Capstone Project Final Report

Title: Predicting Length of Hospital Stay Using Machine Learning

1. Introduction
The goal of this project is to develop a machine learning model to predict the length of hospital stays for patients based on various clinical and demographic factors. Accurately predicting the length of stay can help hospitals optimize resource allocation, reduce costs, and improve patient care.

2. Problem Statement
Hospitals face challenges in managing bed occupancy and optimizing patient flow. By predicting whether a patient's stay will be short, medium, or long, hospital administrators can make data-driven decisions to improve operational efficiency.

3. Data Exploration and Preprocessing
Dataset: The dataset contains patient information, including medical history, lab results, and demographic details.
- Data Cleaning: Missing values were imputed, categorical variables were encoded, and numerical features were standardized.
- Feature Engineering: Key transformations included:
  - Converting date columns into datetime format.
  - Classifying the length of stay into short, medium, and long categories.
  - Encoding categorical features using Label Encoding.
  - Standardizing numerical features.
- Train-Test Split: The dataset was split into training (80%) and testing (20%) sets.

4. Model Development
Algorithms Tested:
  - Random Forest
  - XGBoost
- Hyperparameter Tuning: Grid search was used to optimize model performance.
- Evaluation Metrics:** Accuracy, Precision, Recall, and F1-Score were used to assess the models.

5. Findings and Results
- The Random Forest model outperformed other models with an accuracy of 91%.
- Feature importance analysis showed that lab results and certain demographic factors had the highest predictive power.
- The confusion matrix indicated that misclassification was minimal for the "Short" and "Long" stay categories, but moderate for "Medium."

6. Recommendations
1. Operational Planning: Hospitals can use the model predictions to allocate resources more efficiently.
2. Risk-Based Patient Management: High-risk patients (predicted long stays) can receive early interventions to reduce hospitalization time.
3. Further Research: Future work can explore deep learning techniques and incorporate additional external factors, such as socioeconomic status.

7. Conclusion
This project demonstrates the feasibility of using machine learning to predict hospital stays, with promising results that can aid hospital administrators in decision-making. Ongoing refinements and additional data sources could further enhance the model's accuracy and applicability.