# Heart Failure Prediction

Modern Academy

الاكاديمية الحديثة للهندسة والتكنولوجيا

## TEAM WORK:

| | |
|---|---|
| 1_ALI AHMED ALI FADLOUN | 4230170 |
| 2_MOSTAFA MOHAMED MOSTAFA | 4230165 |
| 3_HAYDI AHMED | 4230047 |
| 4_MALAK MOHAMED | 4230029 |

# 1- INTRODUCTION:

Heart failure is one of the most common and serious chronic medical conditions in the modern era, posing a significant challenge to healthcare systems worldwide. It occurs when the heart is no longer able to pump enough blood to meet the body's demands for oxygen and nutrients. Over time, this condition can lead to the deterioration of vital organ functions, worsening symptoms, and a marked increase in mortality—particularly among elderly individuals and patients with underlying chronic diseases.

Heart failure is associated with a variety of medical and behavioral risk factors, including high blood pressure, diabetes, coronary artery disease, obesity, smoking, and sedentary lifestyles.

## 1- Exploratory Data Analysis:

- **Numerical**: `age`, `ejection_fraction`, `creatinine_phosphokinase`, `platelets`, `serum_creatinine`, `serum_sodium`, `time`
- **Categorical/Binary**: `anaemia`, `diabetes`, `high_blood_pressure`, `sex`, `smoking`, `DEATH_EVENT`

## 2_Visualizations used:

- Histograms & KDE plots for distributions.
- Bar charts & Pie charts for binary features.
- Scatter plots (with `hue=sex`) to explore correlation with `age` and `DEATH_EVENT`

## 3_Objectives:

- Analyze clinical features to understand their impact on heart failure.
- Develop and evaluate multiple machine learning models for predicting heart failure.
- Identify the most significant predictors of heart failure

## 4_ Dataset Overview:

The dataset contains **299 patient records** with **13 clinical features**, including:

- **Demographics:** Age, sex.
- **Medical Conditions:** Anaemia, diabetes, high blood pressure, smoking.

- **Biochemical Markers:** Creatinine phosphokinase, ejection fraction, platelets, serum creatinine, serum sodium.
- **Follow-up Time:** Time until the event (death or survival).
- **Target Variable:** DEATH_EVENT (binary: 1 for death, 0 for survival).

## 2- Data Exploration & Analysis:

### 2.1 Data Representation

- The dataset was loaded and inspected using df.head(), df.tail(), and df.info().
- No missing values were found (df.isna().sum() confirmed completeness).
- Features included both numerical (e.g., age, serum creatinine) and categorical (e.g., diabetes, smoking) variables.

## 2.2 Key Findings from EDA

### Age Distribution

- Patients' ages ranged from **40 to 95**, with a peak around **60–70**.
- Older patients (>70) showed a higher likelihood of death.

### Medical Conditions

- **Anaemia (43%):** Slightly imbalanced (57% no anaemia).
- **Diabetes (42%):** Nearly balanced distribution.
- **High Blood Pressure (35%):** Present in a third of patients.
- **Smoking (32%):** Smokers were less frequent than non-smokers.

### Biochemical Markers

- **Creatinine Phosphokinase:** Highly skewed, with most values <1000.
- **Ejection Fraction:** Most patients had values between **30–60%** (normal range: 50–70%).

- **Serum Creatinine:** Elevated levels (>1.2 mg/dL) correlated with higher mortality.
- **Serum Sodium:** Lower levels (<135 mEq/L) were associated with worse outcomes.

**Target Variable (**DEATH_EVENT**)**

- **Imbalanced Dataset:** 68% survived, 32% died.

### 2.3 Visualizations

- **Histograms & KDE Plots:** Showed distributions of numerical features (e.g., age, serum creatinine).
- **Bar Plots & Pie Charts:** Illustrated categorical feature distributions (e.g., diabetes, smoking).
- **Scatter Plots:** Revealed relationships (e.g., age vs. serum creatinine, stratified by sex).

## 3-Data Preprocessing:

### 3.1 Feature-Target Separation

- **Features (**X**):** All columns except DEATH_EVENT.
- **Target (**y**):** DEATH_EVENT.

### 3.2 Scaling

- **MinMaxScaler** was applied to platelets to normalize values between 0 and 1.

### 3.3 Train-Test Split

- **80% training, 20% testing** with random_state=42 for reproducibility.

## 4- Model Development & Evaluation:

### 4.1 Models Tested

1. **Logistic Regression**
2. **Support Vector Machines (SVC, LinearSVC)**
3. **K-Nearest Neighbors (KNN)**
4. **Decision Tree**
5. **Ensemble Methods:**
    - Bagging
    - Random Forest
    - Extra Trees
    - XGBoost

6. **Voting Classifier** (Logistic Regression, Random Forest, XGBoost)
7. **Stacking Classifier** (SVC + KNN → Logistic Regression)

### 4.2 Performance Metrics

- **Accuracy:** Overall correctness.
- **Precision (Class 1):** Proportion of true positives among predicted positives.
- **Recall (Class 1):** Proportion of actual positives correctly predicted.
- **F1-Score (Class 1):** Harmonic mean of precision and recall.

### 4.3 Results

| **Model** | **Accuracy** | **Precision (1)** | **Recall (1)** | **F1-Score (1)** |
|---|---|---|---|---|
| Random Forest | 0.992 | 0.984 | 0.990 | 0.987 |

| Model | Accuracy | Precision (1) | Recall (1) | F1-Score (1) |
|---|---|---|---|---|
| Extra Trees | 0.992 | 0.984 | 0.990 | 0.987 |
| XGBoost | 0.992 | 0.984 | 0.990 | 0.987 |
| Voting Classifier | 0.991 | 0.980 | 0.990 | 0.985 |
| Stacking Classifier | 0.977 | 0.960 | 0.964 | 0.962 |

**Key Observations:**

- **Top Performers:** Random Forest, Extra Trees, and XGBoost achieved **~99.2% accuracy**.
- **Logistic Regression & SVC:** Underperformed (accuracy <85%).
- **Imbalance Handling:** Ensemble methods effectively addressed class imbalance.