# Final Project Report

**Title: Clinical Data Analysis: Insights and Recommendations**

## Abstract

This study focuses on understanding patient survival outcomes through clinical data analysis. Leveraging datasets from the MIMIC-III clinical database, the research integrates demographic, diagnosis, laboratory, medication, and vital signs data to create comprehensive patient profiles. Advanced data preprocessing, statistical analysis, and machine learning models such as Random Forest and XGBoost were applied. The findings reveal key predictors of survival and demonstrate that the Random Forest model outperforms traditional models, offering promising directions for enhancing clinical decision-making.

## Introduction

The growing availability of healthcare data enables data-driven improvements in patient care. This project explores clinical data analysis techniques aimed at understanding survival outcomes among patients. By integrating multiple datasets from MIMIC-III and applying predictive modeling, the study seeks to uncover key factors influencing survival and propose effective analytical approaches.

## Objective

- To integrate diverse clinical datasets for comprehensive patient profiling.  
- To classify patient survival outcomes across various time horizons.  
- To identify key factors influencing survival and evaluate predictive models' performance.

## Methodology

Data was extracted from the MIMIC-III database using BigQuery SQL and Python 3.0. The datasets included:  
- Demographics  
- Clinical Diagnoses  
- Laboratory Results  
- Medications  
- Vital Signs  
  
Datasets were merged using unique identifiers (`subject\_id`, `hadm\_id`) to create unified patient records. Missing values were handled with median or mean imputation depending on the presence of outliers (detected via the interquartile range method). Numerical features were scaled, and categorical features were encoded to prepare data for modeling.

## Analytical Method

Patients were classified into four survival classes:  
- Class 0: Less than 30 days  
- Class 1: 30 days to 1 year  
- Class 2: 1 to 5 years  
- Class 3: Over 5 years  
  
Statistical tests applied included:  
- ANOVA and Kruskal-Wallis tests for continuous variables  
- Chi-square tests for categorical variables  
  
Machine learning models such as Random Forest, XGBoost, Logistic Regression, SVM, and Naive Bayes were trained. Model performances were evaluated based on Accuracy, Micro-AUC, and Matthews Correlation Coefficient (MCC) scores.

## Results

- Random Forest outperformed other models with Accuracy = 0.909, Micro-AUC = 0.991, and MCC = 0.866, surpassing the results reported in related research papers.  
- Key predictors of patient survival included Age, BMI, and Vital Signs.  
- Compared to the research paper results, the models trained in this project showed significantly better predictive performance across all metrics.

## Conclusion

The integrated clinical data analysis successfully identified major factors influencing patient survival. The superior performance of the Random Forest model suggests it can serve as a strong baseline for future survival prediction studies. The findings offer critical insights for clinical decision-making and highlight the potential of advanced machine learning methods in healthcare analytics.