# Methodology

This project implements a robust machine learning pipeline to predict survival outcomes of patients based on clinical and laboratory data, using binary classification. The methodology consists of the following steps:

## 1. Data Collection

We collected a real-world medical dataset compiled from several Excel sheets. The dataset included diverse clinical measurements and laboratory test results of patients exposed to Zinc Phosphide. All sheets were merged into a single dataset for analysis.

## 2. Data Cleaning and Preprocessing

The dataset underwent thorough cleaning and preparation:

- Duplicate columns were removed.

- Irrelevant features, such as discharge vitals and final outcome indicators, were excluded.

- Missing values were handled by imputing with a constant placeholder value (`-999`) to maintain compatibility with machine learning algorithms.

- Categorical variables were encoded numerically using LabelEncoder.

## 3. Feature Selection (Top 10)

To improve model performance and interpretability:

- A Random Forest model was trained on the full dataset.

- Feature importances were computed.

- The 10 most influential features were selected based on their importance scores.

- Subsequent modeling used only these top 10 features to reduce complexity while retaining predictive power.

## 4. Train/Test Split

The balanced dataset was split into training (80%) and testing (20%) subsets using train\_test\_split, with stratification to preserve the class ratio of survival and death outcomes.

## 5. Data Balancing

Due to class imbalance in the original dataset (i.e., few patients labeled as 'Died'), the minority class was upsampled by replicating its observations to match the majority class. This ensured balanced classes in the training data.

## 6. Model Training

A Random Forest classifier was trained using the top 10 features on the balanced and scaled training set. Hyperparameters were tuned using cross-validation (StratifiedKFold) to achieve optimal performance.

## 7. Model Evaluation

The model was evaluated on the test set using standard metrics:

- Accuracy

- Precision

- Recall

- F1-Score

The Random Forest achieved excellent results, with all metrics indicating a highly accurate and reliable model.

## 8. Feature Importance Analysis

A bar plot was generated to visualize the relative importance of the top 10 features. Features such as RDW, Creatinine, and Poison Severity Score were found to contribute most significantly to the model’s predictions.

## 9. Deployment & Prediction

The final model was saved using pickle for reproducibility. A user-friendly Streamlit web application was developed with two input modes:

- Batch prediction through CSV or Excel file upload.

- Manual data entry via a form interface.

The app outputs survival predictions along with an option to download the results as a CSV file.