**Predicting ICU Mortality: Preliminary Analysis**

**Project Overview**

This project aims to predict in-hospital mortality for ICU patients using a public, time-series dataset from the PhysioNet 2012 Challenge, available on Kaggle. The dataset contains records for 4,000 ICU stays, including static demographic data and dynamic, time-stamped physiological measurements.

The goal of this initial notebook is to perform the necessary data loading, preprocessing, and exploratory data analysis (EDA) to establish a baseline understanding of the data and identify key challenges and features. The findings will inform the subsequent feature engineering and modeling phases.

**Proposed High-Level Plan:**

1. **Feature Engineering:** Extract meaningful features from the time-series data (e.g., summary statistics, trends over time).
2. **Modeling:** Build and compare several classification models (e.g., Logistic Regression, XGBoost) to predict the In-hospital\_death outcome.
3. **Address Challenges:** Implement strategies to handle significant class imbalance and missing data.
4. **Interpretation:** Use SHAP (SHapley Additive exPlanations) to interpret the model's predictions and understand which features are the most influential.
5. **Dashboard (Stretch Goal):** Create a simple interactive dashboard to visualize model predictions.

**Detailed Plan**

**Phase 1: Data Foundation & Feature Engineering (Weeks 1-2)**

* **Week 1: Advanced EDA & Data Cleaning**
  + You've already started this! Finish the initial EDA.
  + **Deep Dive:** Analyze correlations between all features. Are some redundant?
  + **Missing Data Strategy:** The pH feature is missing in over 50% of cases. Do you drop it? Or do you use a sophisticated imputation method (like KNNImputer or MICE) and justify your choice? You'll need to research and experiment.
  + **Outlier Detection:** Check for and handle outliers in your features. Does a patient with a recorded temperature of 25°C make sense, or is it a data entry error?
* **Week 2: Advanced Feature Engineering**
  + **Beyond Aggregates:** Instead of just mean, min, and max, could the *trend* or *slope* of a vital sign be more predictive? (e.g., Is the patient's heart rate rapidly increasing over the first 12 hours?). This requires more complex feature creation.
  + **Interaction Features:** Could the ratio of two features be important (e.g., BUN to Creatinine ratio)?
  + **Domain Research:** Briefly research clinical literature. Are there other known risk factors you can engineer from the existing data?

**Phase 2: Modeling & Rigorous Evaluation (Weeks 3-4)**

* **Week 3: Baseline & Advanced Modeling**
  + **Establish a Baseline:** Build a simple Logistic Regression model. This is the score you have to beat.
  + **Experiment with Models:** Implement and train more powerful models like **XGBoost**, **LightGBM**, and maybe a simple **Neural Network**.
  + **Address Class Imbalance:** The outcome is skewed. You need to implement and compare techniques like SMOTE (Synthetic Minority Over-sampling Technique) or using class\_weights in your models to ensure it learns to predict the minority (death) class effectively.
* **Week 4: Hyperparameter Tuning & Validation**
  + **Systematic Tuning:** For your best-performing model (e.g., XGBoost), perform rigorous hyperparameter tuning using techniques like GridSearchCV or RandomizedSearchCV to find the optimal settings. This is computationally intensive and takes time.
  + **Cross-Validation:** Ensure you are using a robust cross-validation strategy (like StratifiedKFold) to get a reliable estimate of your model's performance and avoid overfitting.

**Phase 3: Interpretation & Insights (Weeks 5-6)**

* **Week 5: Model Interpretation with SHAP**
  + This is a core part of your project. It's not enough to predict; you have to *explain*.
  + **Global Explanations:** Generate SHAP summary plots to identify the top 10-15 most important features for your model overall.
  + **Local Explanations:** Generate SHAP dependence plots to understand *how* a feature impacts the prediction (e.g., does higher age *always* increase mortality risk, or does it level off?).
* **Week 6: Case Studies & Final Evaluation**
  + **Drill Down:** Use SHAP force plots to explain the prediction for specific, individual patients. For example: "For patient X, the model predicted high risk primarily due to their low pH and high age, even though their heart rate was normal."
  + **Final Metrics:** Evaluate your final model using a comprehensive set of metrics beyond just accuracy (AUC-ROC, Precision-Recall Curve, F1-Score). Justify why these metrics are appropriate for an imbalanced medical dataset.

**Phase 4: Synthesis & Communication (Week 7)**

* **Week 7: Reporting and Presentation**
  + Synthesize all your findings into a final report or paper.
  + Create a compelling presentation that tells the story of your project: the problem, the data challenges, your solutions, and the final insights your model provides.
  + Clean, comment, and finalize your code notebook so it's fully reproducible.