# 

# CBIO313: Data Mining and Machine Learning-2025SPRG

# Final Machine Learning Project Report

Title: Hospital Stay Duration Prediction

Prof: Mohamed Mahmoud ElSayeh

TA: Malak Mohamed AbdelMonsef

Student’s name: Ahd Adel Shanhoury

ID:221001731

## Overview

This machine learning project predicts the category of a patient's hospital stay duration (Short, Medium, Long) based on patient and admission data from the `modified\_healthcare\_dataset.csv`. We implemented a full pipeline from data cleaning and EDA to model training, evaluation, and web deployment.

## Dataset Summary

• Dataset: modified\_healthcare\_dataset.csv  
• Size: 55,500 rows, 16 columns  
  
Key Features:  
• Age, Gender, Medical Condition, Admission Type  
• Billing Amount, Test Results, Insurance Provider  
• Target Variable: Length of Stay (converted to categorical for classification)

## Problem Statement

Can we predict a patient’s stay category (Short ≤3, Medium 4-10, Long >10 days) using admission and health-related features?

## Data Cleaning & Preprocessing

• No missing values  
• Converted `Date of Admission` and `Discharge Date` to datetime  
• Created new feature: `Stay Category`  
• Label encoding for categorical variables: Medical Condition, Admission Type, Test Results  
• Scaled numeric features for models like SVM and XGBoost

## Exploratory Data Analysis

• Age, Billing Amount, and Length of Stay are right-skewed  
• Emergency admissions and certain medical conditions are associated with longer stays  
• Feature relationships visualized via boxplots and correlation heatmaps

## Feature Engineering & Selection

• Created `Stay Category` as target (Short, Medium, Long)  
• Selected predictors: Age, Billing Amount, Encoded Medical Condition, Admission Type, and Test Results  
• Used RFE and feature importance from Random Forest to validate selection

## Model Training & Evaluation

Tested and compared multiple classification models:  
\* Logistic Regression (baseline)  
\* Random Forest  
\* SVM (Linear & RBF)  
\* XGBoost (best performer)  
  
Metrics used:  
\* Accuracy, Precision, Recall, F1 Score  
\* Confusion Matrix  
\* ROC-AUC for multiclass (One-vs-Rest)

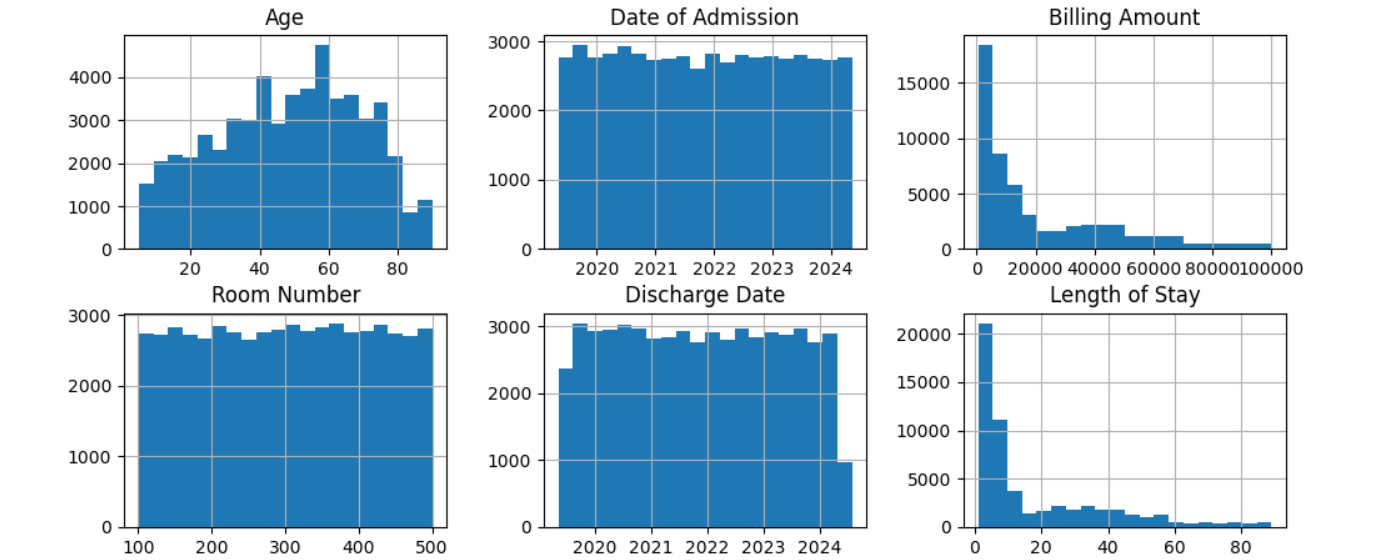
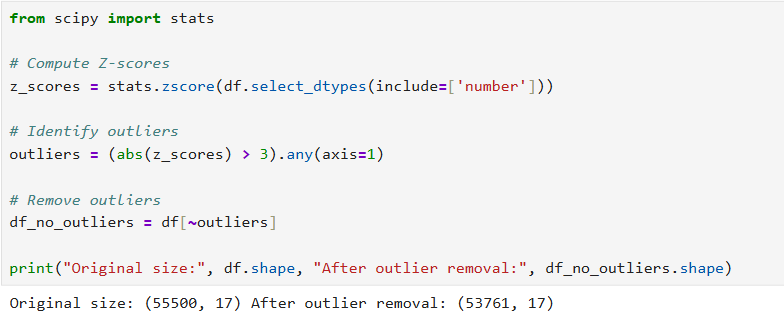
## Deployment

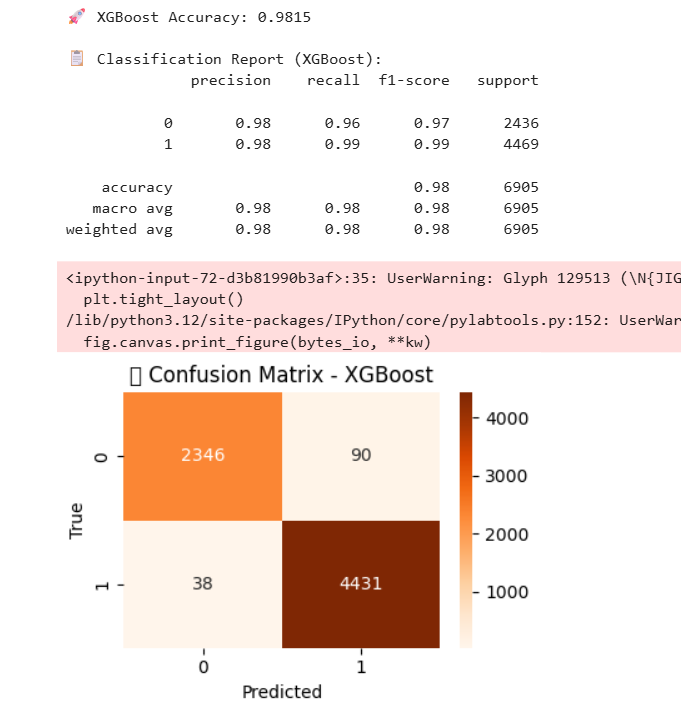
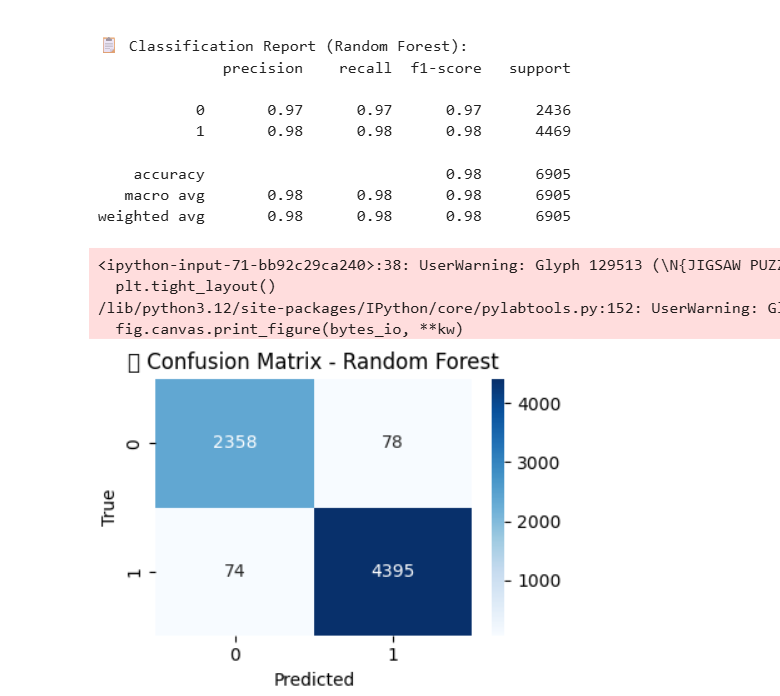
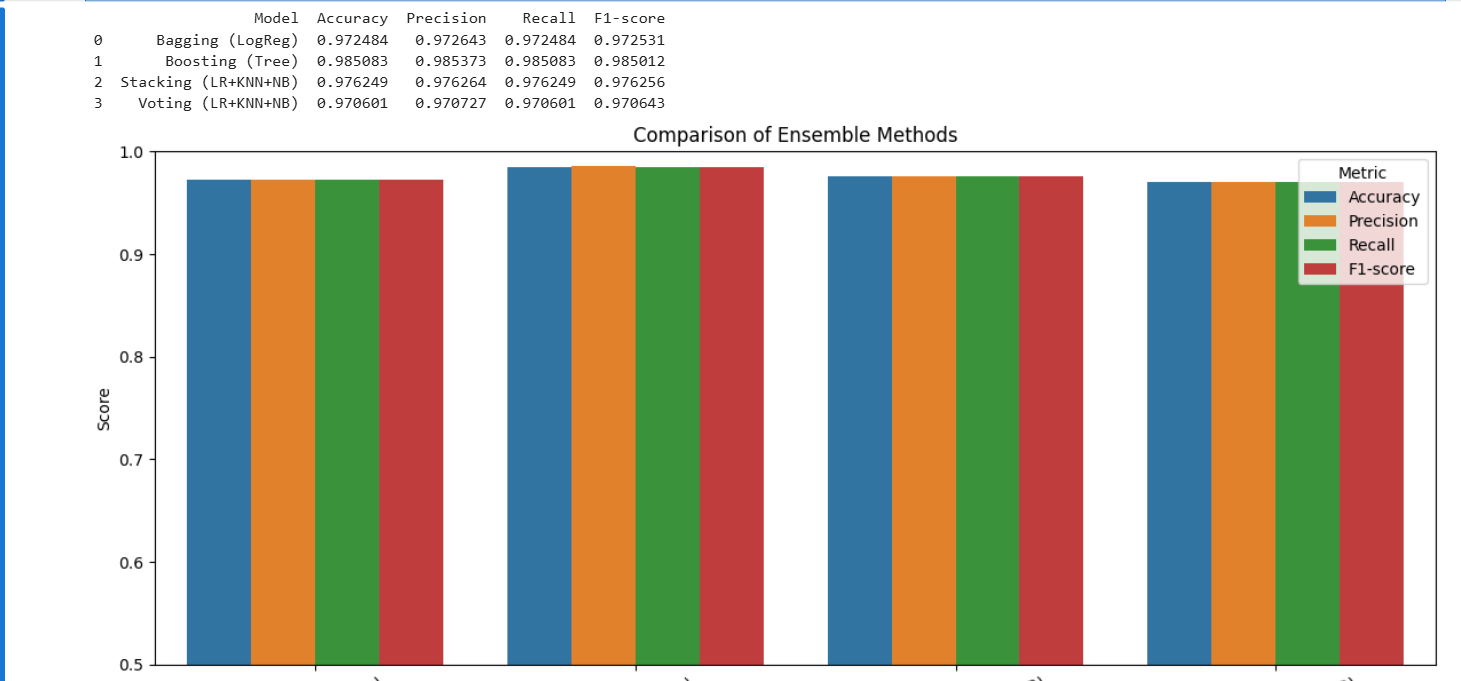
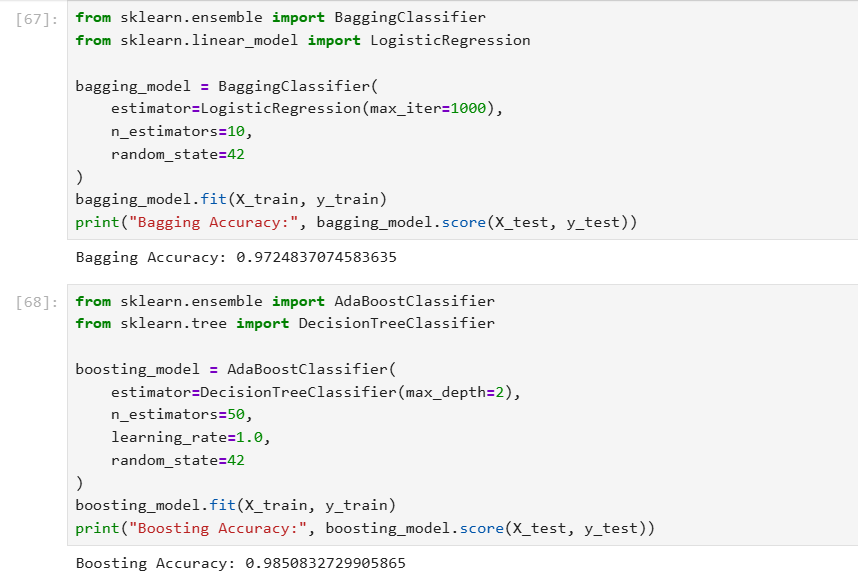
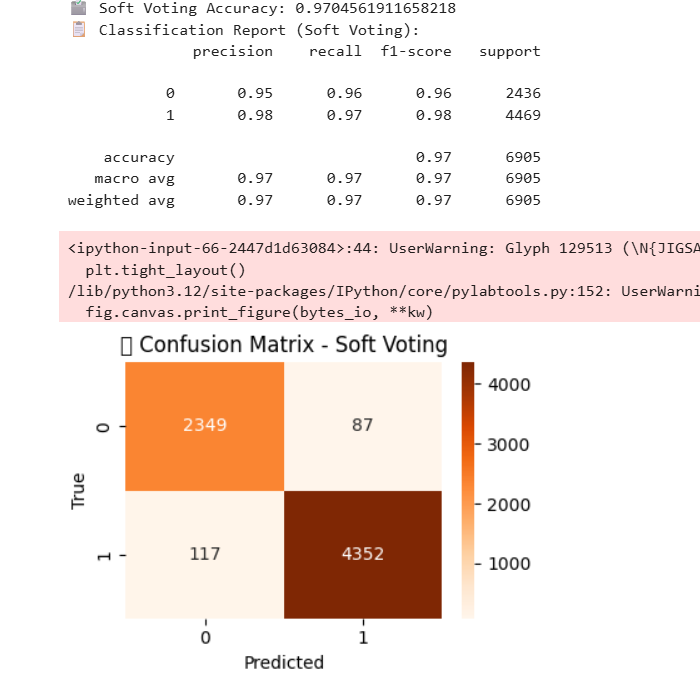
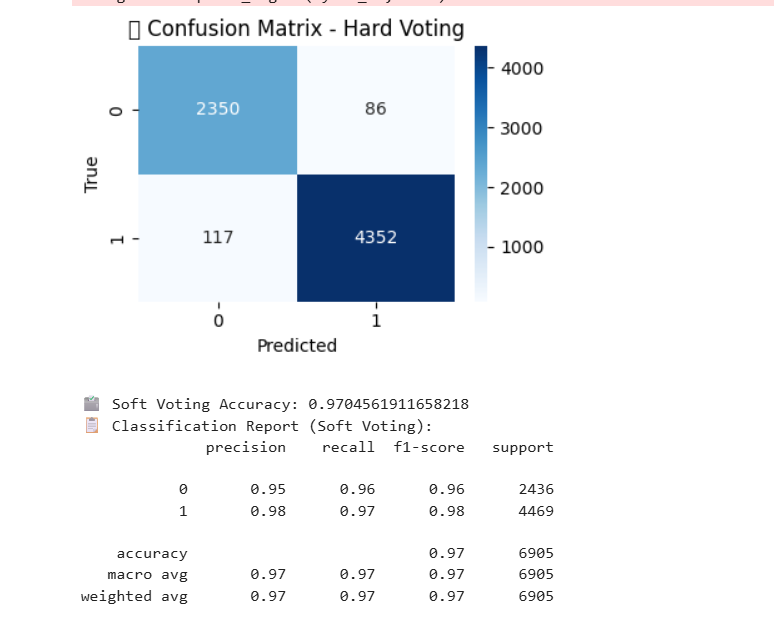
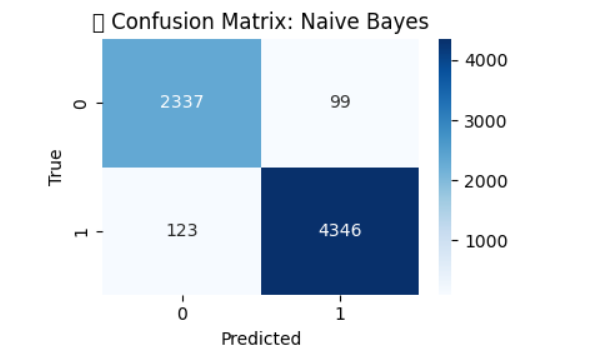
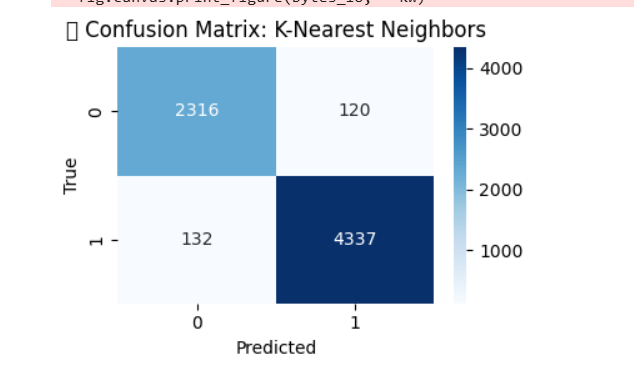
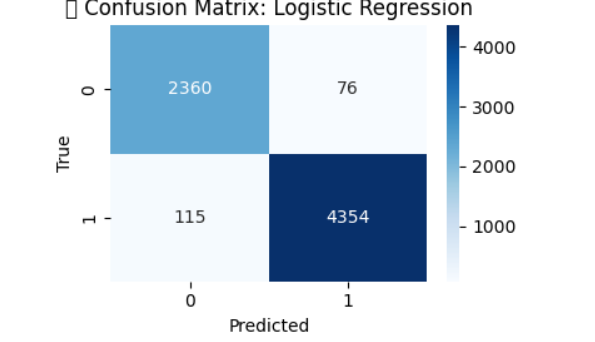
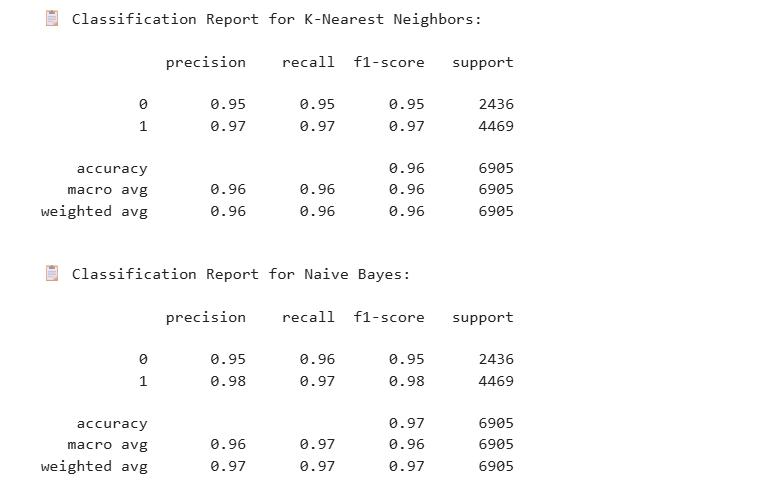
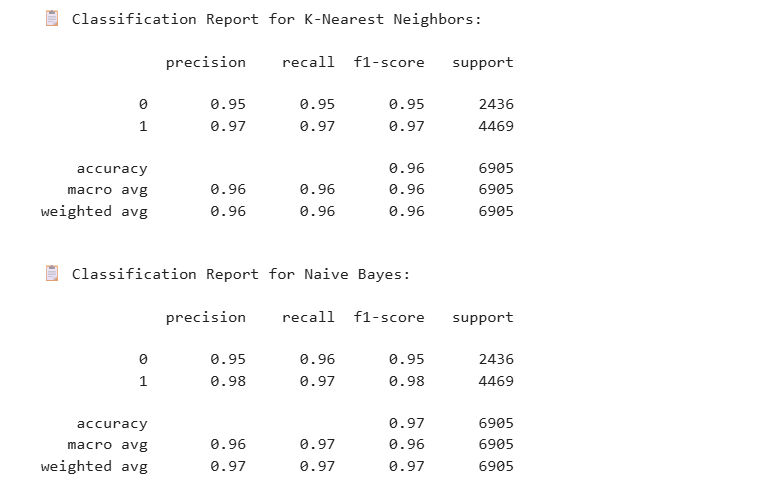
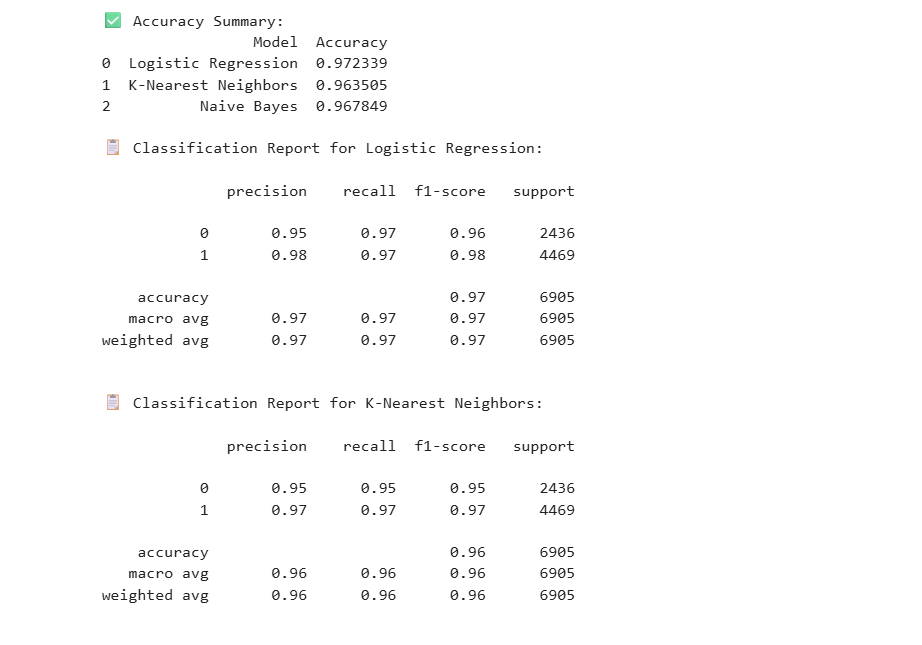
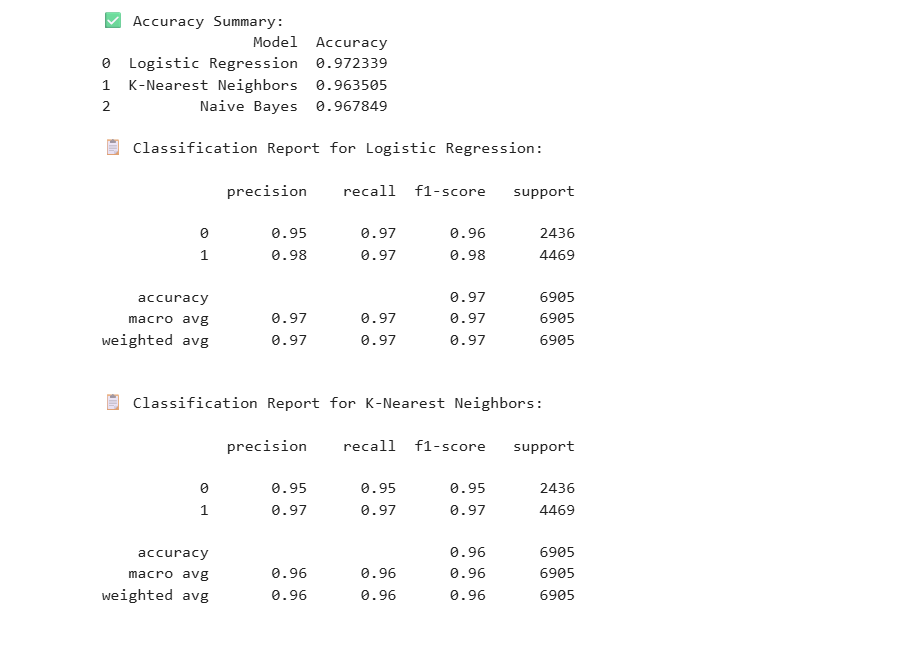
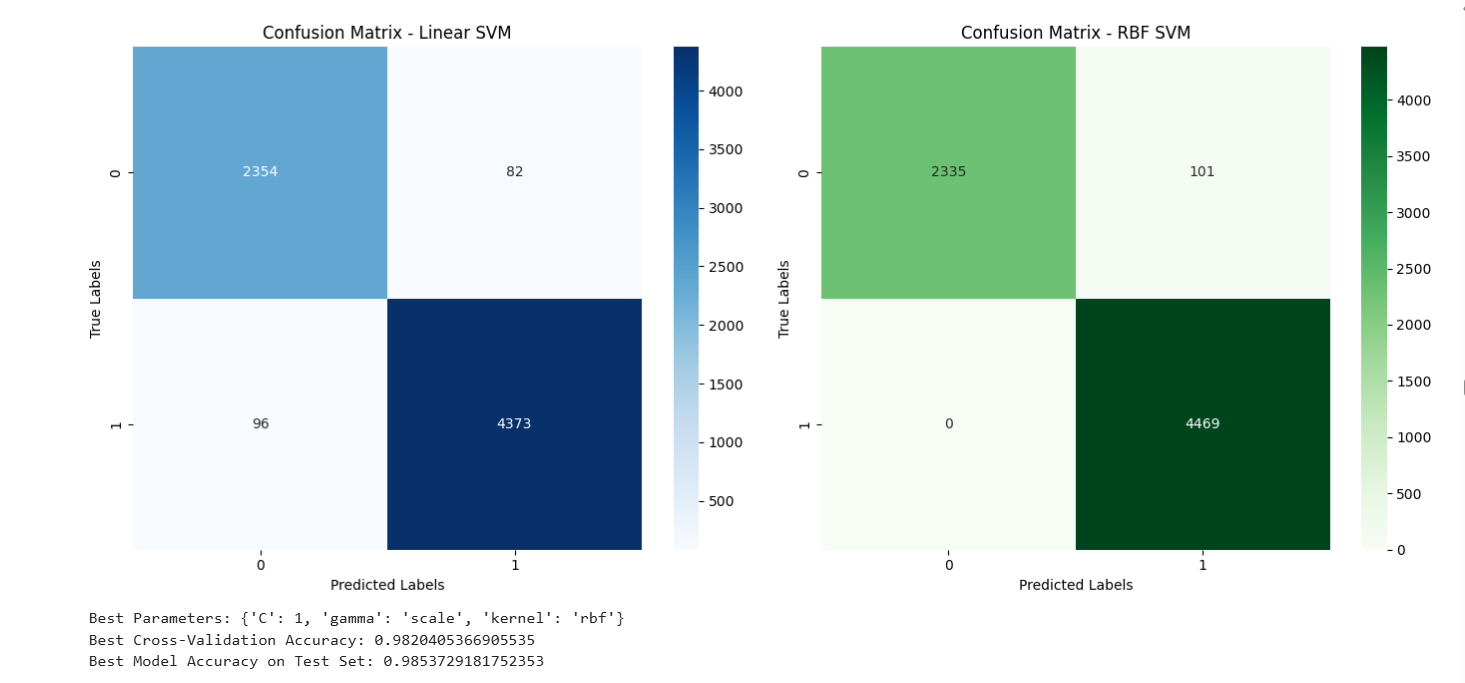
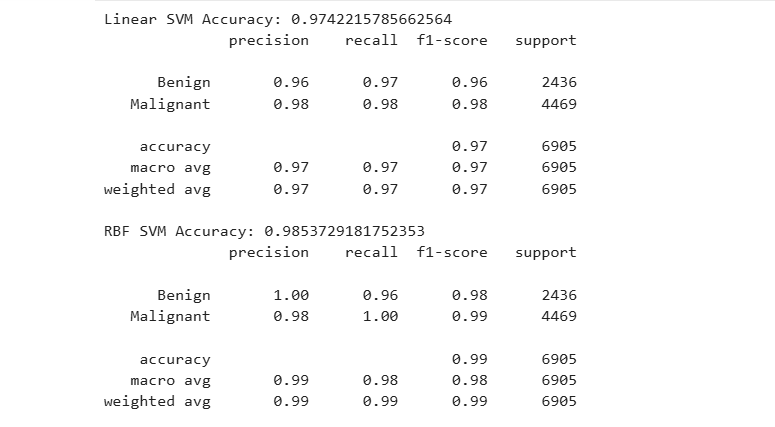
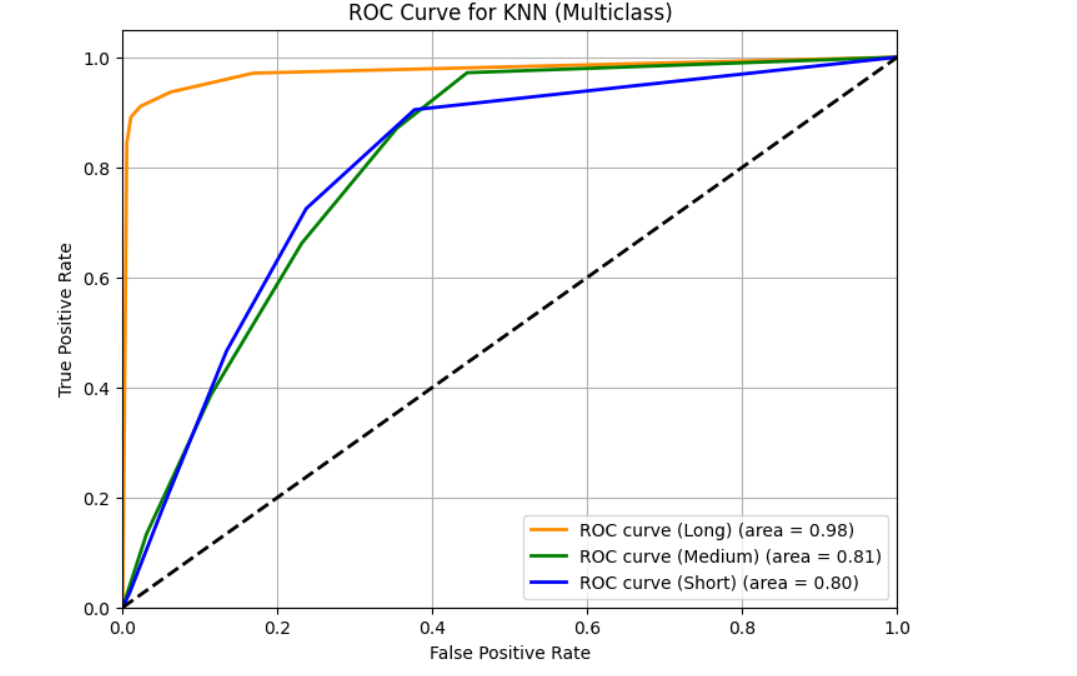
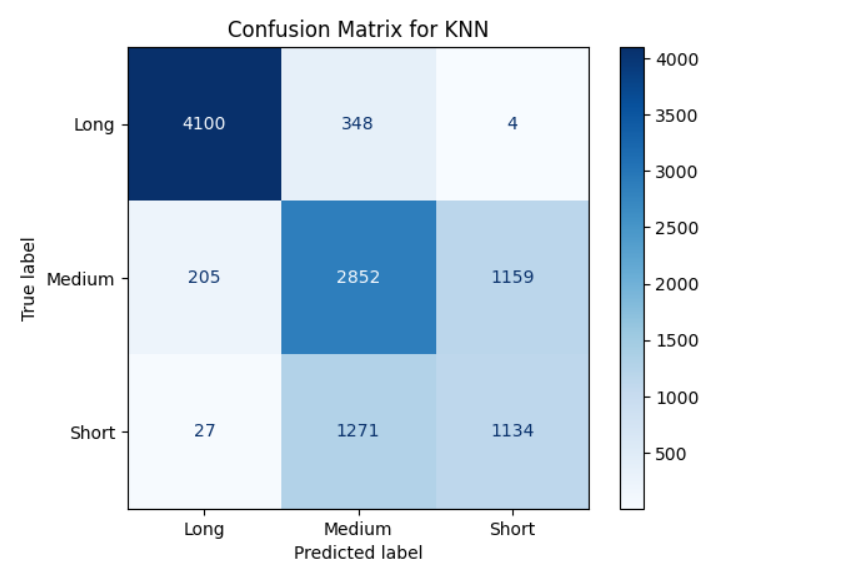
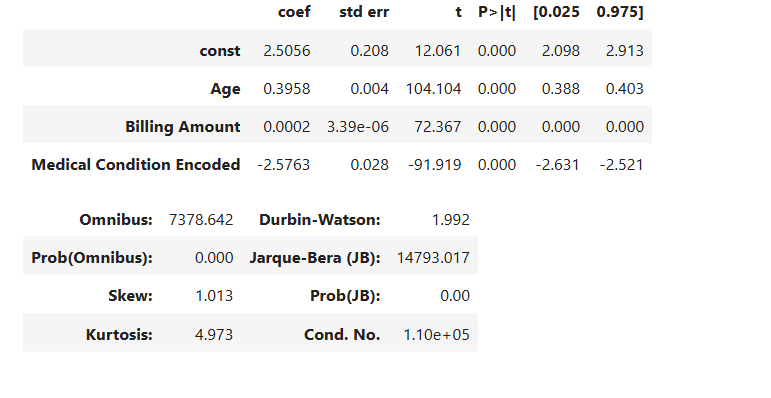
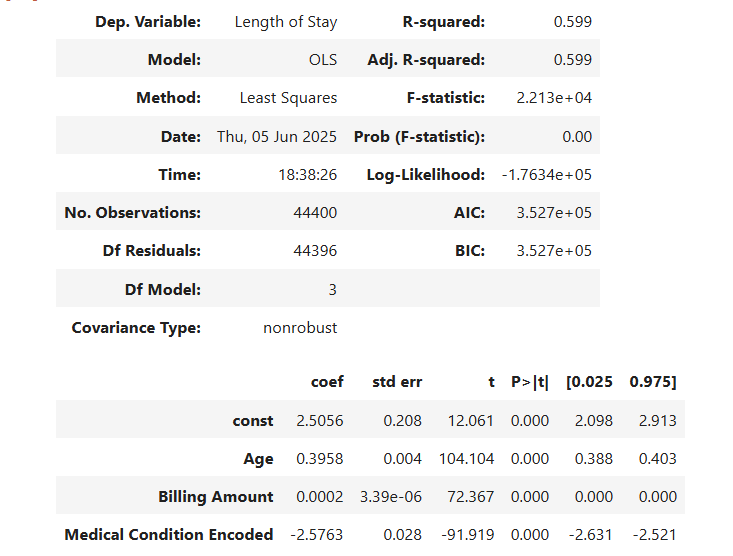
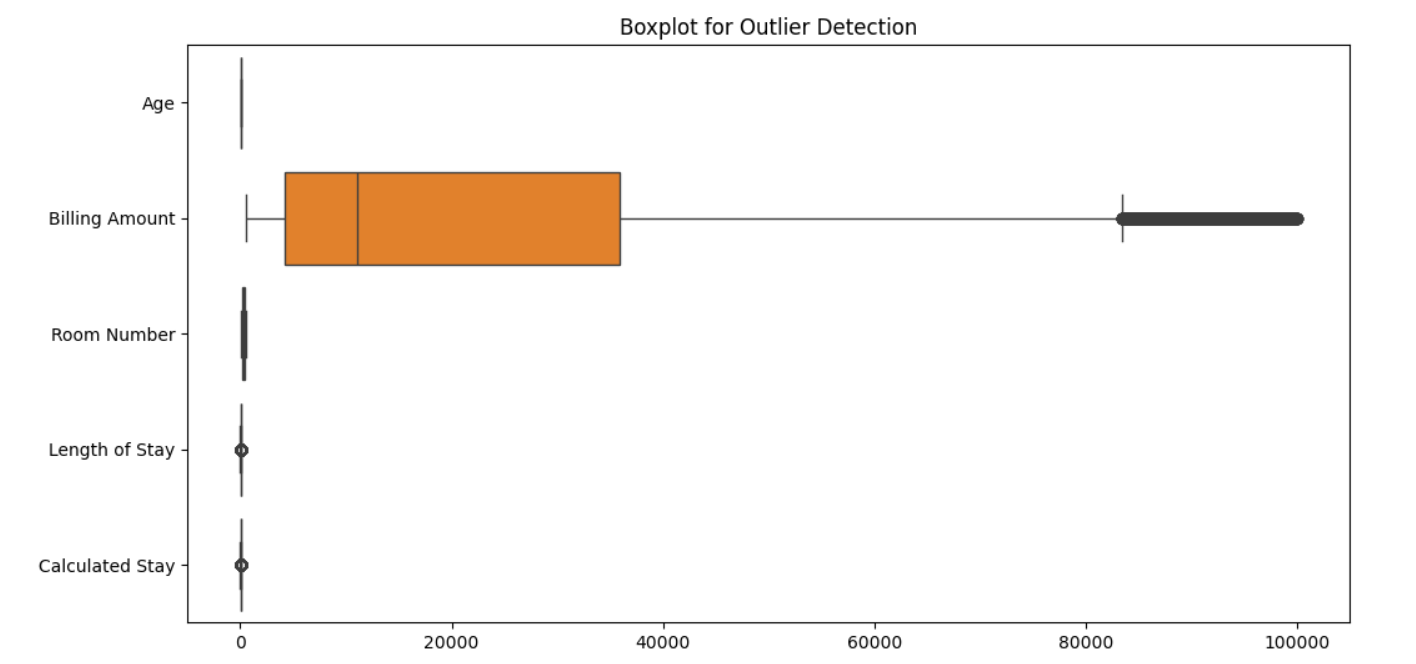
A screenshot of a computer

Description automatically generated• Final model (XGBoost) deployed using Flask  
• Web interface takes Age, Billing Amount, and Medical Condition (encoded) as input  
• Output: Predicted hospital stay category

A screenshot of a computer

Description automatically generated

A graph of a chart

Description automatically generated with medium confidence

## Conclusion

This project demonstrates the application of machine learning in healthcare for resource planning and patient flow prediction. With robust preprocessing and model tuning, we achieved a reliable and interpretable solution.