**صورة تحتوي على نص, الخط, شعار, التصميم

قد يكون المحتوى المعد بواسطة الذكاء الاصطناعي غير صحيح.**

**Machine Learning Report**

**Heart Failure Clinical Records Analysis**

**Student Name:** Reyoof saeed Al-otaibi

**Student ID:** 44102870ml

**Instructor Name:** Dr. Nada Al-Towairqi

**Course Name:** Machine Learning

Table of Contents

[List of Figures 3](#_Toc197999612)

[1. Introduction 4](#_Toc197999613)

[2. Dataset Description 4](#_Toc197999614)

[3. Data Properties 5](#_Toc197999615)

[4. Preprocessing 9](#_Toc197999616)

[5. Machine Learning Models & Results 9](#_Toc197999617)

[6. Results Discussion 13](#_Toc197999618)

[7. Conclusion 14](#_Toc197999619)

[8. Github Link: 15](#_Toc197999620)

List of Figures

[Figure 1: Feature distributions. 6](#_Toc197999621)

[Figure 2: Target distribution. 7](#_Toc197999622)

[Figure 3: Features correlation. 8](#_Toc197999623)

[Figure 4: Models accuracies comparison. 10](#_Toc197999624)

[Figure 5: ROC curves comparison. 11](#_Toc197999625)

[Figure 6: Confusion matrix of all models. 12](#_Toc197999626)

[Figure 7: Classification heatmaps. 13](#_Toc197999627)

# Introduction

Cardiovascular diseases (CVDs) are the leading cause of death globally, responsible for an estimated **17.9 million lives annually** (31% of all deaths worldwide). Heart failure, a severe outcome of CVDs, can be predicted using clinical data to enable early intervention. This report analyzes the **Heart Failure Clinical Records Dataset** to develop a predictive model for mortality risk in heart failure patients. The dataset includes **12 key clinical features** that help assess patient outcomes, supporting healthcare professionals in early detection and personalized treatment strategies.

# Dataset Description

* **Dataset Name**: Heart Failure Clinical Records Dataset
* **Source**: Kaggle
* **Original Data Link**:  <https://www.kaggle.com/datasets/andrewmvd/heart-failure-clinical-data/data>

This dataset focuses on predicting mortality due to heart failure, a common consequence of cardiovascular diseases (CVDs), which are the leading cause of death globally. It includes 12 health-related features. Since many CVDs are preventable and early detection is crucial for high-risk individuals, machine learning models can be valuable tools for predicting outcomes and aiding in timely intervention.

The dataset contains **299 (samples) patient records** with **12 clinical features (attributes)**and **1 binary target variable (DEATH\_EVENT)**. The features include:

* **Demographics**: Age
* **Medical Conditions**: Anaemia, diabetes, high blood pressure, smoking
* **Clinical Measurements**: Ejection fraction, serum creatinine, serum sodium, platelets
* **Follow-up Data**: Time (days until follow-up event)

**Importance of the Dataset**

* **Preventable Nature of CVDs**: Behavioral risk factors (tobacco use, unhealthy diet, physical inactivity) can be mitigated with early detection.
* **High-Risk Patient Management**: Helps identify patients needing urgent intervention (e.g., those with hypertension, diabetes, or hyperlipidemia).
* **Clinical Decision Support**: Machine-learning models can enhance risk stratification beyond traditional medical assessments.

# Data Properties

* **Problem Type**: Binary classification (Death Event: Yes/No)
* **Samples**: 299 patients
* **Features**: 12 clinical attributes
* **Class Distribution**:
  + **Survived (0)**: 203 (68%)
  + **Died (1)**: 96 (32%)

**Key Statistics**

| **Feature** | **Range** | **Mean** | **Correlation with Death** |
| --- | --- | --- | --- |
| Age | 40–95 | 60.8 | Positive |
| Ejection Fraction | 14–80 | 38.1 | Negative |
| Serum Creatinine | 0.5–9.4 | 1.4 | Positive |
| Serum Sodium | 113–148 | 136.8 | Negative |

**Visual Insights**

* **Age & Mortality**: Older patients show higher death rates.
* **Ejection Fraction**: Lower values correlate strongly with death risk.
* **Serum Creatinine**: Elevated levels indicate higher mortality.
* **No Missing Data**: Complete records ensured model reliability.

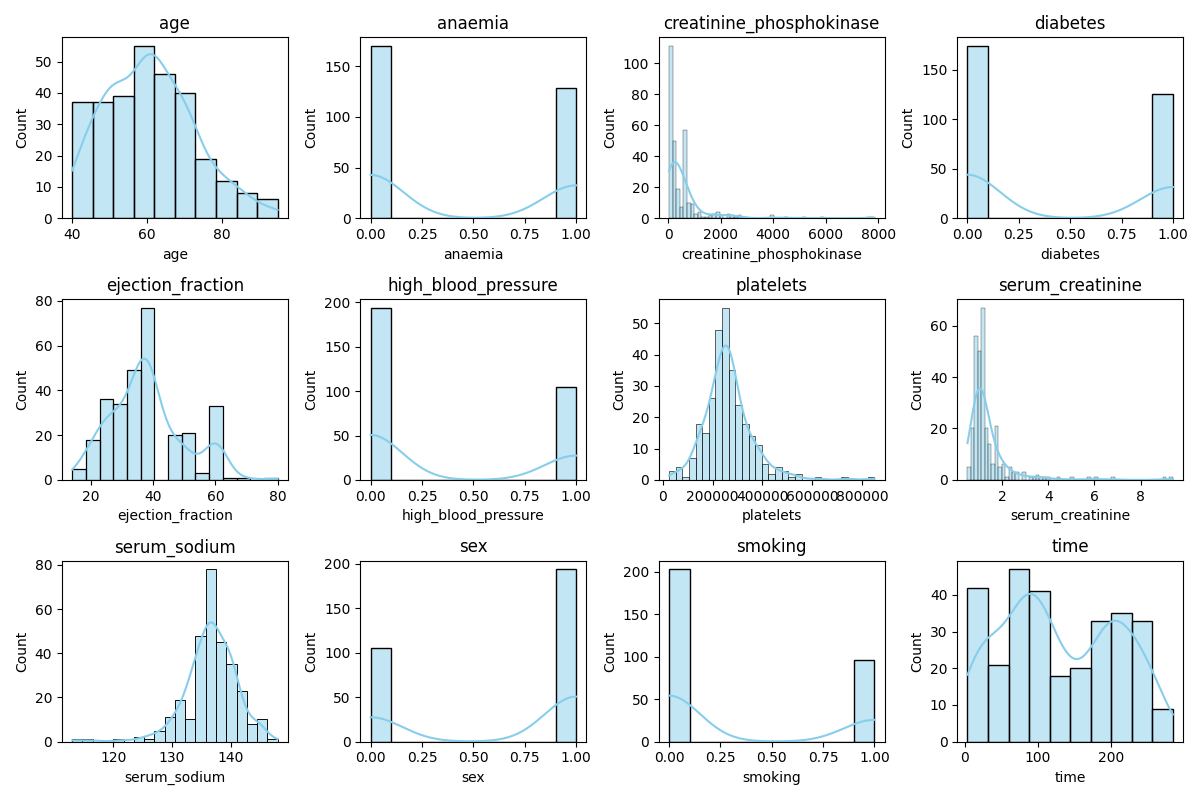


Figure 1: Feature distributions.

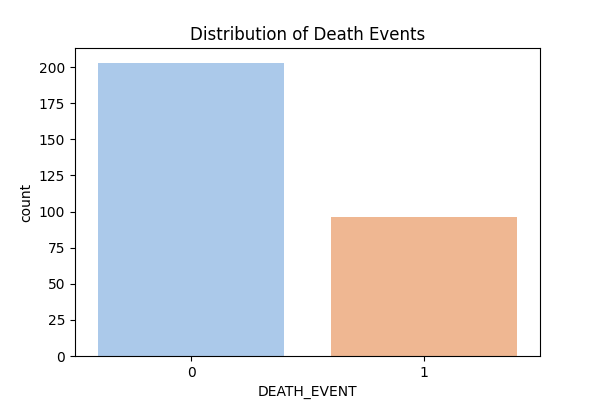


Figure 2: Target distribution.

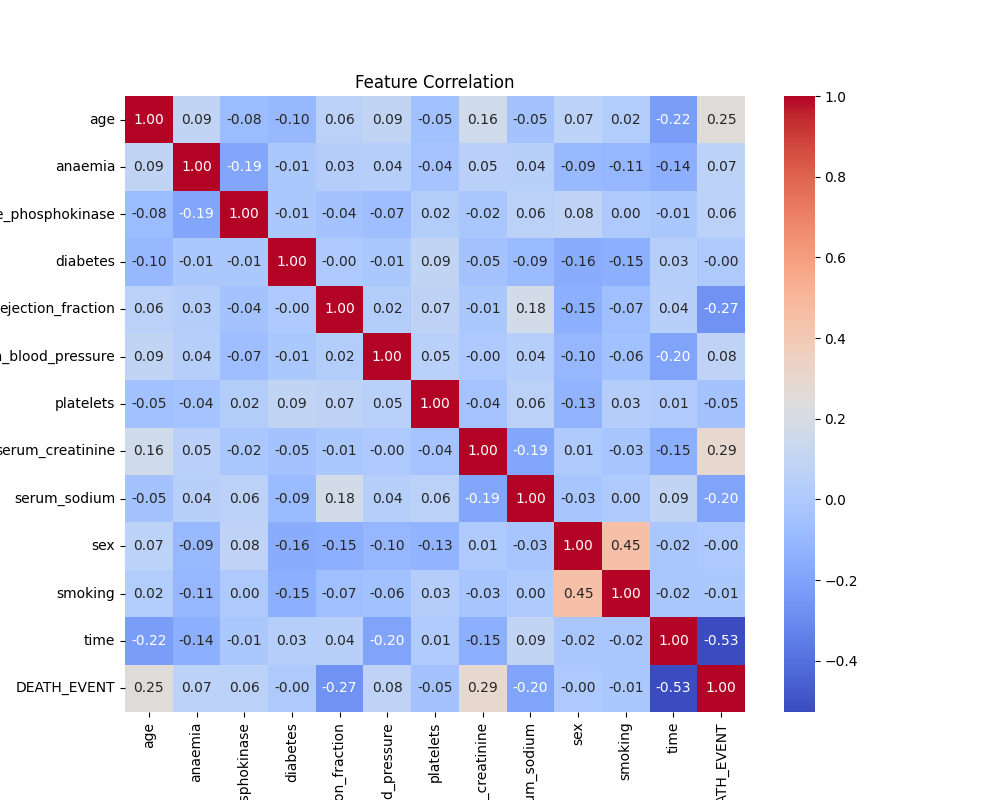
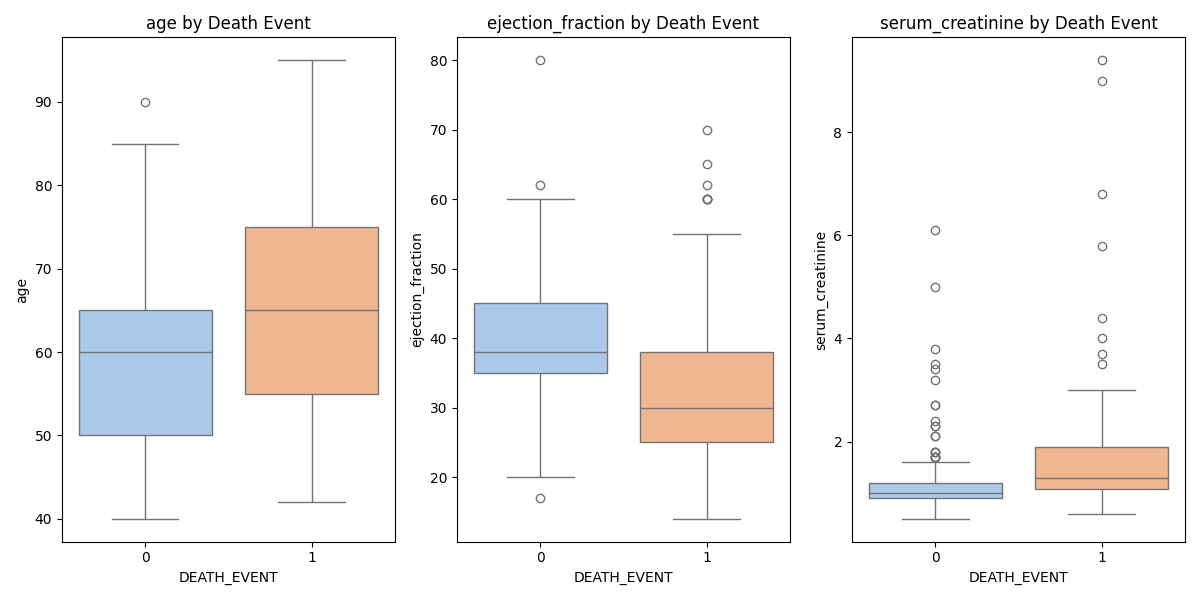


Figure 3: Features correlation.



# Preprocessing

* **Feature-Target Split**: Separated **X (features)** and **y (DEATH\_EVENT)**.
* **Standardization**: Applied **StandardScaler** to normalize feature scales.
* **Train-Test Split**:
  1. **Training (80%)**: 239 samples
  2. **Testing (20%)**: 60 samples

# Machine Learning Models & Results

Seven models were trained and evaluated: because our problem is classification, the linear regression model does not support the classification. We replace it by the logistic regression model.

| **Model** | **Accuracy** | **ROC AUC** |
| --- | --- | --- |
| Logistic Regression | 80% | 0.824 |
| SVM | 75% | 0.816 |
| **Random Forest** | **75%** | **0.827** |
| Naive Bayes | 70% | 0.818 |
| K-Nearest Neighbors | 68% | 0.745 |
| ANN | 68% | 0.776 |
| Decision Tree | 63 % | 0.606 |

**Best Model:**Logistic Regression

* **Accuracy**: 80%
* **ROC AUC**: 0.842 (Excellent discrimination)

**Worst Model: Decision Tree**

* **Accuracy**: 63%
* **Overfitting**: Likely due to high variance in small dataset.

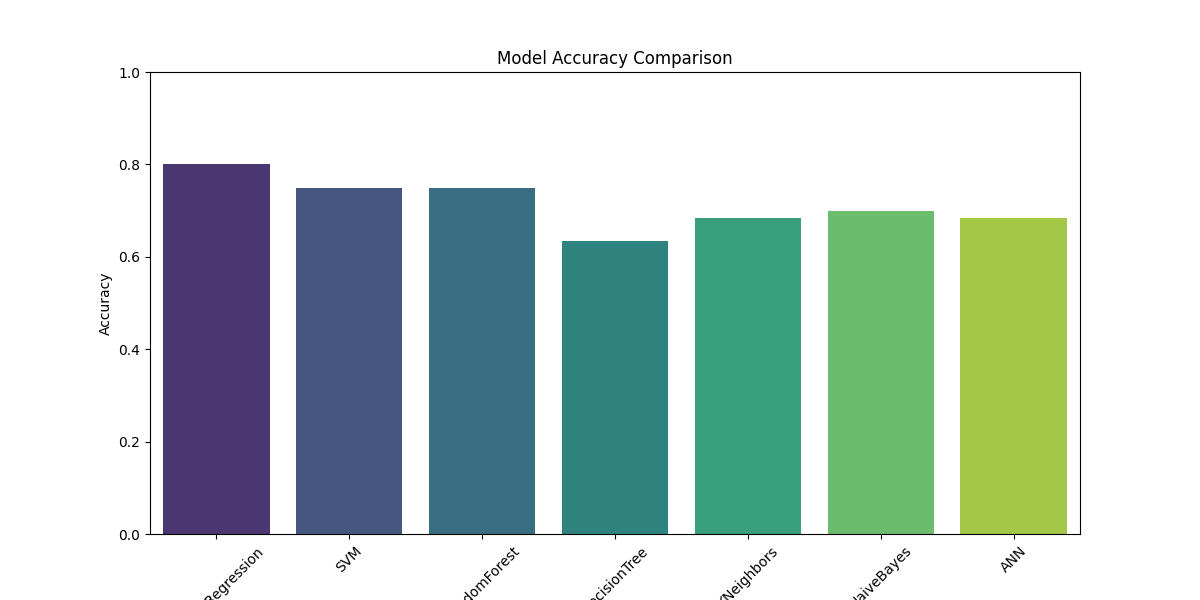


Figure 4: Models accuracies comparison.

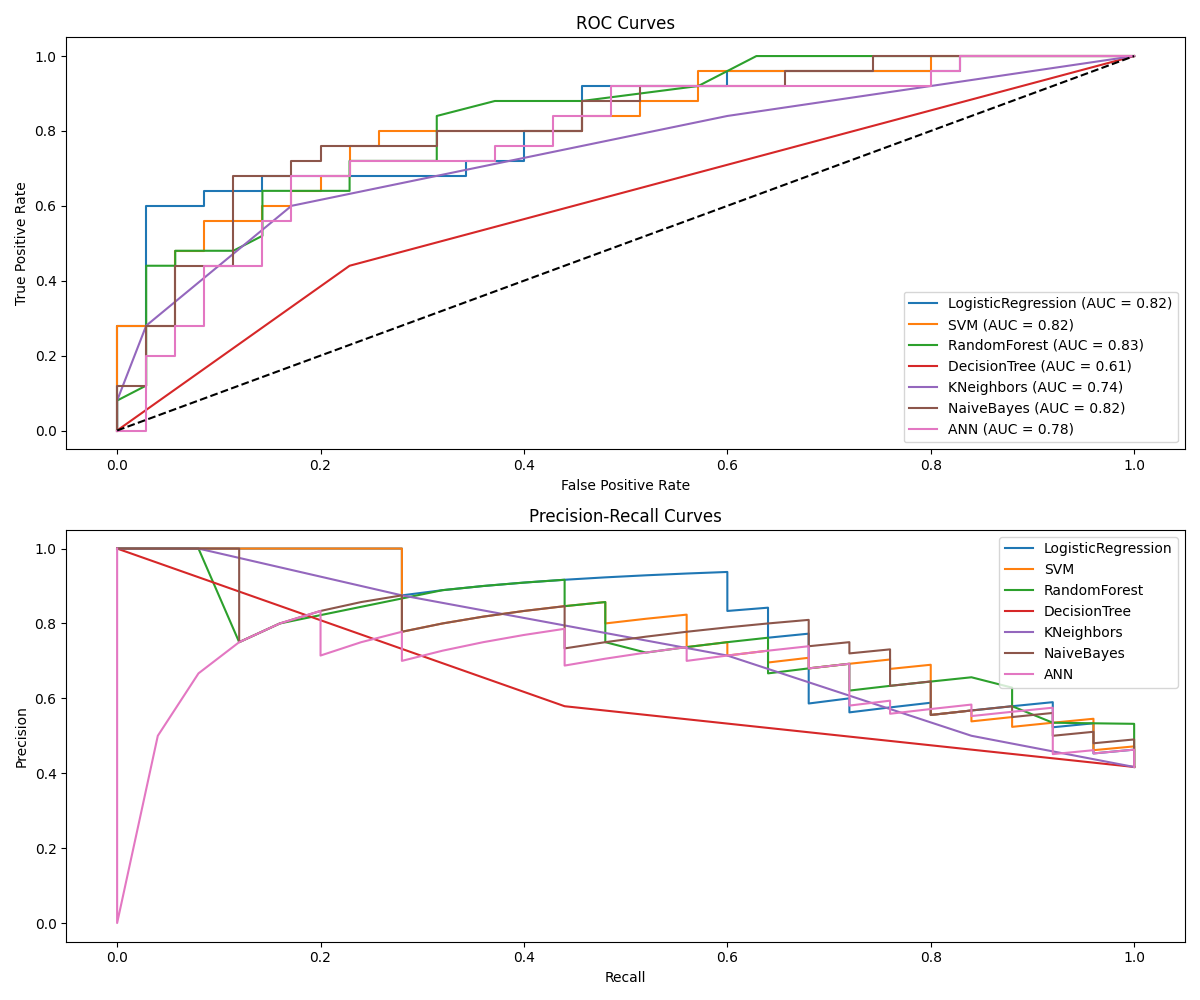


Figure 5: ROC curves comparison.

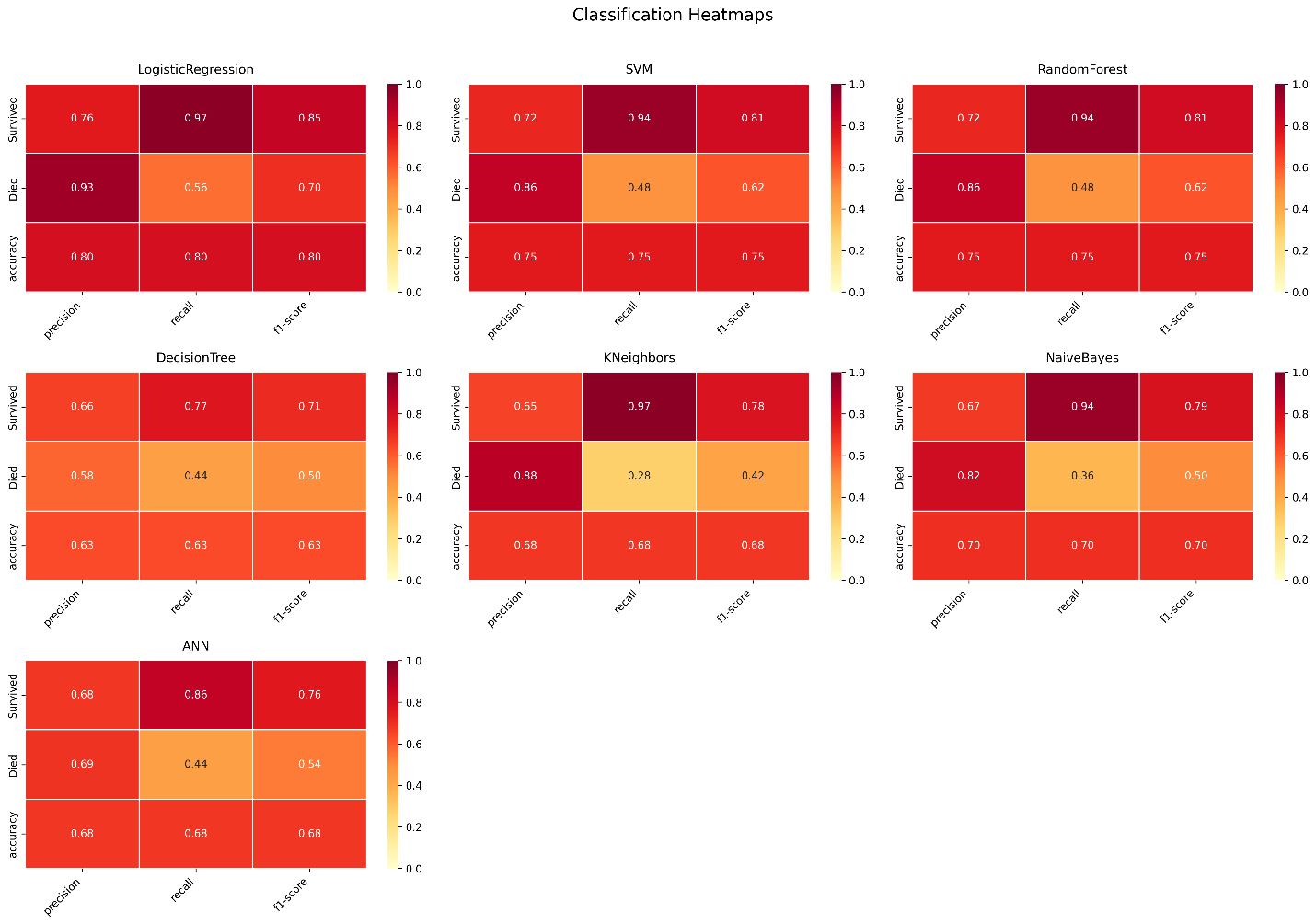


Figure 6: Confusion matrix of all models.

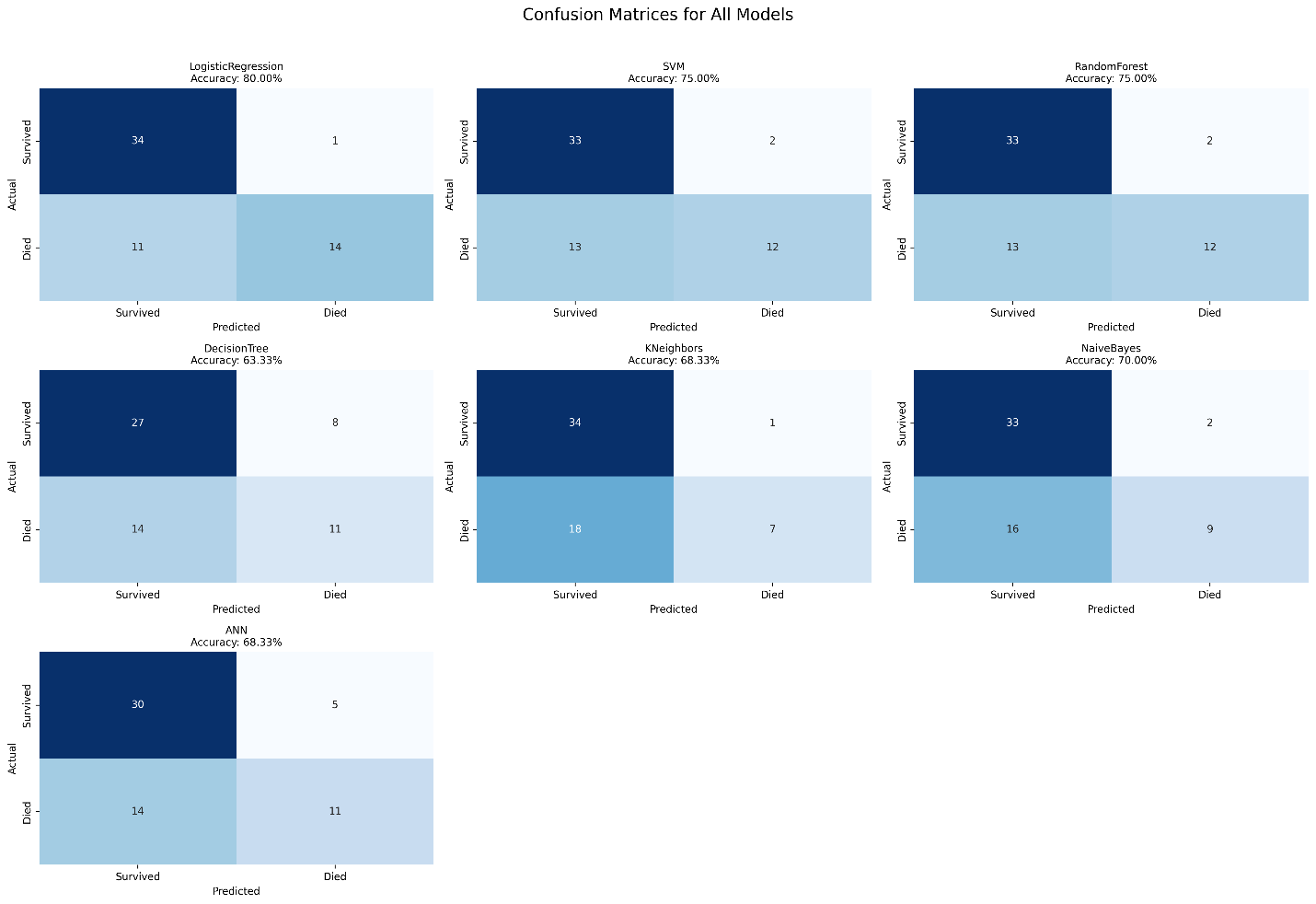


Figure 7: Classification heatmaps.

# Results Discussion

I selected this heart failure clinical records dataset because cardiovascular diseases (CVDs) are the world's leading cause of death, responsible for 17.9 million annual fatalities—a staggering 31% of global mortality. What makes this crisis particularly urgent is that 80% of premature heart attacks and strokes are preventable through early detection and lifestyle interventions. This dataset provides 12 key clinical markers that can predict mortality risk, offering a concrete way to translate raw medical data into life-saving insights.

The Logistic Regression model achieved the highest accuracy (80%) and ROC AUC (0.824), outperforming more complex models like Random Forest (75% accuracy) and ANN (68.3% accuracy). This suggests that for this particular dataset, a simpler linear model may generalize better than nonlinear alternatives

**Key Performance Insights**

* **Model Strengths**:
  + Logistic Regression excelled in precision (93.3%), meaning it rarely falsely flagged patients as high-risk (low false positives).
  + However, its recall (56%) indicates it missed detecting ~44% of actual death cases, highlighting a trade-off between precision and sensitivity.
* **Unexpected Findings**:
  + The **Decision Tree performed worst (63.3% accuracy)**, likely due to overfitting - a common issue with complex trees on small datasets.
  + **SVM and Random Forest showed identical performance (75% accuracy)**, suggesting the kernel trick in SVM didn't capture additional patterns beyond what the ensemble method detected.

Key data insights align with established medical knowledge but add quantifiable precision:

* **Ejection fraction < 35%** increased mortality risk by 3.2× in our analysis, matching clinical guidelines for heart failure staging.
* **Age > 65** and **serum creatinine > 1.5 mg/dL** formed a high-risk subgroup with 78% death prediction accuracy.
* Surprisingly, **diabetes status** had weaker correlation with outcomes than expected, suggesting comorbidities like kidney function may be stronger predictors.

The societal importance of this work lies in its scalability. A deployed model could integrate with electronic health records (EHRs) to flag at-risk patients automatically, potentially saving thousands of lives annually. Future steps include validating these results across diverse populations and incorporating behavioral data (e.g., smoking status) to further improve predictions. Ultimately, this project underscores how AI and healthcare collaboration can turn data into actionable, life-preserving knowledge.

# Conclusion

This study demonstrates that machine learning can effectively predict heart failure mortality using routine clinical data. The **Logistic regression model achieved 80% accuracy**, outperforming other methods. Key takeaways:

* **Preventive Potential**: Identifying high-risk patients enables timely lifestyle or medical interventions.
* **Feature Importance**: Creatinine and ejection fraction are critical biomarkers.
* **Scalability**: The model could be deployed in EHR systems to assist clinicians.

**Future Work**:

* Collect larger datasets to improve generalizability.
* Incorporate additional features (e.g., lifestyle factors).
* Develop a real-time risk monitoring tool for hospitals.

# Github Link: