Predicting a patient's length of stay

Group 20: Project 10 (Project Sponsor: Wrightington, Wigan & Leigh)
This Developing a predictive machine learning model to accurately estimate a patient's length of stay to facilitate informed decision-making regarding hospital bed resources

Ali Alqahtani

School of Computing & Communications, Lancaster University, a.m.alqahtani@lancaster.ac.uk

Jin Cheng Chew

School of Computing & Communications, Lancaster University j.c.chew@lancaster.ac.uk

Sofiia Denysiuk

School of Computing & Communications, Lancaster University, s.denysiuk@lancaster.ac.uk

Xiaomei Ge

School of Computing & Communications, Lancaster University, a.m.alqahtani@lancaster.ac.uk

Issa Ndungu

School of Computing & Communications, Lancaster University, i.m.ndungu@lancaster.ac.uk

1 INTRODUCTION

The National Health Service (NHS) in England faces significant challenges in managing hospital bed resources nationwide. The ever-increasing demand for bed resources has been widely documented over the last 35 years, underlined by the worrying combination of a steadily increasing number of hospital admissions from 11.1 million in 2000/01 to 18.4 million in 2024/25 [1], coupled with steadily decreasing average available bed numbers in hospitals from approx. 299,000 in 1987/88 to 131,875 in 2024/25 [2]. As a result, hospitals across the country are constantly operating near, at, or beyond capacity, with bed occupancy rates consistently exceeding the 85% safety threshold [3]. This is clearly unsustainable, as spare capacity is essential to accommodate variations in demand driven by external factors such as area demographics, seasonal changes, natural disasters, pandemics/epidemics, among others [4] [5] [6] [7] [8]. Furthermore, when comparing the UK with other members of the Organization for Economic Cooperation and Development (OECD), we see a significantly lower number of hospital beds per 1000 inhabitants than the average (2.4 vs 4.6). This issue is compounded by the overall decrease in bed stocks across NHS England hospitals, seeing the total number of beds decreasing by 8.3% between 2010/11 and 2019/20.

The Wrightington, Wigan and Leigh Teaching Hospitals NHS Foundation Trust, a group of hospitals in the Greater Manchester area, exemplifies some of the operational challenges faced by hospitals nationwide. Bed managers have the tough responsibility of making crucial decisions on how to allocate these limited resources, often relying on insight gained through years of experience, personal initiative and forecasting data. This approach is unsustainable mainly due to the largely subjective and error-prone nature of human decision-making under load [9], which could lead to suboptimal resource allocation, bias [10], and increased pressure on healthcare staff. Therefore, there is an equally pressing need to develop more proactive bed management strategies to help alleviate departmental pressures and allow hospitals to better cope with expected high demand.

1.1 Motivation

Significant advancements in artificial intelligence, machine learning, and data science allow for the development of robust, data-driven solutions that leverage existing patient data to support informed decision-making and the efficient management of hospital bed resources. One key metric this can be applied to is a patient's length of stay (LoS), a seemingly crucial indicator for effective bed resource allocation. Accurate prediction of a patient's length of stay is valuable as it gives bed managers better insight into how long beds are likely to be occupied, which they can then use to inform future resource allocation decisions. The downstream effects of this are improved patient flow, operational efficiency, reduced waiting times, and ultimately the enhanced ability to provide a better standard of care. With all this in mind, the aim is to develop a program that preprocesses patient data, extracts the most predictive features, and feeds them into a bespoke machine learning model to predict patient length of stay, thereby supporting informed decision-making and the efficient management of hospital bed resources.

1.2 Project Objectives & Research Questions

1 To understand the major trends in data such as demographics, diagnoses with the length of stay.
- What are the data types, distributions & correlations of the variables in the dataset?
- Are there any significant data quality issues in the dataset and what are the techniques to address them?

2 To develop, evaluate and select the most effective model for predicting a patient's length of stay.
- What is the most suitable evaluation criteria to measure model performance in a hospital setting?
- Which model best meets the defined criteria?

3 To identify and study the key features that significantly contribute to the prediction of the length of stay.
- Which factors contribute the most to predicting the length of stay in each model?
- What relationships (i.e., positive or negative) exist between our independent and dependent variables?

4 To deliver practical recommendations for the bed management team based on the findings and to improve the data collection framework.
- How can the chosen model output be used or applied in the hospital workflow?
- What are the specific actions that the bed management team can take based on the results?
- What are the additional data points that are currently not collected and would improve the model performance by doing so?

2 METHODOLOGY

2.1 Research Strategy

We adopted a Comparative Modeling Strategy grounded in supervised machine learning principles. This approach involves developing multiple predictive models using different algorithms and comparing their performance using consistent evaluation metrics [11]. This strategy is particularly appropriate for healthcare prediction tasks where model interpretability, generalization capability, and robust performance across diverse patient subgroups are critical considerations [12].

The research followed an iterative, agile-inspired workflow with regular team meetings every 2-3 days, incorporating stand-ups and clearly defined task distributions tracked using a Gantt chart. A GitHub repository was used for the version control and code sharing.

2.2 Data Initial Exploration

After the Initial Exploration, it was discovered that the dataset contained 41846 patient records with 101 features encompassing demographic information, clinical indicators, admission circumstances, and temporal variables. Initial exploration has also revealed revealed:

- Target Variable: The spell_los_hrs (LoS - Length of Stay) was identified as the target variable. It is a continuous variable representing the number of hours a patient remained hospitalized.
- Feature Diversity: The dataset contained a mixture of numerical continuous variables, categorical variables with varying cardinality (ranging from binary indicators to features with over 5,000 unique categories), and temporal features.
- Missing Data Patterns: Several features had more than 70% missing values.

2.3 Exploratory Data Analysis (EDA) and Data Cleaning

Each team member conducted EDA on the assigned feature subsets following a standard procedure:

1. Assessing Initial Data Assessment: 20 features had more than 70% missing values, and some showed systematic missingness patterns correlated with admission type (e.g., emergency vs. planned admissions).
2. Describing: Computed descriptive statistics for the numerical variables in the dataset. The LoS distribution was substantially positively skewed, with extreme outliers (extended hospital stays). Analysis of the distribution indicated that approximately 1% of cases represented extreme values, which may potentially distort model training. For the categorical variables, the unique values and their frequencies were counted, and rare or dominant categories were identified.

Key findings from EDA were then used to proceed with the data cleaning, where we performed the following:

- Handling missing values: Features with >70% missing values were considered to be removed if they were weakly correlated with our target variable, and if missingness could not be reliably addressed through as imputation methods. For the categorical feature NEWS2, which has substantial missingness but contains important context, we created binary "missing" indicator flags before imputation. This approach captures potential informative missingness patterns, particularly relevant in emergency admission scenarios where certain data may be systematically unavailable [13].
- Removing duplicates: no duplicated rows were identified.

2.4 Data Preprocessing

2.4.1 Handling Categorical Variables

High-Cardinality Features (>50 categories): Three categorical features had extremely high cardinality (1400+, 2600+, 2700+ unique categories). We employed CatBoost Encoding in conjunction with k-folds to mitigate the risk of data leakage [14]. It is effective as it prevents the classic dimensionality explosion associated with methods such as One-Hot Encoding, whilst also mitigating against overfitting [15].

Medium-Cardinality Features (5-50 categories): For features in this range, we applied Label Encoding. Introducing false ordinality is a risk with this method, but research demonstrates that tree-based models are more

robust to this, since they partition the feature space using threshold splits rather than assuming linearity between variables [16] [17], making it a suitable option.

Low-Cardinality Features (<10 categories): For features with few categories, we implemented One-Hot Encoding, creating binary flag variables for each category. It is effective for low-cardinality features, as it avoids imposing any ordinal structure, and has demonstrated strong performance across various ML models [18].

2.4.2 Data Leakage Prevention

To help ensure model validity and prevent overly inflated model performance, we thoroughly reviewed, identified, and removed all features that were seen to contain information that would not be available to the model at the time of prediction. These included features such as those representing patient outcomes, discharge information, or post-admission events. This prevents data leakage and ensures that the model is trained on genuine predictive features [19].

2.4.3 Target Variable Transformation

After visualizing our target variable using histograms (Figure1), we found its distribution to be highly right-skewed and dispersed (std:135, max: 3022). There are 95% patients staying from 0 to 10 days, while other 5% is from 30 days to 125 days. Such extreme skewness and dispersion can bias model training, making it difficult to learn stable patterns and leading to poor generalization in rare but long-stay cases. To control this, clipping the target variable by capping all values above its 95th percentile to that percentile value (627) to handle the extreme cases, and then log1p transformation applied to the target variable. After clipping and transformation, it can be seen from Figure1 that distribution of the target variable becomes more concentrated and substantially less right-skewed. Furthermore, research shows that 'outlier' patients with exceedingly long lengths of stay consume a disproportionate share of hospital bed resources, so it may be sensible to treat them as an entirely separate sub-population in future model design considerations.

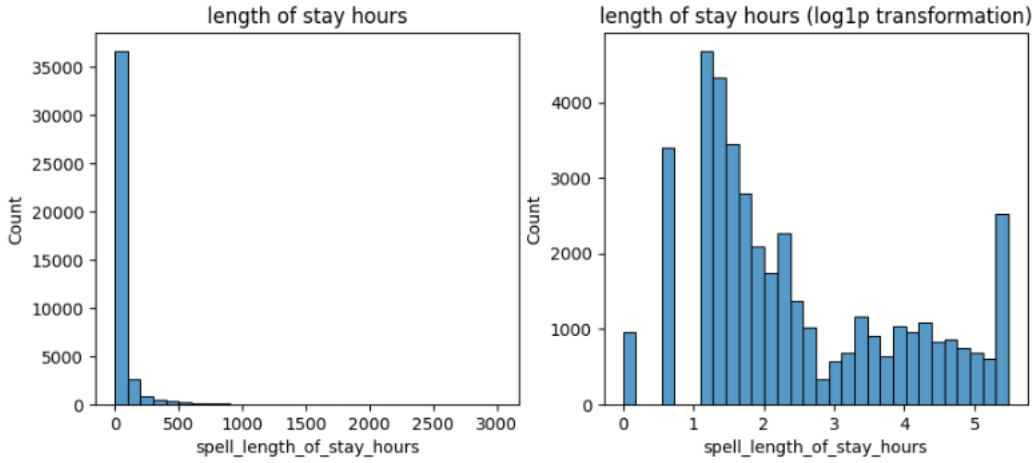

Figure 1: LoS distribution before and after the transformation

2.4.4 Feature Engineering

Beyond data cleaning, we created derived features to capture extra information, that can be derived:

1. Emergency Admission Flag: A binary indicator inferred from missing NEWS2 patterns. Emergency cases typically have longer, less predictable stays [20], and this feature showed strong predictive value.
2. Season of Admission: Derived from admission dates to capture known seasonal pressures on hospitals, especially winter-related increases in respiratory admissions [21].
3. Future Feature Opportunities: Documented options such as interaction terms between high-importance variables and finer temporal features (e.g., day of week, holiday periods).
4. Metadata Flags for Encoded Variables: For each CatBoost-encoded category, we added {var}_frequency and {var}_is_rare to capture category prevalence and improve model performance (source). New features were created after one hot encoding as well.

2.5 Feature Selection

After feature engineering and data cleaning, there are 94 features. Using all the 94 features for training may introduce unnecessary computational cost and noise to the training. Therefore, a structured feature selection pipeline was implemented to select the most relevant predictors.

1. Correlation-based pruning: The first step removes the correlated features to minimize the redundancy and avoid multicollinearity [22]. As two features that carry the same information may bring biases to model training, applying a correlation threshold of 0.7 reduced the number of features from 94 to 71. This gets to ensure a cleaner and more independent feature set.
2. Random Forest feature selection: The next step is to evaluate the predictive contribution of each remaining feature using a RandomForest (RF) model. RH is chosen because its tree-based structure allows it to identify non-linear relationships and complex interactions. As the final selected predictive models are also tree-based, using the RF model to estimate the feature importances ensures better alignment with the model behavior [23]. RF provides built-in measures of feature importance, and 15 features with above-average importance scores were selected, indicating that they have significant effects on a patient's length of stay. The selected features and corresponding importance scores can be found in Table 1

Table 1: The top 15 important features

Search Space	Feature importance
IP_admission	0.263797
spell_primary_diagnosis_description	0.149847
Admission_Hour	0.060558
NEWS2_missing	0.056771
spell_dominant_proc_count	0.050992
discharge_letter_status	0.049863
spell_dominant_proc_description	0.047107
patient_age_on_admission	0.044203
spell_secondary_diagnosis_encoded	0.032226
ward_type_admission	0.023686
hrg_sub_group_encoded	0.021119
comorbidity_score	0.021061
ward_code_admission_2	0.019896
spec_direc_Womens_Health	0.014468
general_medical_practice_desc	0.014156

2.6 Model Development

2.6.1 Model Selection Reasoning

We selected Random Forest and XGBoost because they align well with the dataset's mixed characteristics, including numerous categorical features—with some having high cardinality—and multiple binary variables. Both models handle heterogeneous data and missing values with minimal preprocessing while capturing complex nonlinear relationships important for predicting LOS. Random Forest offers strong robustness, reduced overfitting through bagging, and clear feature importance insights. While XGBoost provides state-of-the-art performance, built-in regularization, and efficient handling of sparse and high-cardinality categorical data.

2.6.2 Evaluation Metrics

Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) were main metrics used for the model evaluation. MAE and RMSE are scale-dependent, so their absolute values can be hard to interpret. To account for the scale of the target variable, normalized errors were also reported which is obtained using MAE/RMSE divided by the mean of target variable. Additionally, MAE/ RMSE of the selected best model have compared with simple baseline predictors that always predict the mean or median LOS. This allows us to verify that the model performs better than naive, non-learning strategies.

Some plots were used to observe predicting outcomes, such as QQ plot of residuals, residual histogram, and predicted VS actual values plot.

2.6.3 Cross-Validation Strategy

A shuffled 5-fold cross-validation was implemented in the hyperparameter tuning phase to ensure that the output is reliable and prevent model overfitting [24]. The dataset is shuffled and divided into five folds, with four folds used for training and the remaining one-fold used for validation in each iteration. This reduces the risk of the model being biased toward a particular train-test split and the shuffled data also ensures that each fold will be a random and representative sample of the data [25], resulting in a more comprehensive performance evaluation. By integrating this shuffled 5-fold cross-validation technique into the Optuna-based hyperparameter tuning, the model is validated five times with different folds of the data. The average Root Mean Squared Error (RMSE) across five folds is then used as the metrics for evaluating the performance of a model with different hyperparameters value, ensuring that the selected hyperparameters are based on consistent generalization performance.

2.6.4 Hyperparameter Tuning (HT)

The Optuna library was used for tuning hyperparameter. The default search algorithm in Optuna is Tree-structured Parzen Estimator (TPE). TPE is an algorithm based on the Bayesian optimization, based on which TPE dynamically adapts its sampling distribution according to historical trial results and effectively identify a better hyperparameter combination within specified trials (300 trials in this report).

The Table 2 and Table 3 shows the selected hyperparameters of XGBoost and RandomForest for tuning. During the tuning process, the Optuna chose the best model and hyperparameter combination according to minimum MAE. XGBoost performs better than RandomForest and the best hyperparameters can be found in Table 2.

Table 2: Hyperparameters tuned for XGBoost

Hyperparameters	Search Space	Description	Best Hyperparameters
eta	0.01 – 0.1 (float)	Learning rate	0.0999
max_depth	2 – 4 (int)	Maximum tree depth	4
min_child_weight	3 – 15 (int)	Minimum sum of instance weight in a child	7
gamma	0.0 – 10.0 (float)	Minimum loss reduction required for a split	0.0078
subsample	0.7 – 1.0 (float)	Row subsampling ratio	0.9073
colsample_bytree	0.7 – 1.0 (float)	Column subsampling ratio	0.4625
reg_lambda	0.1 – 10 (log-uniform)	L2 regularization term	0.4625
reg_alpha	0.1 – 10 (log-uniform)	L1 regularization term	5.6223
n_estimators	300 – 600 (int)	Number of boosting rounds	551

Table 3: Hyperparameters tuned for RandomForest

Hyperparameter	Search Space	Description
n_estimators	300 – 600 (int)	Number of trees
max_depth_rf	3 – 5 (int)	Maximum tree depth
min_samples_split	2 – 10 (int)	Minimum number of samples required to split a node
min_samples_leaf	1 – 5 (int)	Minimum number of samples required at a leaf node
max_features	{"sqrt", "log2"}	Number of features considered at each split
bootstrap	{True, False}	Whether bootstrap sampling is used

3 RESULTS

3.1 Model Performance Evaluation

In the train data set, the MAE and RMSE of the best model are 17.7 and 40.4, respectively. While MAE and RMSE coming from the test data set are 20.3 and 46.0 respectively. This means it would have around 20-hours error of length of stay (LOS) if the model is used to predict each patient's LOS. A normalized prediction error is

a 20-hours error accounted for 55% of mean LOS of a patient (36h). 55% is not small compared to a generally acceptable range (20%-30%). However, the std. of LOS is way larger than mean LOS, showing distribution of LOS is heavy-tailed and huge fluctuations, so, from this perspective, 20h prediction error is not abysmal. Two naive baseline predictors that always predict mean and median LOS are created. Compared with the two simple baseline predictors, the model's MAE and RMSE decreased 39.8% and 35.2%. This means the model has captured some patterns of the data set.

Another noticeable discrepancy between MAE and RMSE is vast (around 26h). RMSE is sensitive to outliers; the vast discrepancy indicates that some highly inaccurate predictions exist. Figure2 and Figure3 provide evidence for this. It can be seen from the Figure3 some points are diverged far from the red line, indicating that residuals of some predictions are enormous (reached 300h). Moreover, Figure2 exhibits a funnel-shaped pattern and suggests that the longer the LOS, the less accurate the prediction. From Figure4, the residuals exhibit clear non-normality and heavy tails, which is consistent with the heteroscedastic residual pattern observed in the Figure2&3.

In a nutshell, the chosen best model can capture some patterns of data set, but prediction can be very inaccurate when LOS is long. The model provides reasonably reliable predictions for cases with LOS below 100 hours. Although 5% extreme LOS values (e.g. above 3000h) have been capped to 95th percentile value (239h); the LOS is still quite discrete. This attribute of target variable brings large challenges for model prediction.

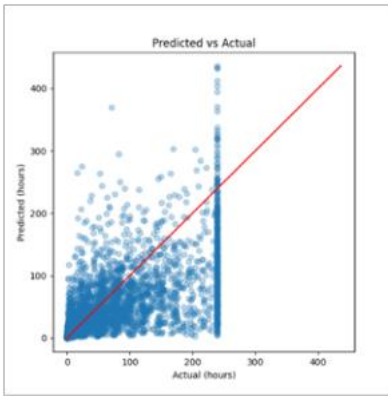

Figure 2: Predicted vs Actual

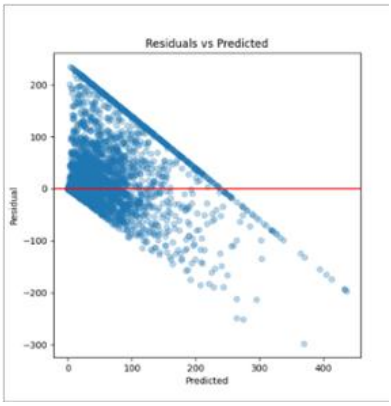

Figure 3: Residuals vs Predicted

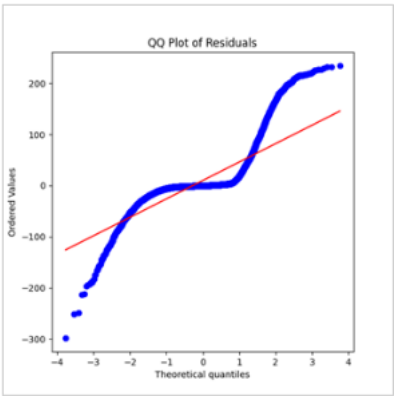

Figure 4: QQ plot Residuals

3.2 Addressing the Research Questions

Our research questions can mainly be grouped into 4 groups, as shown below:

1. Understanding the data trends: The dataset has a combination of mixed numerical, binary and categorical variables with uneven distributions. Therefore, different rules were established such as applying different encoding method based on cardinality, and cleaning highly correlated, irrelevant, or leakage-prone features. Feature selection is also used to narrow down the most meaningful predictors. Throughout the analysis, EDA showed weak patterns with no single factor dominating LOS, highlighting the complexity of the problem.

2. Model development and evaluation: RMSE and MAE were selected because they fit into hospital length-of-stay prediction, where the target is continuous and shows irregular patterns [26]. Following the cleaning, the data still contain outliers and skewed values, so these metrics help to describe the operational impact of prediction errors. MAE offers a simple quantification of the mean error, while RMSE puts a little more emphasis on larger errors [27]. Both provide a balanced view of model accuracy. Based on these measures, XGBoost demonstrated the best and most reliable overall performance for predicting patient length of stay.

3. Identifying key predictors: This analysis revealed the most significant predictors for patient stay length predictions through feature selection and feature importance. The five most significant predictors are: IP_admission, spell_primary_diagnosis_description, Admission_Hour, NEWS2missing and

spell_dominant_proc_count. The correlation analysis indicated that majority of the top features are positively related to LOS and NEWS2missing have negative relationship. These results mean that the variables are the most influential in predicting early LOS and they can be used to explore insights in operational planning.

4. Recommendations: By examining the feature of importance from selected features, it provides early insights for bed management. A more direct prediction of a patient's length of stay can enable the staff to predict better for the bed demand considering patients with high probability of longer stay. Additionally, the dataset mainly contains short-stay cases and thereby limited more accurate long-stay prediction. Therefore, the further improvement focused in collecting more long-stay data is recommended to strengthen future modelling and improve hospital resource planning.

4 PAPERS EVALUATION AND ANALYSIS OF THE RESULTS

4.1 Implications of Findings

Going beyond the model accuracy, the top predictors point out a number of operational factors associated with longer stay. The high importance of IP_admission suggests that full inpatient admissions can be associated with longer care episodes compared to day cases. Similarly, the high score of spell_primary_diagnosis_description suggests that the underlying clinical condition is a major factor in determining bed demand and predicting patient's stay length. Admission_Hour is also significantly influencing, meaning that the arrival timing may affect patient flow and downstream delays, thereby extending the patient's stay length. Although these trends were not clinically proven, they are early indicators that can assist bed managers to determine groups with which they need more attention or plan ahead.

4.2 Potential Biases and Limitations

The model's errors mainly occur in predicting patient's with long stay period. For instance, the model shows the sign of under or overestimates on longer patient's stay length because the data has too minimal instances presented in the dataset. By referring to the model evaluation result, it suggests that the model lacks sufficient examples of long-term patient stay length and therefore cannot fully capture the unexpected pattern of an extended patient stay.

4.3 Defence of validity:

The validity of the study is supported by the fact that the all the used methods are justified with reputable academic and industry sources, not the opinion-based forums. This guarantees that the modelling decisions, preprocessing decisions and evaluation processes are in line with the best practice, and it enhances the reliability and accuracy of the research output.

5 CONCLUSION

This project compared models Random Forest and XGBoost ML models for predicting patient length of stay. After full preprocessing, feature engineering, and cross-validated tuning, XGBoost perfomance was the best overall, reducing error substantially compared to baseline predictors. Key predictors such as admission type, primary diagnosis, and admission hour were the most influential.

However, the following limitations remain: a single model cannot accurately predict both short- and long-stay patients, as these groups behave fundamentally differently, data quality issues and limited long-term stay data.

Future work should explore alternative algorithms, refine the feature engineering procedure, and especially, separate modelling strategies for different patient groups, alongside collecting more long-stay cases. Data on metabolic disorders and patient vitals, such as blood pressure, pulse rate, and oxygen saturation, can be added to improve model performance, as they contain significant medical information that can inform model predictions [].

Overall, we achieved our goal of implementing and comparing two models and producing reliable results. The chosen approach was appropriate, and the findings provide a solid basis for future refinement.

REFERENCES

[1] Statista, *"Hospitals in the United Kingdom: Statista dossier,"* 2025. [Online]. Available: https://www.statista.com/study/26236/hospitals-and-hospital-departments-in-the-united-kingdom-statista-dossier. Accessed: Dec. 11, 2025.

[2] L. Ewbank, J. Thompson, H. McKenna, S. Anandaciva, and D. Ward, *"NHS hospital bed numbers: Past, present, future,"* The King's Fund, Nov. 5, 2021. [Online]. Available: https://www.kingsfund.org.uk/insight-and-analysis/long-reads/nhs-hospital-bed-numbers?utm_source=chatgpt.com. Accessed: Dec. 11, 2025.

[3] National Guideline Centre (UK), *"Chapter 39: Bed occupancy,"* in *Emergency and Acute Medical Care in Over-16s: Service Delivery and Organisation (NICE Guideline 94)*. London, U.K.: National Institute for Health and Care Excellence, 2018. [Online]. Available: https://www.nice.org.uk/guidance/ng94/evidence/39.bed-occupancy-pdf-172397464704. Accessed: Dec. 11, 2025.

[4] Nuffield Trust, *"Hospital bed occupancy,"* Nuffield Trust, Resource. [Online]. Available: https://www.nuffieldtrust.org.uk/resource/hospital-bed-occupancy. Accessed: Dec. 9, 2025.

[5] M. K. Hasan, S. M. Nasrullah, A. Quattrocchi, P. Arcos González, and R. Castro-Delgado, *"Hospital surge capacity preparedness in disasters and emergencies: a systematic review," Public Health*, online ahead of print, 2023.

[6] K. Suhail and R. Cochrane, *"Seasonal variations in hospital admissions for affective disorders by gender and ethnicity," Social Psychiatry and Psychiatric Epidemiology*, vol. 33, no. 5, pp. 211–217, May 1998, doi: 10.1007/s001270050045.

[7] K. J. Fong, C. Summers, and T. M. Cook, *"NHS hospital capacity during COVID-19: overstretched staff, space, systems, and stuff," BMJ*, vol. 385, p. e075613, 2024, doi: 10.1136/bmj-2023-075613.

[8] A. Koch, B. J. Jones, and C. A. Perez, *"Explaining variation in hospital admission rates between general practices: cross-sectional study," BMJ*, vol. 319, pp. 98–102, 1999, doi: 10.1136/bmj.319.7202.98.

[9] E. C. Tacker and M. T. Silvia, *"Decision making in complex environments under conditions of high cognitive loading: A personal expert systems approach," Expert Systems with Applications*, vol. 2, no. 2/3, pp. 121–127, 1991.

[10] A. D. Banasiewicz, *Evidence-Based Decision-Making: How to Leverage Available Data & Avoid Cognitive Biases*. New York, NY, USA: Routledge, 2019.

[11] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer.

[12] Shillan, D., Sterne, J. A., Champneys, A., & Gibbison, B. (2019). Use of machine learning to analyse routinely collected intensive care unit data: A systematic review. Critical Care, 23(1), 284.

[13] Little, R. J., & Rubin, D. B. (2019). Statistical analysis with missing data (3rd ed.). Wiley.

[14] J. Olusegun, "Handling Categorical Variables in Ensemble Algorithms," *ResearchGate*, Feb. 2025.

[15] Micci-Barreca, D. (2001). A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems. ACM SIGKDD Explorations Newsletter, 3(1), 27-32.

[16] Hancock, J. T., & Khoshgoftaar, T. M. (2020). CatBoost for big data: An interdisciplinary review. Journal of Big Data, 7(1), 94.

[17] M. Kuhn and K. Johnson, Applied Predictive Modeling. New York, NY, USA: Springer, 2013.

[18] Potdar, K., Pardawala, T. S., & Pai, C. D. (2017). A comparative study of categorical variable encoding techniques for neural network classifiers. International Journal of Computer Applications, 175(4), 7-9.

[19] Kaufman, S., Rosset, S., Perlich, C., & Stitelman, O. (2012). Leakage in data mining: Formulation, detection, and avoidance. ACM Transactions on Knowledge Discovery from Data, 6(4), 1-21.

[20] Chaou, C. H., Chen, H. H., Chang, S. H., Tang, P., Pan, S. L., & Yen, A. M. (2017). Predicting length of stay among patients discharged from the emergency department—Using an accelerated failure time model. PloS One, 12(1), e0165756

[21] Rosychuk, R. J., Klassen, T. P., Voaklander, D. C., Senthilselvan, A., Rowe, B. H., & Bulloch, A. G. (2011). Seasonality patterns in croup presentations to emergency departments in Alberta, Canada: A time series analysis. Pediatric Emergency Care, 27(4), 256-260

[22] V. K. Jha, "Multicollinearity in Data," GeeksforGeeks, Aug. 7, 2025. https://www.geeksforgeeks.org/machine-learning-multicollinearity-in-data/

[23] A. Orlenko and J. H. Moore, "A comparison of methods for interpreting random forest models of genetic association in the presence of non-additive interactions," BioData Mining, vol. 14, art. no. 9, 2021.

[24] D. Wilimitis and C. G. Walsh, "Practical Considerations and Applied Examples of Cross-Validation for Model Development and Evaluation in Health Care: Tutorial," JMIR AI, vol. 2, no. 1, e49023, Dec. 18, 2023. https://ai.jmir.org/2023/1/e49023 doi:10.2196/49023

[25] V. Chugani, "A Comprehensive Guide to K-Fold Cross Validation," DataCamp, Jun. 21, 2024. https://www.datacamp.com/tutorial/k-fold-cross-validatio

[26] G. Almeida, F. B. Correia, A. R. Borges, and J. Bernardino, "Hospital Length-of-Stay Prediction Using Machine Learning Algorithms—A Literature Review," Applied Sciences, vol. 14, no. 22, Art. no. 10523, Nov. 2024. https://www.mdpi.com/2076-3417/14/22/10523

[27] M. Harris, "Understanding RMSE, MSE, and MAE," StatsWithR, Sep. 20, 2024.

https://www.statswithr.com/foundational-statistics/understanding-rmse-mse-and-mae

Contributions Sheet

Group ID : 20

Project Title : Predicting a patients' length of stay

Project Host : Wrightington, Wigan & Leigh

Group Member Contributions

Student Name	Student ID	Contribution/ Role Description	% Contribution
Xiaomei Ge	36973303	1. EDA (20 columns) 2. Feature Engineering (target variable) 3. Model building and evaluation (tune hyperparameters, model comparison) 4. Data leakage Prevention (identification & removal) 5. Build model evaluation criteria 6. Writing report (Model evaluation, Hyperparameter tuning and Evaluation metric)	22.5%
Chew Jin Cheng	37041332	1. EDA (20 columns) 2. Data Leakage prevention (removal) 3. Feature Engineering (Categorical handling) 4. Feature Selection 5. Model building and evaluation (tune hyperparameters, model comparison) 6. Report writing (Hyperparameter tuning, cross validation, addressing research questions and evaluations and analysis of the results)	20%
Issa Ndungu	37022498	1. Project Spokesperson 2. EDA (20 columns) 3. Data Leakage Prevention (removal) 4. Feature Engineering 5. Feature Selection 6. Video editing 7. Report writing (Introduction, Motivations & Data Pre-processing)	20%
Sofiia Denysiuk	36956514	1. EDA (20 columns) 2. Codebase for merging the EDA 3. Model Building 4. Data Leakage Prevention (removal) 5. Model Evaluation (evaluation criteria, hyperparameter tuning) 6. Report writing (Research Strategy, Data Exploration, Exploratory Data Analysis (EDA) and Data Cleaning, Feature Engineering, Conclusion) Report formatting, references maintenance 7. Presentation Template Design	22.5%
Ali Alqahtani	36975589	1. EDA (21 columns) 2. Data Cleaning 3. Features Engineering 4. Features Selection 5. Report Writing (Conclusion)	15%

Confirmation of Agreement

Student ID	Signature	Date

36973303	**Xiaomei Ge**	12/12/2025
36956514	**Sofiia Denysiuk**	12/12/2025
37022498	**Issa Ndungu**	12/12/2025
37041332	Chew Jin Cheng	12/12/2025
36975589	Ali Alqahtani	12/12 /2025