Header

Collin Lindsay

MSBD 566

Predictive Modeling and Analytics

October 20, 2025

Report - Heart Failure Clinical Records

Project Description

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

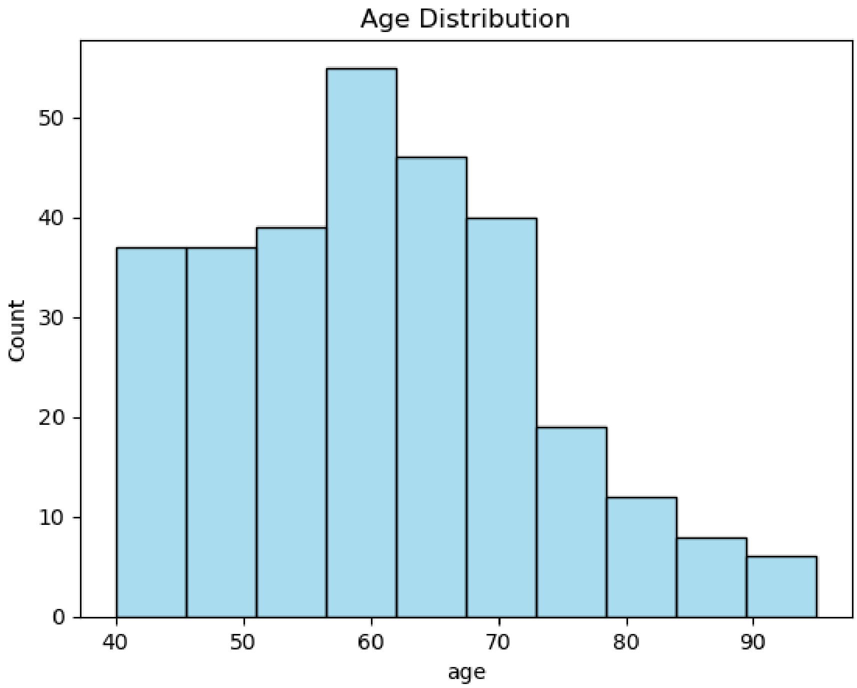
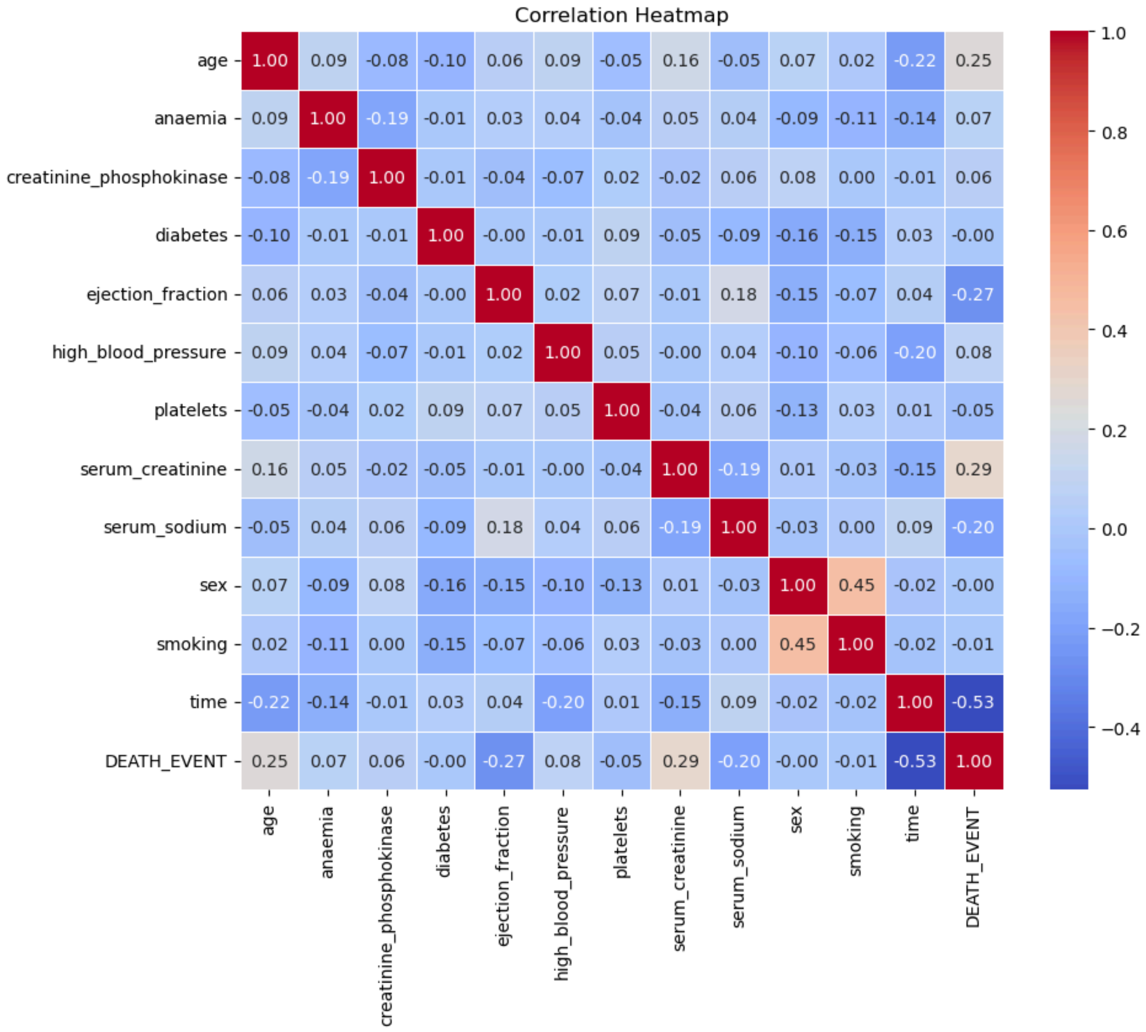


Figure 1: Age distribution of all 299 patients in the dataset. The top demographic consisted of patients ages 60-70.

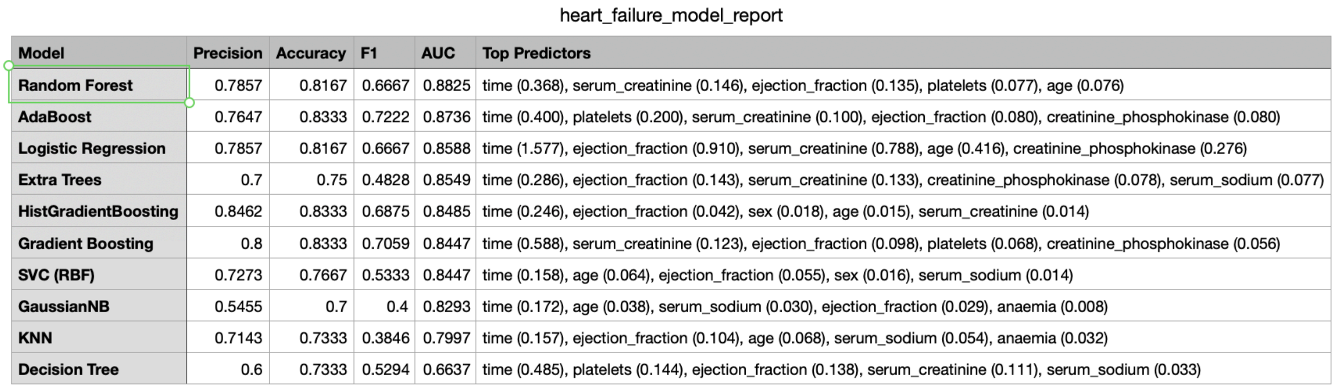


	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sodium	sex	smoking	time	DEATH_EVENT
age	1.00	0.09	-0.08	-0.10	0.06	0.09	-0.05	0.16	-0.05	0.07	0.02	-0.22	0.25
anaemia	0.09	1.00	-0.19	-0.01	0.03	0.04	-0.04	0.05	0.04	-0.09	-0.11	-0.14	0.07
creatinine_phosphokinase	-0.08	-0.19	1.00	-0.01	-0.04	-0.07	0.02	-0.02	0.06	0.08	0.00	-0.01	0.06
diabetes	-0.10	-0.01	-0.01	1.00	-0.00	-0.01	0.09	-0.05	-0.09	-0.16	-0.15	0.03	-0.00
ejection_fraction	0.06	0.03	-0.04	-0.00	1.00	0.02	0.07	-0.01	0.18	-0.15	-0.07	0.04	-0.27
high_blood_pressure	0.09	0.04	-0.07	-0.01	0.02	1.00	0.05	-0.00	0.04	-0.10	-0.06	-0.20	0.08
platelets	-0.05	-0.04	0.02	0.09	0.07	0.05	1.00	-0.04	0.06	-0.13	0.03	0.01	-0.05
serum_creatinine	0.16	0.05	-0.02	-0.05	-0.01	-0.00	-0.04	1.00	-0.19	0.01	-0.03	-0.15	0.29
serum_sodium	-0.05	0.04	0.06	-0.09	0.18	0.04	0.06	-0.19	1.00	-0.03	0.00	0.09	-0.20
sex	0.07	-0.09	0.08	-0.16	-0.15	-0.10	-0.13	0.01	-0.03	1.00	0.45	-0.02	-0.00
smoking	0.02	-0.11	0.00	-0.15	-0.07	-0.06	0.03	-0.03	0.00	0.45	1.00	-0.02	-0.01
time	-0.22	-0.14	-0.01	0.03	0.04	-0.20	0.01	-0.15	0.09	-0.02	-0.02	1.00	-0.53
DEATH_EVENT	0.25	0.07	0.06	-0.00	-0.27	0.08	-0.05	0.29	-0.20	-0.00	-0.01	-0.53	1.00

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 were developed using machine learning models to predict patient mortality (feature ‘DEATH_EVENT’) from clinical features within the heart failure dataset. The initial process began with data preprocessing. The dataset was imported into a Jupyter notebook, all of the clinical features were checked for missing values using the MSNO import. An initial EDA analysis showed that there were no significant outliers in the data for any of the clinical features present. To prevent skew during modeling, the target variable was removed from the DataFrame. Since all the predictors were numeric, the features were standardized using a standard scaler pipeline. The data set was then split into training and testing sets using an 80/20 stratified split to enter your balance representation of the target variable. When selecting models for machine learning, 10 common supervised models and learning algorithms were implemented using the scikit-learn library. These models included Logistic Regression, Random Forest, Extra Trees, Gradient Boosting, HistGradientBoosting, AdaBoost, Support Vector Classifier (RBF kernel), K-Nearest Neighbors, Gaussian Naïve Bayes, and Decision Tree. Its model was then trained on the training set and evaluated on the test set using four key performance metrics: precision, accuracy, F1 score, an area under the curve (AUC). Next feature importance was extracted using the models built in methods such as feature importance. Finally, the models were sorted from best to worst AUC 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.

Simple Evaluation

After running all ten models, the model with the best performance 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 there is a strong reliability in the data. Based on the top predictors, older patients with lower ejection fraction, 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.