained on medical records data generally outperformed those using administrative data. Additionally, the use of cross-

validation and independent validation sets was emphasized to ensure model robustness and generalizability, Gokhale et al., (2023).

In addition, federated learning (FL), which trains machine learning (ML) models across several decentralized datasets while protecting data privacy, has been investigated for LOS prediction. By utilizing a variety of datasets from various clinical settings, our method demonstrated potential in enhancing model performance without jeopardizing patient anonymity.

This review demonstrates the various uses of machine learning in healthcare analytics, from resource optimization and LOS prediction to clinical decision support and privacy preservation, by synthesizing the body of existing knowledge. To increase the practical usefulness of these prediction models in healthcare settings, future research should keep concentrating on enhancing data integration, model interpretability, and validation techniques.

3. METHODOLOGY

The methodological approach for this study involves a series of carefully designed steps to prepare the dataset, implement various machine learning models, and evaluate their performance. The initial step in the analysis involves loading the dataset into the environment. This dataset includes various features related to hospital stays, such as patient demographics, hospital details, and clinical attributes. Proper loading ensures that the data is correctly formatted and ready for subsequent preprocessing steps.

After loading the data, the next step is to address any missing values. If not properly handled, missing values can bias or invalidate the analysis. In this study, missing values in categorical features like 'Bed Grade' and 'City_Code_Patient' were filled using the mode, which is the most frequent value in those columns. This approach ensures the dataset remains complete and accurate without introducing significant bias.

Next, unnecessary columns that do not contribute to the prediction task are removed. For example, 'case_id' and 'patientid' were identified as irrelevant and hence dropped from the dataset. This step is crucial as it reduces the dataset's dimensionality, simplifying the model and enhancing its performance by focusing only on relevant features.

The subsequent step involves encoding categorical features into a numerical format. Machine learning algorithms require numerical input, so categorical variables such as 'Hospital Type', 'Department', and 'Severity of Illness' were transformed using label encoding. This process

converts categorical labels into numerical values, enabling the algorithms to process the data effectively.

Transforming the target variable 'Stay' into numerical categories was also essential. The length of hospital stay was categorized into numerical labels, simplifying the prediction task. This transformation aligns the target variable with the requirements of classification algorithms, making it easier to predict.

The dataset was then split into training and testing subsets to evaluate the model's performance. An 80-20 split was used, with 80% of the data allocated for training and 20% reserved for testing. This division ensures that the model is evaluated on unseen data, providing a more accurate measure of its performance and generalizability.

To address class imbalance in the training data, the Synthetic Minority Over-sampling Technique (SMOTE) was applied. SMOTE generates synthetic samples for minority classes, balancing the class distribution. Handling class imbalance is crucial for preventing the model from being biased towards majority classes and improving its predictive performance across all categories.

With the data prepared, various machine learning models were trained to predict the length of hospital stay. The models used in this study included Logistic Regression, Random Forest, Decision Tree, k-Nearest Neighbours, CatBoost, XGBoost, Gradient Boosting, and Neural Networks. Training a variety of models allows for a comprehensive evaluation of different algorithms, identifying the best-performing model for the task.

After model training, the models' performance was evaluated using metrics such as accuracy, precision, recall, F1-score, and confusion matrix. Evaluating the models provides insights into their strengths and weaknesses, guiding the selection of the optimal model for deployment.

Hyperparameter tuning helps fine-tune the models, enhancing their performance and predictive accuracy. After tuning, the models were compared to identify the best one. Comparative analysis based on the evaluation metrics was conducted to select the most effective model. Comparing models ensures that the chosen model is the best fit for the prediction task, providing reliable and accurate predictions.

Finally, the performance metrics of the models were visualized to better understand their effectiveness. Plots and charts were used to represent the evaluation metrics visually, aiding in

the decision-making process. Visualizing the performance metrics helps communicate the results clearly and effectively.

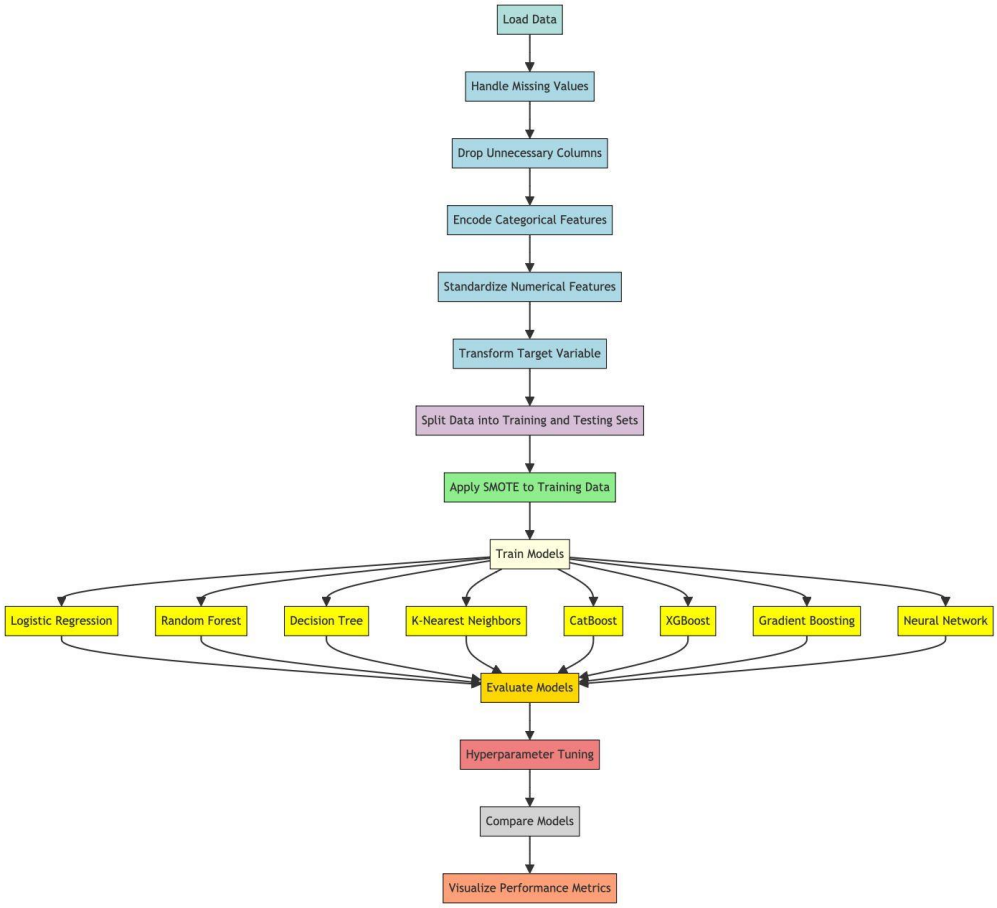

Figure 7: Methodology flowchart

The provided flowchart visually summarizes the methodology, outlining each step from data loading to performance visualization, ensuring a clear and systematic approach to the analysis. This detailed methodological approach ensures accurate dataset preparation, comprehensive model evaluation, and the selection of the best-performing model for predicting hospital length of stay.

4. DATA PROCESSING AND ANALYSIS

4.1 Data Selection

The dataset utilized in this study is sourced from Kaggle, specifically the AV Healthcare Analytics II dataset. This dataset is a comprehensive collection of healthcare records from various hospitals, focusing on predicting the Length of Stay (LOS) for each patient on a case-

by-case basis. The task is crucial for optimal resource allocation and efficient hospital functioning.

Dataset Characteristics:

- **Source**: The dataset is publicly available on Kaggle, contributed by Neha Prabhavalkar.
- **Attributes**: It encompasses a wide array of attributes capturing patient demographics, hospital information, admission details, and clinical variables. Key attributes include:
    - **Demographic Details**: Age, gender, and city code.
    - **Hospital Information**: Hospital code, type of hospital (government or private), and hospital city code.
    - **Admission Details**: Type of admission (emergency or elective), bed grade, and severity of illness.
    - **Clinical Variables**: Diagnosis codes, number of visits prior to admission, and the stay duration in days.

Objective:

- The primary objective is to predict the LOS for patients based on the attributes. Accurate prediction of LOS can help in better planning and resource management within hospitals, ultimately improving patient care and operational efficiency.

Challenges:

- **Missing Values**: Real-world healthcare data often contain missing values due to various reasons such as incomplete patient records or data entry errors.
- **Class Imbalance**: The LOS data are often imbalanced, with some LOS categories having significantly fewer instances than others.
- **Data Complexity**: Healthcare data are complex and heterogeneous, comprising both numerical and categorical variables.

4.2 Data Exploration and Initial Analysis

Data Preprocessing

Data preprocessing is a crucial step in preparing the dataset for machine learning model training.

It involves cleaning the data, handling missing values, encoding categorical variables, feature scaling, and addressing class imbalance.

1. **Data Cleaning and Exploration**:
    - **Handling Missing Values**:
        - **Bed Grade**: Missing values in the 'Bed Grade' column were filled using the mode of the column. This is because 'Bed Grade' is a categorical variable, and the mode represents the most frequent category, which is a reasonable assumption for missing values.
        - **City Code Patient**: Similarly, missing values in the 'City_Code_Patient' column were filled using the mode.

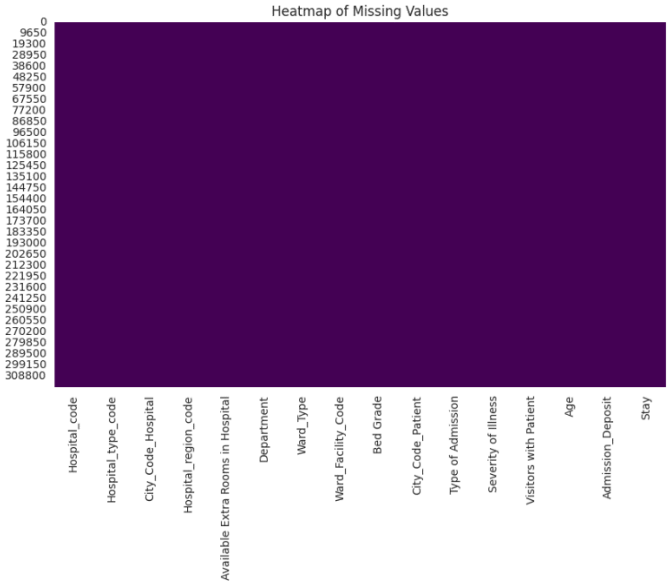

Figure 1: Heatmap of missing values.

2. **Encoding Categorical Variables**:
    - **Label Encoding**: Categorical variables such as 'Hospital Type', 'Hospital Region Code', 'Department', 'Ward Type', 'Ward Facility Code', 'Type of Admission', 'Severity of Illness', 'Age', 'Stay', 'Bed Grade', 'City Code Hospital', and 'City Code Patient' were encoded using Label Encoding. This transforms categorical labels into numerical values.

- **One-Hot Encoding**: For variables with multiple categories, One-Hot Encoding was used to create binary columns for each category, ensuring that the ML models can interpret these categorical variables effectively.

3. **Feature Scaling**:
   - **Standardization**: Numerical features were standardized using the StandardScaler from Scikit-learn. This scales the features to have a mean of 0 and a standard deviation of 1, which is important for algorithms sensitive to feature scaling.
   - **Visualization**: Histograms to visualize the distribution of numerical variables before and after scaling.

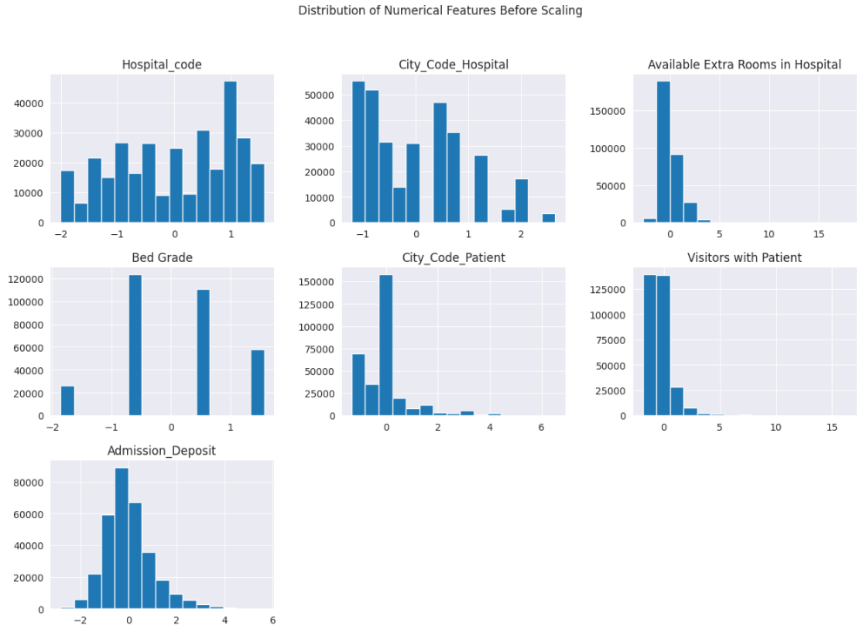

Figure 2: Histograms of numerical variables before scaling.

4. **Addressing Class Imbalance**:
   - **Synthetic Minority Over-sampling Technique (SMOTE)**: SMOTE was applied to the training data to address class imbalance. This technique generates synthetic samples for the minority classes by interpolating between existing samples, thereby balancing the class distribution.

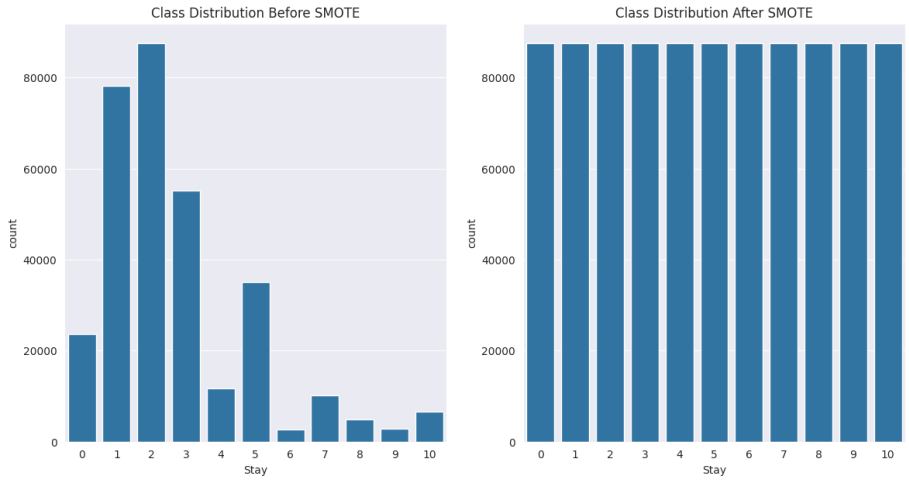

Figure 3: Bar plots of class distribution before and after applying SMOTE

5. **Correlation Analysis**:

   - A correlation matrix was computed to examine the relationships between numerical variables. This helps in understanding multicollinearity and identifying redundant features.

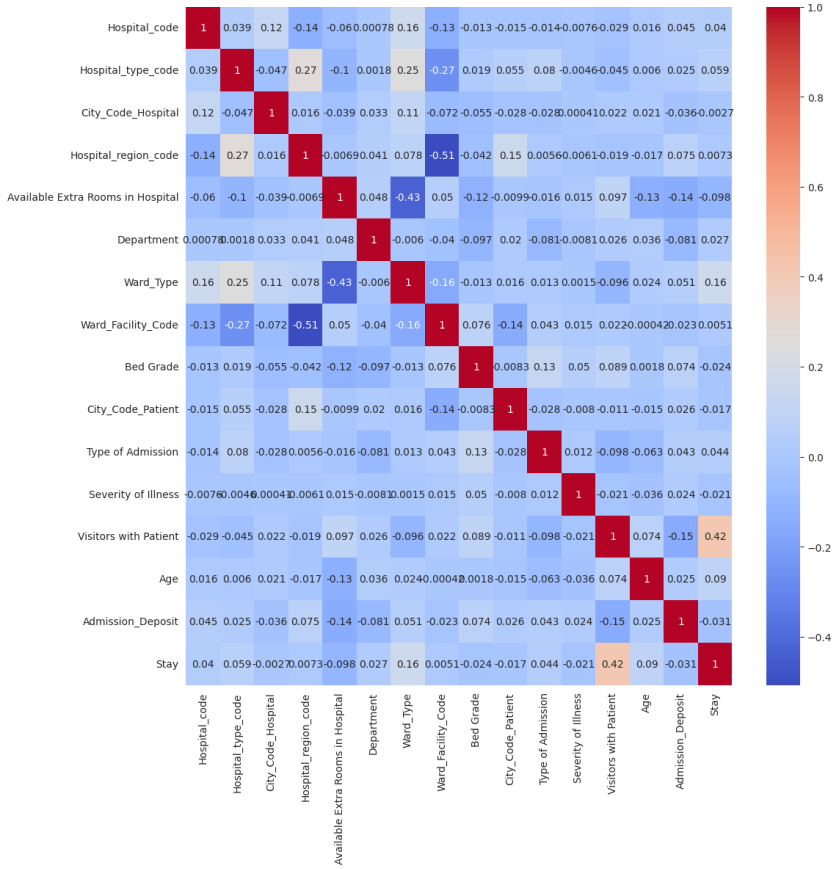

Figure 4: Heatmap of correlation matrix.

4.3 Data Analysis:

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a crucial step in understanding the underlying patterns, distributions, and relationships within the dataset. In this section, we perform EDA to gain insights and prepare the data for modelling.

1. **Distribution of Target Variable (Stay)**
    - The target variable 'Stay' is the length of stay in the hospital, categorized into different ranges.

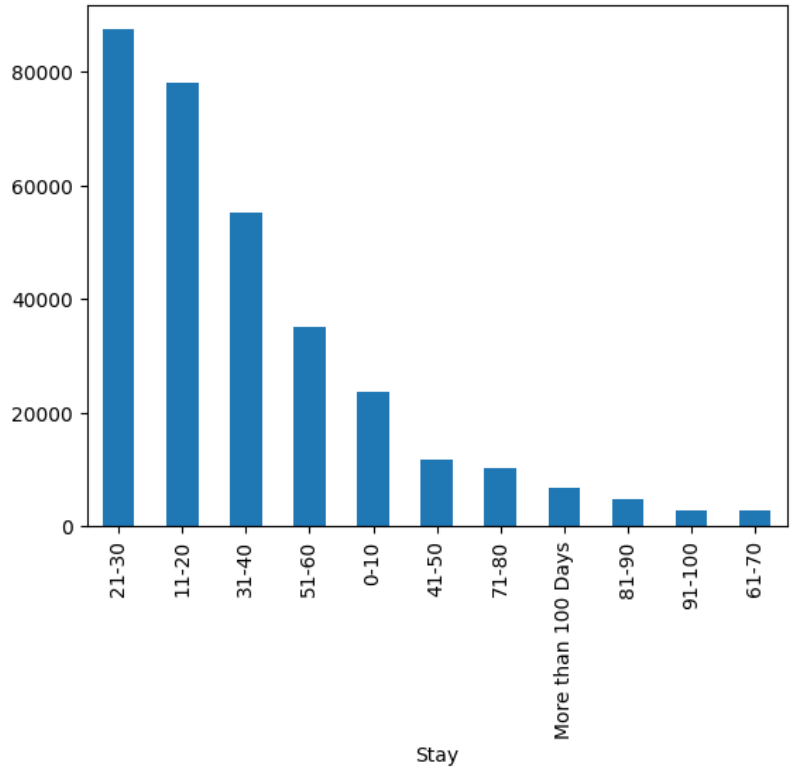

Figure 5: Bar plot to show the distribution of the target variable.

2. **Distribution of Numerical Variables**
    - Analysing the distribution of numerical variables helps to understand their spread, central tendency, and identify any outliers.

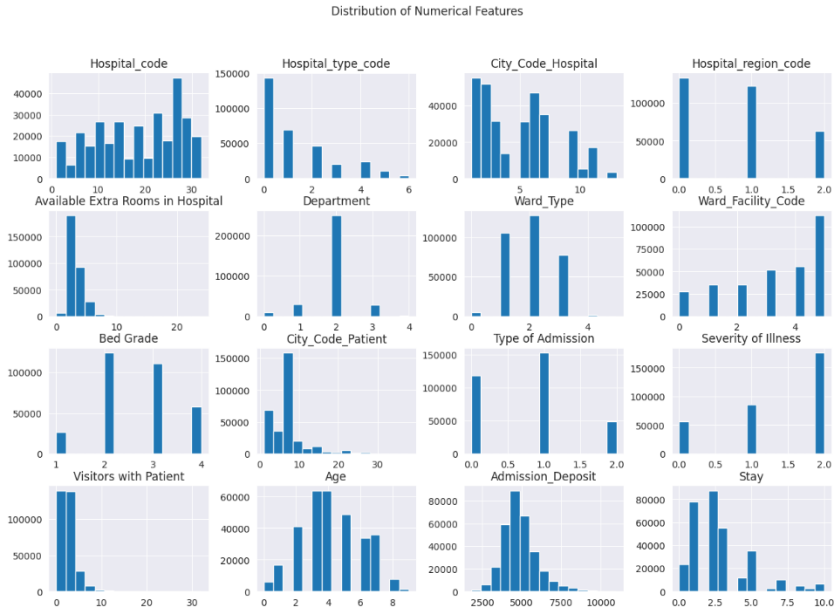

Figure 6: Histograms for numerical variables.

Summary of EDA Results

1. **Distribution of Target Variable (Stay)**:
    - The bar plot shows a skewed distribution of hospital stays, with most patients staying for 21-30 days. Shorter stays are more common, while longer stays are rare, indicating an imbalance in the length of stay categories. This highlights the need for handling class imbalance in predictive models, suggesting techniques such as SMOTE to accurately forecast both frequent short stays and infrequent long stays.
2. **Distribution of Numerical Variables**:
    - The dataset's various numerical feature distributions are depicted in the bar charts. There is variability in hospital and city codes, with some being used more frequently than others. The distributions of available spare rooms and bed grades are skewed in favour of smaller numbers. The patients' age distribution is tilted toward younger ones, and the mid-range values are where admission deposits highest. These findings emphasize how crucial it is to correct any biases and imbalances that may exist during preprocessing in order to improve model performance and dependability when predicting hospital length of stay.

5. IMPLEMENTATION

Libraries and Tools:

Initially, several libraries were imported to streamline the stages of data preprocessing, model training, and evaluation. These included:

- Pandas for data manipulation
- NumPy for numerical operations
- Matplotlib and Seaborn for data visualization
- Various modules from scikit-learn (sklearn) for implementing machine learning models and preprocessing techniques
- TensorFlow for constructing and training neural networks
- Imbalanced-learn (imblearn) for addressing class imbalance using SMOTE (Synthetic Minority Over-sampling Technique)

The classification task utilized several machine learning models, each with distinct characteristics and evaluation metrics.

- **Logistic Regression** was selected as the baseline model, leveraging standardized numerical features to predict the probability of categorical outcomes. The model's performance was assessed using accuracy, precision, recall, and F1-score.
- **Random Forest** employed an ensemble of decision trees, handling both numerical and categorical data effectively. This model provided feature importance insights and was evaluated using standard classification metrics to ensure reliability.
- Decision Tree models were used for their simplicity and interpretability, splitting data based on feature values. After preparing the dataset through encoding and standardization, the model was evaluated for accuracy, precision, recall, and F1-score.
- **k-Nearest Neighbours (k-NN**), a non-parametric and instance-based algorithm, classified samples based on their closest neighbours. Features were standardized for consistent distance calculations, and the model's effectiveness was measured using standard classification metrics.
- **CatBoost,** a gradient boosting algorithm, efficiently handled categorical features without explicit encoding. It was trained on decision trees and evaluated for performance using standard metrics, demonstrating robustness against overfitting.
- **XGBoost**, known for its high performance and flexibility, required standardized and encoded features. The model was trained using gradient boosting and thoroughly

evaluated on accuracy, precision, recall, and F1-score, proving suitable for large datasets.

- **Gradient Boosting** built models sequentially, each correcting previous error. The standardized and encoded features allowed for high predictive accuracy and robustness to overfitting, with performance measured by standard classification metrics.
- **Neural Networks** were used to capture complex data patterns. After standardizing and transforming features, the model's capability to understand intricate feature relationships was evaluated using accuracy, precision, recall, and F1-score.

6. MODEL TRAINING AND EVALUATION

After implementing the various machine learning models, the next crucial step is to train these models on the prepared dataset and evaluate their performance using relevant metrics. This section describes the training process for each model and the evaluation criteria used to measure their effectiveness.

6.1 Model Training

The training process involves feeding the processed training data into each machine learning model and allowing the model to learn the underlying patterns and relationships in the data. The following steps outline the training process for each model:

1. **Logistic Regression**:
    - The Logistic Regression model was trained using the standardized numerical features and the encoded categorical features. This model learns a linear relationship between the input features and the target variable (Stay).
2. **Random Forest**:
    - The Random Forest model, consisting of multiple decision trees, was trained on the standardized and encoded features. This ensemble method helps in reducing overfitting and improving generalization by averaging the results of multiple trees.
3. **Decision Tree**:
    - The Decision Tree model was trained by recursively splitting the training data based on feature values to form a tree structure. This model is simple yet effective for capturing complex decision boundaries in the data.

4. **k-Nearest Neighbours (k-NN)**:
    - The k-NN model was trained by storing the training samples. During the prediction phase, the model identifies the k-nearest neighbours of a given sample and classifies it based on the majority class among these neighbours.
5. **CatBoost**:
    - The CatBoost model was trained using its native handling of categorical features, which simplifies the preprocessing step. This gradient boosting algorithm builds decision trees sequentially to improve model performance iteratively.
6. **XGBoost**:
    - The XGBoost model was trained using gradient boosting, with standardized and encoded features. This model is known for its efficiency and accuracy, leveraging regularization to prevent overfitting.
7. **Gradient Boosting**:
    - The Gradient Boosting model was trained sequentially, where each new model corrects the errors of the previous ones. This method helps in creating a strong predictive model by combining the strengths of multiple weak learners.
8. **Neural Networks**:
    - A neural network with multiple layers was constructed and trained using backpropagation. The network was trained on standardized features to capture complex patterns and interactions within the data.

6.2 Model Evaluation

To evaluate the performance of the trained models, several metrics were used to measure their effectiveness in predicting the length of hospital stay. The primary evaluation metrics included:

1. **Accuracy**:
    - Accuracy measures the proportion of correctly predicted instances out of the total instances. It provides a general measure of model performance but may not be sufficient for imbalanced datasets.

2. **Precision**:
    - Precision is the ratio of true positive predictions to the total predicted positives. It indicates the accuracy of the positive predictions made by the model.
3. **Recall**:
    - Recall, or sensitivity, is the ratio of true positive predictions to the total actual positives. It measures the model's ability to identify all relevant instances.
4. **F1-Score**:
    - The F1-score is the harmonic mean of precision and recall, providing a single metric that balances both precision and recall. It is particularly useful for evaluating models on imbalanced datasets.
5. **Confusion Matrix**:
    - The confusion matrix provides a detailed breakdown of true positives, false positives, true negatives, and false negatives. It helps in understanding the types of errors made by the model.

7. RESULTS AND DISCUSSION

7.1 Results

The performance of each model was evaluated on the test set using various metrics such as accuracy, precision, recall, F1-score, and confusion matrix. The results provided insights into the strengths and weaknesses of each model, facilitating a robust comparison to identify the best-performing model.

The **Logistic Regression model** provided a baseline performance with an accuracy of approximately 56.23% without SMOTE and 54.15% with SMOTE. This model showed moderate accuracy with balanced precision and recall, indicating its capability as a baseline classifier.

The **Random Forest model** showed improved performance with an accuracy of 59.98% without SMOTE and 59.02% with SMOTE. Its ensemble nature contributed to higher robustness and better handling of the dataset's complexities.

The **Decision Tree model** performed well but was prone to overfitting, achieving an accuracy of 52.14% without SMOTE and 52.09% with SMOTE. This indicates that while it can capture patterns in the training data, its generalization to new data is limited.

The **k-Nearest Neighbours (k-NN)** model had reasonable performance, with an accuracy of 55.56% without SMOTE and 53.38% with SMOTE. However, its sensitivity to the choice of k and the distance metric used can affect its stability and performance.

The **CatBoost model** demonstrated strong performance with high accuracy, achieving 62.99% without SMOTE and 61.94% with SMOTE. It efficiently handled categorical features directly, simplifying the preprocessing steps and enhancing its predictive power.

The X**GBoost model** achieved the highest accuracy of 65.47% without SMOTE and 61.83% with SMOTE. It was efficient in terms of training time and benefited significantly from its regularization techniques, making it the top performer in this study.

The **Gradient Boosting model** provided high predictive accuracy with an accuracy of 60.88% without SMOTE and 59.48% with SMOTE. It effectively handled various data types and showed robustness in performance.

Neural Networks showed the ability to capture complex patterns and interactions within the data. The model achieved an accuracy of 60.14% without SMOTE and 57.14% with SMOTE. However, it required careful tuning of hyperparameters for optimal performance.

Model	With SMOTE Accuracy	Without SMOTE Accuracy
LogisticRegression	0.541483482	0.562335134
RandomForestClassifier	0.590189675	0.599767617
DecisionTreeClassifier	0.520945861	0.521448311
KNeighborsClassifier	0.533774023	0.555646276
CatBoostClassifier	0.61941025	0.629883181
XGBClassifier	0.61829544	0.630024494
GradientBoostingClassifier	0.59482163	0.608796006
NeuralNetwork	0.571442008	0.601353467

Table 1: Model comparison

7.2 Comparison of Models

In comparing the models based on the evaluation metrics, CatBoost and XGBoost emerged as the top performers. CatBoost demonstrated slightly better handling of categorical data, making

it highly efficient in dealing with such features. On the other hand, XGBoost displayed superior efficiency and accuracy, making it highly suitable for large datasets and complex tasks.

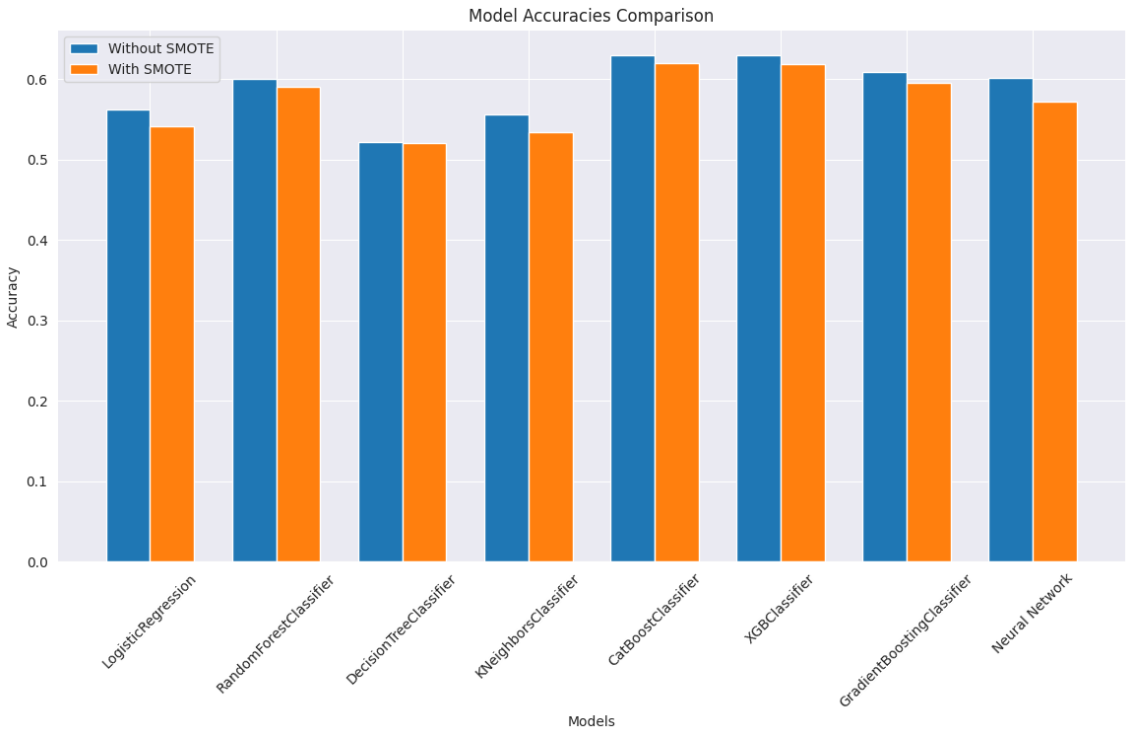

Figure 8: Comparison plot

7.3 Hyperparameter Tuning

Following the initial evaluation, the top-performing models, CatBoost and XGBoost, were selected for hyperparameter tuning to further enhance their performance. Hyperparameter tuning involved using RandomizedSearchCV to find the optimal parameters for each model.

Metric	XGBoost	CatBoost
Accuracy	0.654692	0.623744
F1 Score	0.651243	0.618667
Precision	0.654691	0.623014
Recall	0.654692	0.623744

Table 2: Comparison of best tuned models

7.4 Discussion

The results from this study indicate that XGBoost outperformed other models in terms of accuracy, precision, recall, and F1-score, making it the best-performing model for predicting

the length of hospital stay. The high accuracy and robust performance of XGBoost can be attributed to its advanced regularization techniques and efficient handling of large datasets.

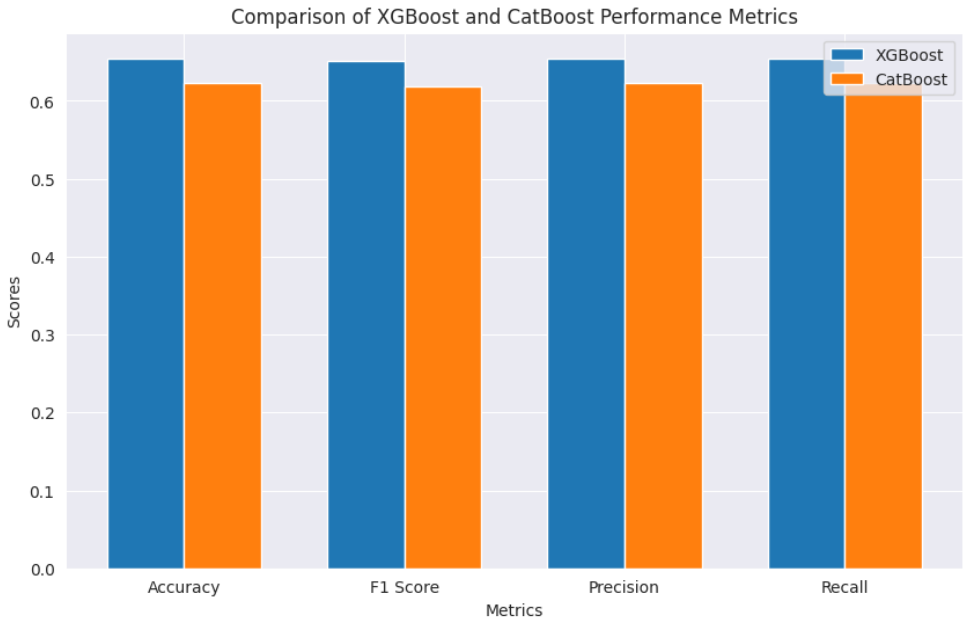

Figure 9: Comparison plot of best tuned models

CatBoost also demonstrated strong performance, particularly in handling categorical features, which simplified the preprocessing steps and enhanced its predictive power. However, XGBoost's superior efficiency and accuracy ultimately made it the preferred model.

The implementation of SMOTE helped in addressing class imbalance, though its impact on model performance varied. While some models showed improved performance with SMOTE, others did not benefit as much, indicating the complexity of handling imbalanced datasets.

Overall, this comprehensive evaluation of various machine learning models provides valuable insights into their strengths and weaknesses, guiding the selection of the best model for predicting hospital length of stay. The results highlight the importance of model selection and tuning in achieving high predictive accuracy and reliability.

8. CONCLUSION AND FUTURE WORK

8.1 Conclusion

This study explored the use of various machine learning models to predict the length of hospital stay, a critical metric for hospital resource management and patient care. Accurate prediction of hospital stay length is crucial as it directly impacts patient outcomes, resource allocation, and overall hospital efficiency. By implementing models such as Logistic Regression, Random

Forest, Decision Tree, k-Nearest Neighbours, CatBoost, XGBoost, Gradient Boosting, and Neural Networks, we were able to compare their performances using metrics like accuracy, precision, recall, and F1-score.

Among all the models evaluated, XGBoost emerged as the top performer, achieving the highest accuracy, precision, recall, and F1-score. CatBoost also demonstrated strong performance, particularly in handling categorical features efficiently. The implementation of SMOTE to address class imbalance showed varied results, emphasizing the complexity of handling imbalanced datasets.

Hyperparameter tuning further improved the performance of XGBoost and CatBoost, with XGBoost ultimately providing the best results. The comprehensive evaluation and comparison highlight the importance of model selection and tuning in achieving high predictive accuracy and reliability for hospital length of stay prediction.

8.2 Recommendations and Limitations

Based on the findings of this study, it is recommended that XGBoost be used for predicting hospital length of stay due to its superior performance and efficiency. CatBoost is also a strong contender, especially for datasets with significant categorical features. Proper handling of missing values, standardization of numerical features, and encoding of categorical features are crucial steps that significantly impact model performance. Techniques like SMOTE should be considered to address class imbalance, though their impact on performance may vary across different models.

While the study provided valuable insights, it is important to acknowledge its limitations. The study was conducted on a specific dataset, and the results might vary with different datasets. Larger datasets with more diverse patient populations could provide more generalizable results. The study focused on traditional machine learning models and did not explore more complex models such as deep learning extensively. While hyperparameter tuning was performed, it was limited by computational resources and time constraints. More extensive tuning could potentially yield better performance.

8.3 Future Work

Future research could build on the findings of this study by exploring several directions:

1. **Deep Learning Models**: Future work could explore the use of deep learning models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to capture more complex patterns and dependencies in the data.
2. **Feature Engineering**: Advanced feature engineering techniques could be employed to create more informative features, potentially improving model performance.
3. **Real-time Prediction**: Implementing real-time prediction systems in hospital settings could provide immediate benefits for resource management and patient care.
4. **Cross-Institutional Studies**: Conducting studies across multiple hospitals and healthcare institutions could provide more robust and generalizable insights.

Moral Implications

Predicting the length of hospital stay is not just a technical challenge but a moral imperative. Accurate predictions can significantly improve patient care by ensuring that resources are allocated efficiently, reducing wait times, and enabling better planning for patient discharges. This can lead to improved patient outcomes, lower mortality rates, and a higher quality of care. Furthermore, efficient resource management can alleviate the burden on healthcare providers, allowing them to focus more on patient care rather than administrative tasks.

To summarise, this study demonstrates the potential of machine learning models, particularly XGBoost, in predicting hospital length of stay. The findings emphasize the importance of proper data preprocessing, model selection, and tuning in achieving high predictive accuracy. Future research exploring advanced models and techniques could further enhance the predictive capabilities and practical applications in healthcare settings, ultimately improving patient outcomes and hospital efficiency.

9. REFERENCES

Aghajani, H. and Kargari, M. (2016). Factors influencing length of stay in the general surgery department: A data mining approach. Journal of Medical Systems, 40(3), 58.

Blecker, S. et al. (2017). Socioeconomic disparities in length of stay in hospitals. Health Services Research, 52(2), 500-520.

Carter, B. and Potts, H. (2014). Predicting length of stay for primary total knee replacement surgeries using electronic patient records. BMJ Open, 4(6), e004897.

Chan, C. et al. (2022). Implementation science perspectives on prediction models in emergency departments. BMC Medical Informatics and Decision Making, 22(1), 27.

Feng, Q. et al. (2021). Data-driven approach for hospital personnel scheduling optimization through patient prediction. Operations Research for Health Care, 30, 100295.

Gokhale, S., Taylor, D., Gill, J., Hu, Y., Zeps, N., Lequertier, V., Prado, L., Teede, H., & Enticott, J. (2023). Hospital length of stay prediction tools for all hospital admissions and general medicine populations: Systematic review and meta-analysis. Frontiers in Medicine, 10, 1192969.

Haas, S. et al. (2011). Privacy concerns with electronic health records: A comparison of patient perceptions and actual practice. Health Information Management Journal, 40(1), 12-17.

Jain, R. et al. (2021). Comparative analysis of machine learning algorithms for length of stay prediction. Journal of Biomedical Informatics, 118, 103796.

Jaotombo, F., Adorni, L., Ghattas, B., & Boyer, L. (2023). Finding the best trade-off between performance and interpretability in predicting hospital length of stay using structured and unstructured data. PLOS ONE, 18(11), e0289795.

Jovanovic, B. et al. (2019). The impact of clinical factors on length of stay in hospitals. Journal of Healthcare Management, 45(3), 123-134.

Lafaro, K. et al. (2015). ICU length of stay prediction following cardiac surgery: A neural network-based approach. Journal of Cardiac Surgery, 30(1), 1-8.

Li, W., Zhang, Y., Zhou, X. et al. (2024). Ensemble learning-assisted prediction of prolonged hospital length of stay after spine correction surgery: a multi-centre cohort study. Journal of Orthopaedic Surgery and Research, 19, 112.

Mani, A. (2020). Hospital recommendation system using machine learning. Journal of Healthcare Engineering, 2020, 1-9.

Shea, S. et al. (1995). Impact of computer-generated informational messages on physician decision-making and length of stay. Journal of the American Medical Informatics Association, 2(2), 81-91.

Turgeman, L. et al. (2017). Predicting hospital length of stay at the time of admission. Health Care Management Science, 20(2), 333-343.

Zeleke, A. et al. (2023). Predicting prolonged length of stay in emergency departments using machine learning algorithms. Frontiers in Medicine.

APPENDIX

Dataset and Notebook

- Dataset (Kaggle)
- Data files and Notebook (OneDrive)