**I. THU THẬP & HIỂU DỮ LIỆU**

* **Nguồn dữ liệu:** Patient Survival Prediction Dataset từ Kaggle (91,713 mẫu, 85 cột).
* **Các loại dữ liệu:**
  + Số liệu sinh lý: nhịp tim, huyết áp, nhiệt độ, GCS, lab test...
  + Thông tin nhân khẩu học: tuổi, giới, chủng tộc.
  + Tình trạng bệnh lý nền: tiểu đường, ung thư di căn, suy gan.
  + Dự đoán APACHE IV về khả năng tử vong (từ hệ thống bệnh viện).
  + Biến mục tiêu: hospital\_death (0 hoặc 1).

**🧹 II. LÀM SẠCH & TIỀN XỬ LÝ DỮ LIỆU**

* **Xử lý thiếu dữ liệu:**
  + Loại bỏ cột trống hoàn toàn (Unnamed: 83).
  + Áp dụng KNN Imputer / Mean Imputation / Iterative Imputer cho các cột quan trọng (glucose, potassium, apache\_4a\_hospital\_death\_prob...).
* **Xử lý biến phân loại:**
  + One-hot encoding: icu\_type, ethnicity, icu\_admit\_source.
  + Label encoding: gender.
* **Chuẩn hóa biến số:**
  + Dùng StandardScaler hoặc RobustScaler để chuẩn hóa biến liên tục.
* **Xử lý mất cân bằng dữ liệu:**
  + Tỷ lệ tử vong chỉ ~8.63% → cần dùng SMOTE, hoặc thêm class weights vào mô hình.

**📊 III. PHÂN TÍCH DỮ LIỆU (EDA)**

* **Thống kê mô tả:** Trung bình tuổi, nhịp tim, glucose...
* **Trực quan hóa:** histogram, boxplot, heatmap tương quan.
* **Phân tích mối liên hệ:**
  + GCS thấp → tăng tử vong.
  + APACHE score cao → tăng tử vong.
  + Tuổi cao, glucose cao, kali bất thường → tăng tử vong.

**🤖 IV. XÂY DỰNG MÔ HÌNH MACHINE LEARNING**

* **Mô hình thử nghiệm:**
  + Baseline: Logistic Regression
  + Nâng cao: Random Forest, XGBoost, LightGBM
* **Chiến lược đánh giá:**
  + Chia train/test, hoặc K-Fold CV.
  + Đánh giá bằng: Accuracy, ROC-AUC, F1-score, Confusion Matrix.
* **Tối ưu mô hình:**
  + Dùng GridSearchCV để tìm bộ siêu tham số tốt nhất cho XGBoost/LightGBM.

**🧠 V. GIẢI THÍCH MÔ HÌNH (EXPLAINABLE AI)**

* **Dùng SHAP:**
  + Xác định đặc trưng ảnh hưởng mạnh nhất đến tử vong.
  + Biểu đồ: summary plot, force plot, dependence plot.
* **LIME (nếu cần):**
  + Giải thích từng bệnh nhân cụ thể → mô hình hoạt động như thế nào cho từng cá nhân.