# Header

Collin Lindsay

MSBD 566

Predictive Modeling and Analytics

December 11, 2025

Report - Heart Failure Clinical Records

# Project Description

The goal of this machine learning project was to identify which health factors are the main contributors of heart failure and ultimately death. There is a high likelihood that patient survival can be predicted using time-series data derived from heart failure patients in Pakistan.

# Data Description:

The dataset contains the medical records of 299 heart failure patients collected at the Faisalabad Institute of Cardiology and at the Allied Hospital in Punjab, Pakistan, during April - December 2015. The patients consisted of 105 women and 194 men aged 40-95 years old. All 299 patients had a medical history of left ventricular systolic dysfunction and had previously experienced heart failure.

Dictionary:

The dataset itself consisted of thirteen (13) clinical features:

- age: age of the patient (years)

- anaemia: decrease of red blood cells or hemoglobin (boolean)

- creatinine phosphokinase  (CPK): level of the CPK enzyme in the blood (mcg/L)

- diabetes: if the patient has diabetes (boolean)

- ejection fraction: percentage of blood leaving the heart at each contraction  (percentage)

- high blood pressure: if the patient has hypertension (boolean)

- platelets: platelets in the blood (kiloplatelets/mL)

- sex: woman or man (binary)

- serum creatinine: level of serum creatinine in the blood (mg/dL)

- serum sodium: level of serum sodium in the blood (mEq/L)

- smoking: if the patient smokes or not (boolean)

- time: follow-up period (days)

- [target] death event: if the patient died during the follow-up period (boolean)

Dataset Journal and Citation

https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-020-1023-5#Sec2

Davide Chicco, Giuseppe Jurman: "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone". BMC Medical Informatics and Decision Making 20, 16 (2020). https://doi.org/10.1186/s12911-020-1023-5

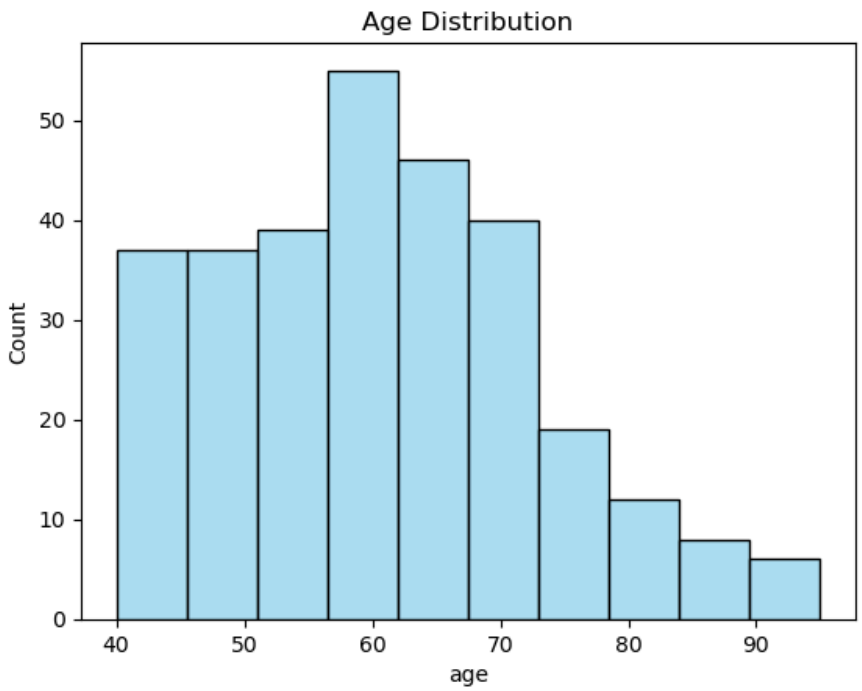


Figure 1: Age distribution of all 299 patients in the dataset. The top demographic consisted of patients ages 60-70.

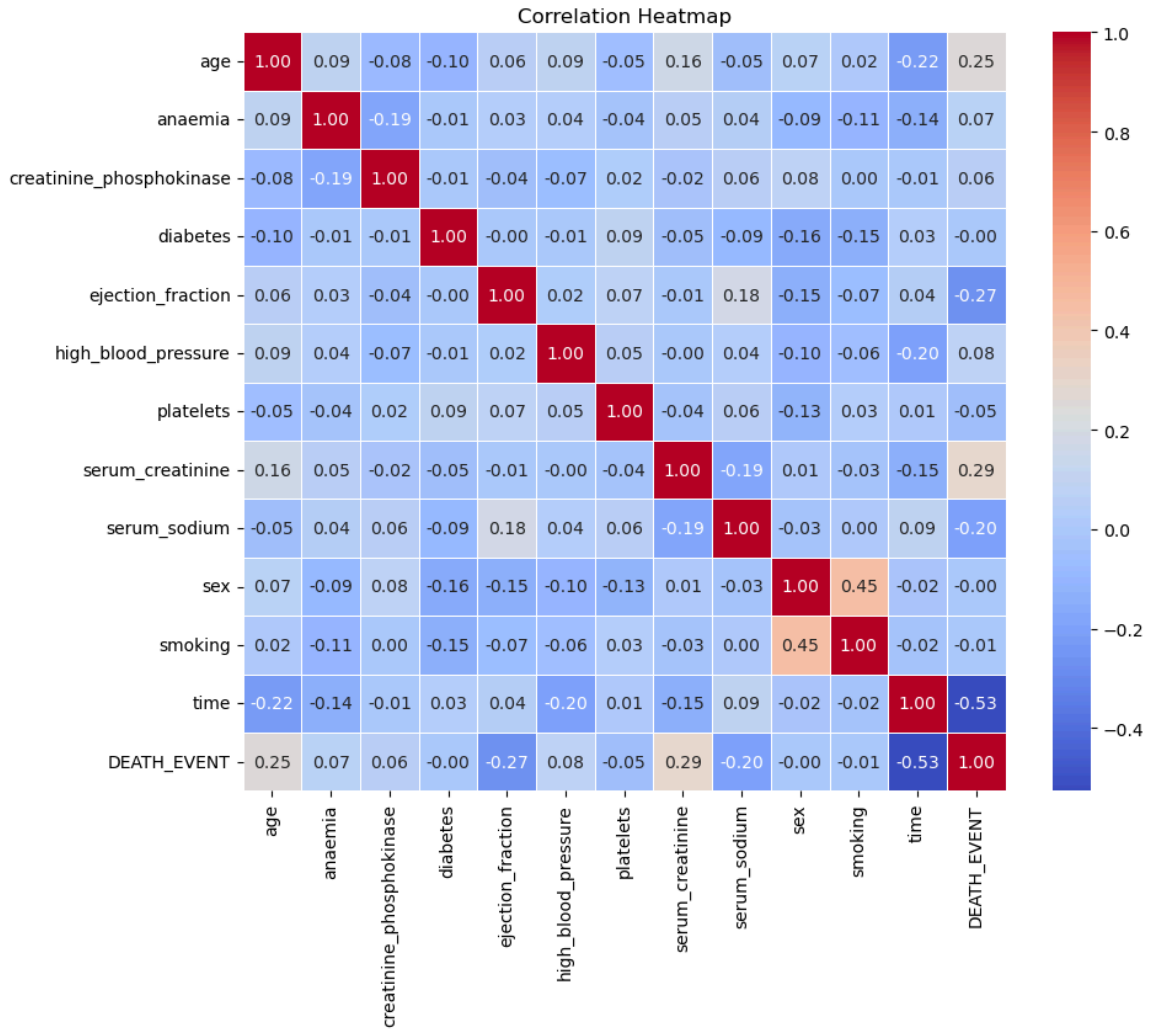


Figure 2: An initial EDA Multicollinearity analysis showing relationships between numeric features. The strong correlation is between sex and smoking (0.45), followed by serum_creatinine (0.29) and age and Death_Event (0.25).

# Methods and Analysis

The methods for this project used machine learning models to predict patient mortality (feature 'DEATH_EVENT') from clinical features in the heart failure dataset. The initial process began with data preprocessing. The dataset was imported into a Jupyter notebook, and all of the clinical features were

checked for missing values using the MSNO import. An initial EDA analysis showed no significant outliers in the data for any of the clinical features. To prevent skew during modeling, the target variable was removed from the DataFrame. Since all predictors were numeric, the features were standardized using a StandardScaler pipeline. The data set was split into training and test sets using an 80/20 stratified split, ensuring balanced representation of the target variable. The random state was also set to 42. When selecting models for machine learning, 12 common models and learning algorithms were implemented using the scikit-learn library. These models included 10 supervised models: Logistic Regression, Random Forest, Extra Trees, Gradient Boosting, HistGradientBoosting, AdaBoost, Support Vector Classifier (RBF kernel), K-Nearest Neighbors, Gaussian Naïve Bayes, Decision Tree, and 2 unsupervised models: Dimensionality Reduction, Principal Component Analysis, and Neural Network (MPLC Classifier). The supervised models were trained on the training set and evaluated on the test set using four key performance metrics: precision, accuracy, F1 score, and area under the receiver operating characteristic (ROC) curve (AUC). Next, feature importance was extracted using methods such as feature importance. Finally, the models were sorted by AUC from best to worst to assess performance, and a representative decision tree from the top performer, Random Forest, was plotted to provide interpretability of the model's decision-making process.

heart_failure_model_report

| Model | Precision | Accuracy | F1 | AUC | Top Predictors |
|---|---|---|---|---|---|
| Random Forest | 0.7857 | 0.8167 | 0.6667 | 0.8825 | time (0.368), serum_creatinine (0.146), ejection_fraction (0.135), platelets (0.077), age (0.076) |
| AdaBoost | 0.7647 | 0.8333 | 0.7222 | 0.8736 | time (0.400), platelets (0.200), serum_creatinine (0.100), ejection_fraction (0.080), creatinine_phosphokinase (0.080) |
| Logistic Regression | 0.7857 | 0.8167 | 0.6667 | 0.8588 | time (1.577), ejection_fraction (0.910), serum_creatinine (0.788), age (0.416), creatinine_phosphokinase (0.276) |
| Extra Trees | 0.7 | 0.75 | 0.4828 | 0.8549 | time (0.286), ejection_fraction (0.143), serum_creatinine (0.133), creatinine_phosphokinase (0.078), serum_sodium (0.077) |
| HistGradientBoosting | 0.8462 | 0.8333 | 0.6875 | 0.8485 | time (0.246), ejection_fraction (0.042), sex (0.018), age (0.015), serum_creatinine (0.014) |
| Gradient Boosting | 0.8 | 0.8333 | 0.7059 | 0.8447 | time (0.588), serum_creatinine (0.123), ejection_fraction (0.098), platelets (0.068), creatinine_phosphokinase (0.056) |
| SVC (RBF) | 0.7273 | 0.7667 | 0.5333 | 0.8447 | time (0.158), age (0.064), ejection_fraction (0.055), sex (0.016), serum_sodium (0.014) |
| GaussianNB | 0.5455 | 0.7 | 0.4 | 0.8293 | time (0.172), age (0.038), serum_sodium (0.030), ejection_fraction (0.029), anaemia (0.008) |
| KNN | 0.7143 | 0.7333 | 0.3846 | 0.7997 | time (0.157), ejection_fraction (0.104), age (0.068), serum_sodium (0.054), anaemia (0.032) |
| Decision Tree | 0.6 | 0.7333 | 0.5294 | 0.6637 | time (0.485), platelets (0.144), ejection_fraction (0.138), serum_creatinine (0.111), serum_sodium (0.033) |

Figure 3: a comprehensive table showing all model depictions and performance as well as top predictors sorted by highest predictor to least high predictor.

Regarding the unsupervised models, Principal Component Analysis (PCA) was used to address possible multicollinearity and simplify the model. The minimum number of main components needed to account for 95% of the variance was retained when PCA was fitted solely on the standardized training data. Both training and test sets were subjected to the ensuing transformation. The entire standardized dataset was used to train a separate two-component PCA model for visualization, allowing the data distribution to be displayed in two dimensions and colored by survival outcome.

The PCA-transformed training set was used to train a multilayer perceptron (MLP) classifier. The Rectified Linear Unit (ReLU) activation function was used in the neural network's two hidden layers, which included 64 and 32 neurons, respectively. To minimize overfitting, L2 regularization was used in conjunction with the Adam optimizer during training. By tracking performance on a validation subset derived from the training data, early halting was implemented. To ensure model generalizability, training was stopped after a specified number of iterations without improvement.

To aid in the interpretation of the models, two main visuals were produced. First, the differences in survival outcomes in the reduced feature space were shown using a two-dimensional PCA scatter plot. Second, a ROC curve provided an assessment of predictive ability by illustrating the trade-off between sensitivity and specificity for the trained neural network.

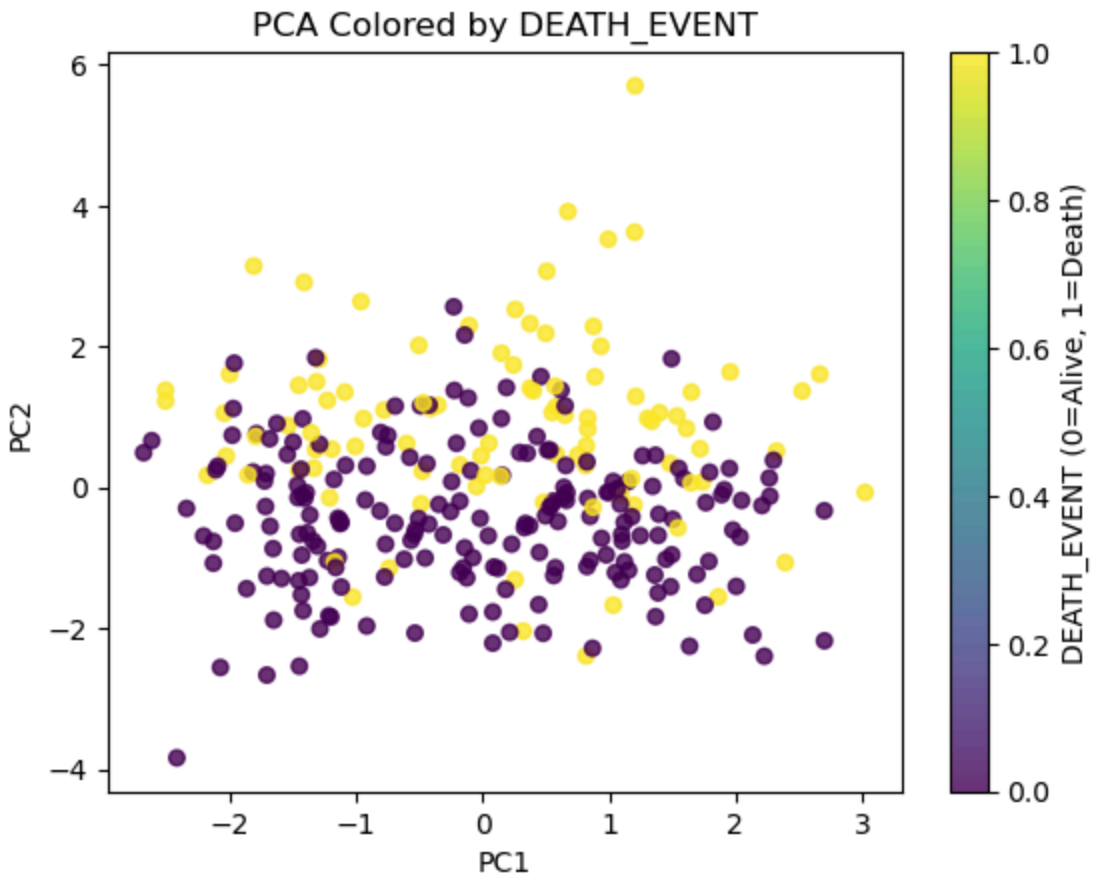


Figure 4: PCA components kept were 11, The total explained variance was identified as 0.958.

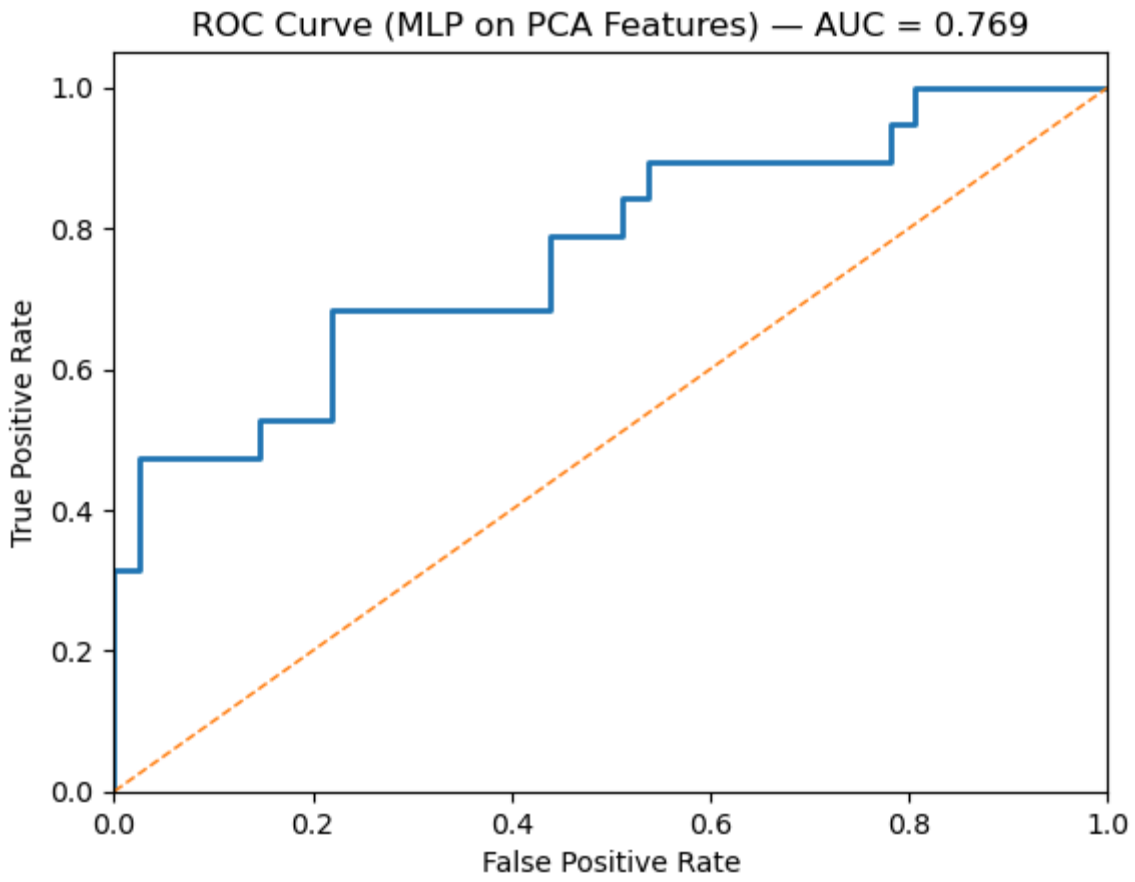


Figure 5: ROC curve illustrating the trade-off between sensitivity and specificity for the trained neural network.

# Simple Evaluation

After running all 10 models, the best-performing model was the Random Forest model. The AUC score was 0.88, indicating strong confidence. The strong Random Forest model performance indicates that patterns in the data reflect meaningful physiological relationships and that the data are highly reliable. Based on the top predictors, older patients with lower ejection fraction and high serum creatinine were more likely to experience a death event. In contrast, patients with better stabilized ejection fraction, lower

blood pressure, and lower age survived. The model's high AUC value supports that the risk factors form a reliable basis for predicting the survival outcome.

Overall, a trustworthy predictive framework for heart failure mortality was produced for the unsupervised models by combining a neural network classifier with PCA-based dimensionality reduction. The neural network successfully used the relevant structure that the reduced feature representation preserved in the data to model intricate, nonlinear relationships. These findings lend support to the viability of combining deep learning with dimensionality reduction for clinical outcome prediction.