()

Matters of the Heart: A Data-Driven Dive into Heart Failure and its Predictors

By Alix CHARPENTIER and Clothilde DUGROS

November 8, 2024

ABSTRACT

Our hospital's cardiology department is conducting a study to pinpoint factors that can foretell deadly heart failure. By doing so, the department will be better prepared to identify patients at risk and provide them with adequate care. Age, ejection fraction and blood creatinine levels were found to be the most significant risk factors for mortality in heart failure patients in our sample data.

1. Introduction

This study is focused on evaluating the key factors contributing towards increased risk of mortality among heart failure patients. After cleaning and validating the received data, we studied the sample of patients. We looked at their characteristics, and how these relate to each other. We then shifted our attention to the main objective of this study, and looked at the mortality rate among the sample, the pattern of survival using graphical representations, and regressions to pinpoint the foretelling factors that most accurately predict mortality risk. We provided some additional robustness check using different machine learning models.

The research question we address is well established in the literature and has been extensively explored in previous studies. We will attempt to test our data using models commonly used in this context (Cox regressions).

Before diving into the analysis, we checked and cleaned our data. We started by visualising the initial observations to get a sense of the structure of the dataset. After describing the size of the dataset, we checked the types of variables and recoded some where necessary. Data cleaning also included checking for missing values or duplicates (of which there were none). We also clarified the meaning of the dummies variables. In addition, we created new variables from existing ones, such as categorising age groups based on quartiles. For a better understanding of the data, a summary statistic of the dataset is provided (see Table 1 below and Table 2 in the Annex section).

2. Charting the Patients' Landscape

2.1. Metrics' Distributions

Current study is based on 299 patients. Follow up time was ranging from 4 to 285 days. About 35% of patients are female and 65% are male. One third of our sample are smokers, about 43% are anemic (low level of red blood cells), 42% have diabetes (type unknown) and 45% have low blood pressure (Figure 1).

Creatine phosphokinase (CPK) is an enzyme found in muscle tissue, including the heart, that play a role in muscles' energy. A high level of this enzyme typically means an injury to a muscle. A normal level of creatinine ranges from 10 to 120 micrograms per liter (mcg/L), which is what the density bar plot shows (Figure 3): more than one third of the patients are in this range. The distribution is, to a great extent, positively skewed, as extreme higher values outliers can be observed. Specifically, the individual with a 8000 mcg/L of CPK is curious to say the least, suggesting either an error in the data, or a patient having a heart attack linked with a massive tear in the heart tissue when his metrics were taken.

Creatinine is a normal waste product of the body. It is produced when a muscle is used, then filtered out of the bloodstream by the kidneys and excreted in the urine. A normal level is between 0.6 and 1.3 mg/dL. The distribution is again heavily skewed to the left, with outliers to the right, with extreme values of more than 8 mg/dL. In heart failure, reduced heart function can lead to reduced blood flow to the kidneys. As a result, the kidneys struggle to excrete creatinine effectively in the urine, causing it to build up and lead to elevated blood creatinine levels. That may explain why some patients have high blood creatinine levels compared to normal. This idea emphasizes the caution needed with our models: we will show correlation, not causation. If the coefficient of creatinine proves to be significant, it could indicate that heart problems are indeed leading to the abnormal levels. Distinguishing what influences what and in which direction cannot be demonstrated with our current models and descriptive statistics.

The normal range for blood **sodium** levels (salt) is about 135 to 145 milliequivalents per liter (mEq/L), which our sample is

coherent with. The distribution seems to follow a normal distribution, slightly negatively skewed, with some outliers at 115 mEq/L, patients with what is called in these cases, severe hyponatremia.

The **platelets**' primary job is to stop the bleeding if you are injured. A normal number of platelets in the blood is 150,000 to 400,000 kiloplatelets per microliter (mcL), a range that explains the spread normal distribution. A couple of outliers can still be observed, at 800,000 kiloplatelets per microliter.

Ejection fraction is a measurement of the percentage of blood leaving the heart each time it contracts. We usually state that a normal ejection fraction in a healthy heart ranges from 50% to 70%. In the case of our patient sample, it seems like many have a lower percentage than the normal range. We can speculate early on from this observation, that a low ejection fraction might be a good predictor of heart failure.

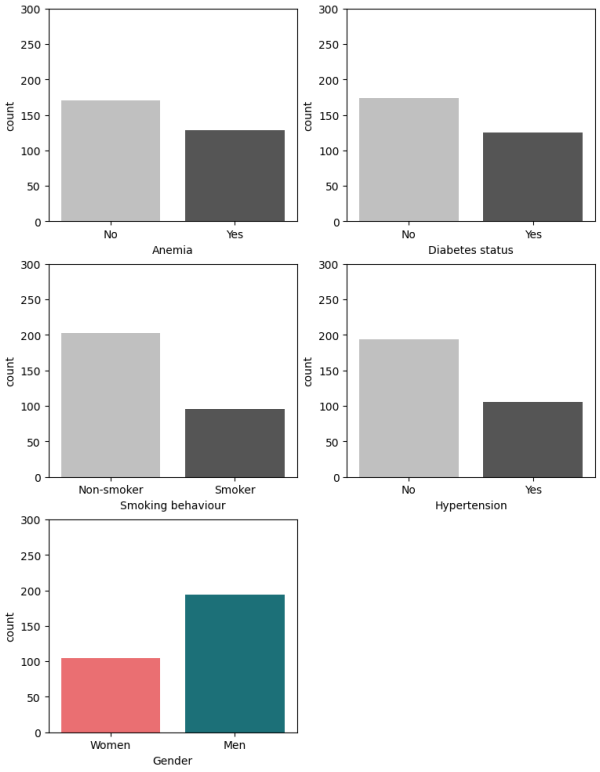


Fig. 1: Distribution of categorical variables

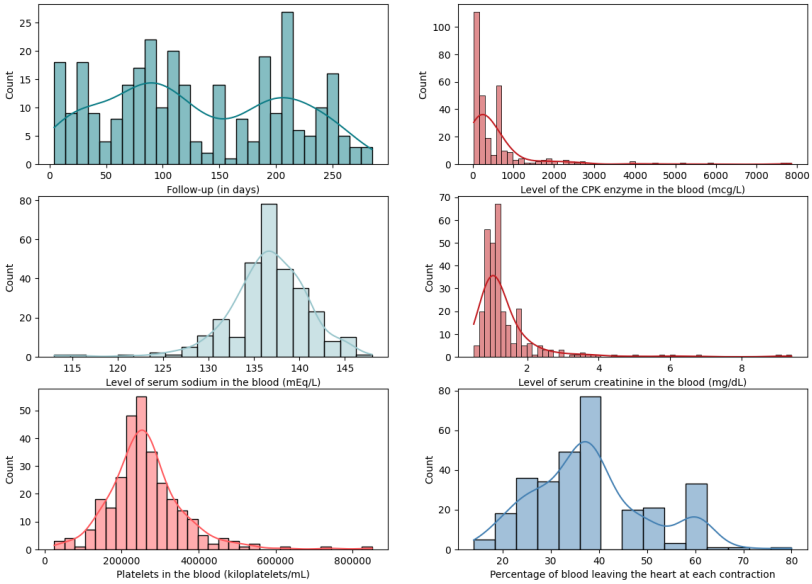


Fig. 2: Distribution of continuous variables

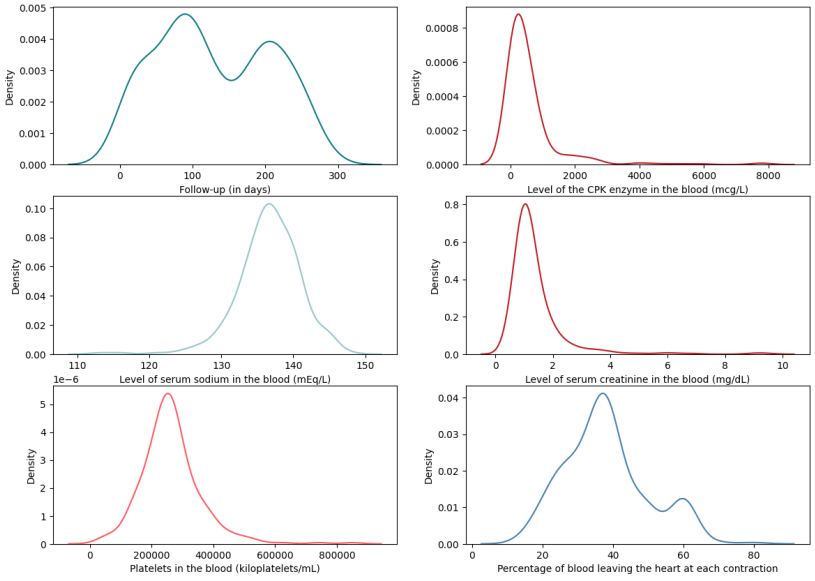


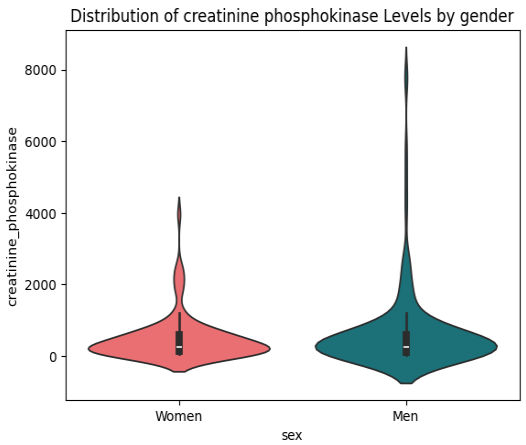
Fig. 3: Smoothed distribution of continuous variables

2.2. *Combinations*

Our dataset is designed to evaluate predictors of heart failure, yet its wealth of health metrics, combined with information on lifestyle habits and sex, offers the potential to explore patterns in comorbidities. Beyond enhancing insights, it's crucial to examine whether some predictors may be interdependent—a phenomenon known as multicollinearity.

2.2..1. *Health Metrics with Age and Gender*

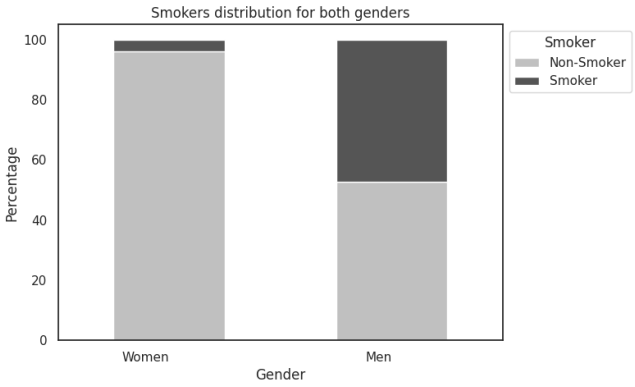
The typical levels of various components often vary significantly depending on the patient's sex. This is the case, for instance, of creatinine, since this chemical component is linked to muscle mass.



In our sample, the highest point density in the dataset for creatinine phosphokinase levels is approximately the same for men and women, as are the quantiles and the median. However, the maximum data value for men is higher than for women (this may be not significiant and due to an outlier, as a CPK value of 8000 is quite abnormal).

These gender effects can be extended to smoking. According to the *crosstab*, smokers are mainly found among male patients: 47.4% of men are smokers, while only 3.8% of women smoke. However, since more males are present in the sample, these findings need to be taken with caution.

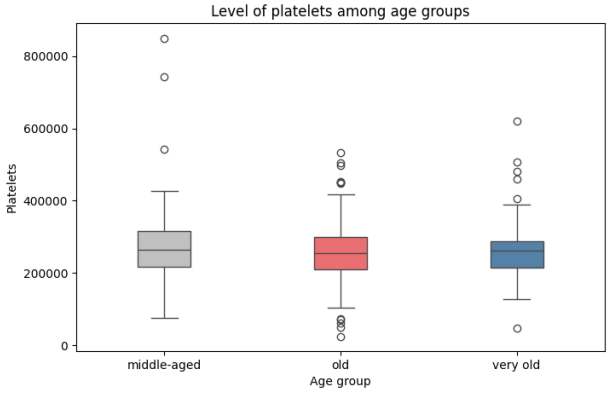
smoking sex	0	1
0	96.2	3.8
1	52.6	47.4
All	67.89	32.11



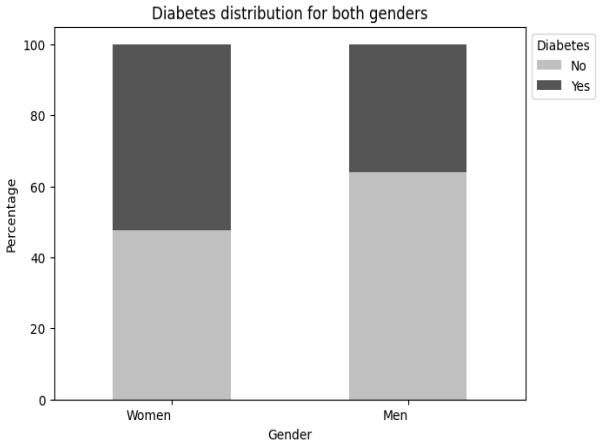
We can also examine how platelet levels vary across age groups. We define three age categories: 'middle-aged' (ages below the first quartile), 'old' (ages between the first and third quartile), and 'very old' (ages above the third quartile). According to the boxplots, the distribution of platelet levels appears to be roughly the same across these age groups.

		Dead (n = 96)	**Censored (n = 203)**
Continuous Variables	Creatinine (mg/dL)	1.840000	1.180000
	Sodium (mEq/L)	135.380000	137.220000
	CPK (mcg/L)	670.200000	540.050000
	Age (years)	65.220000	58.760000
	Platelets (k/mL)	256381.040000	266657.490000
	EF (%)	33.470000	40.270000
Categorical Variables	Male	62 (65%)	132 (65%)
	Smoking	30 (31%)	66 (33%)
	Diabetes	40 (42%)	85 (42%)
	Low BP	39 (41%)	66 (33%)
	Anemia	46 (48%)	83 (41%)

Table 1: Summary of Continuous and Categorical Variables by Survival Status



Finally, we look at how diabetes status varies by gender. The proportion of men and women with diabetes differs: 52.3% of women in our sample have diabetes compared to 36.1% of men. It would be interesting to distinguish between people with type I and type II diabetes to further comment.



2.2..2. Combinations of Diseases

Correlations plots (Figure 4) are a way of visually see associations between two variables. In addition, we constructed an UpSet plot (Figure 5), which provides the number of observations with multiple chosen attributes (combining smoking, high ejection fraction and diabetes for example).

From the correlation plot, we can, for instance, see that the death event is not correlated with diabetes, nor with gender or smoking behaviour. This is something that we will also show in the next section. However, it seems positively correlated with age, the level of serum creatinine in the blood and, to a lesser extent, to hypertension. It is negatively correlated with the percentage of blood leaving the heart at each contraction and the level of serum sodium in the blood. This is not informational yet on significance, but provides some ideas on the "direction" of the relationships.

When looking at the combinational possibilities of smoking, high blood pressure, diabetes, and anemia, the most common combination found is the two latest : diabetes and anemia. This finding in our data can also be found in the litterature (Thomas S, Rampersad M., 2004). Diabetes can damage the kidney (the high sugar level damages the blood vessels of the organ). When kidneys fail, they stop producing a hormone necessary to trigger the production of red blood cells (anemia). In addition, according to the *Centers for Disease Control and Prevention*, many people with diabetes also develop high blood pressure, leading to failure of the kidney and its consequences. This might be why we see a number of patients with diabetes, high blood pressure and anemia. However, our sample remain very small, which might explain why associations do not show up in the correlation plot. For instance, out of 299, 17 individuals have the three illnesses (high blood pressure, diabetes, and anemia). This is five percent of our sample. Therefore, while we can recover some of the litterature's findings, we cannot make hard conclusions on commorbidities either.

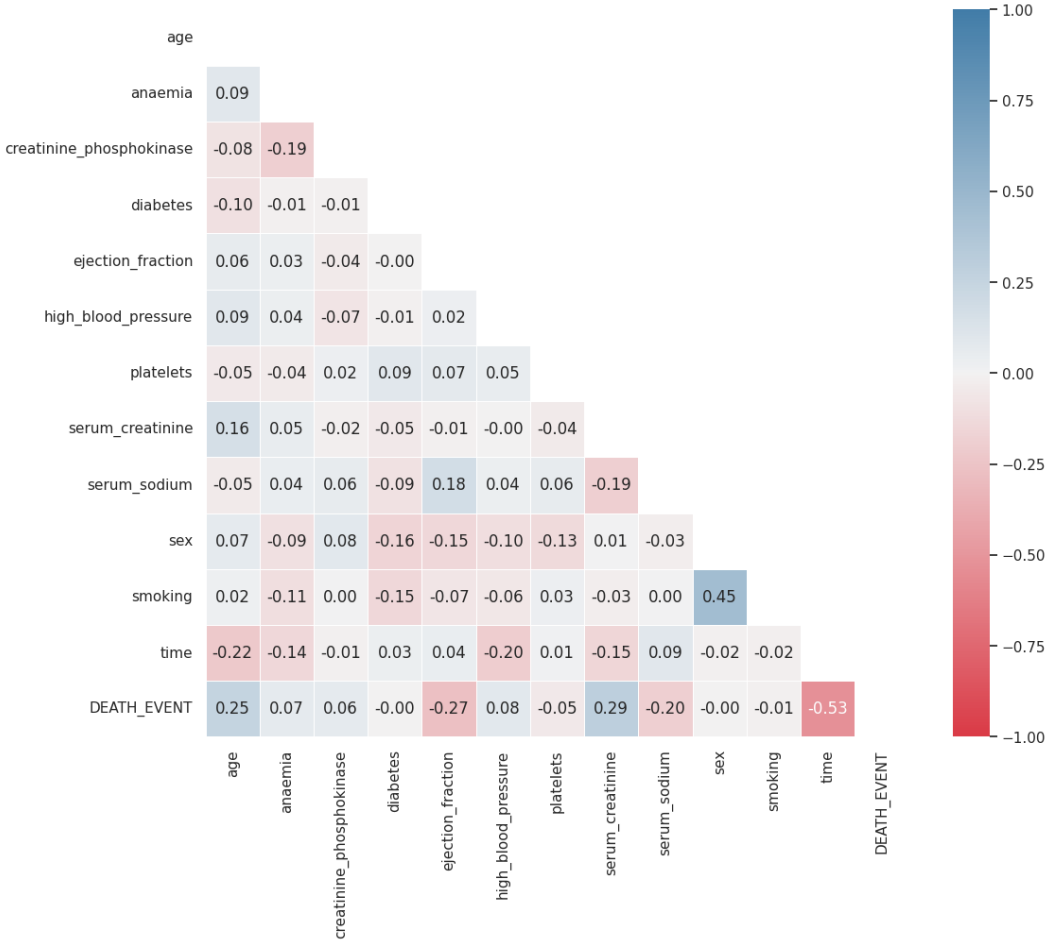


Figure 4: Correlation Heatmap of all the variables

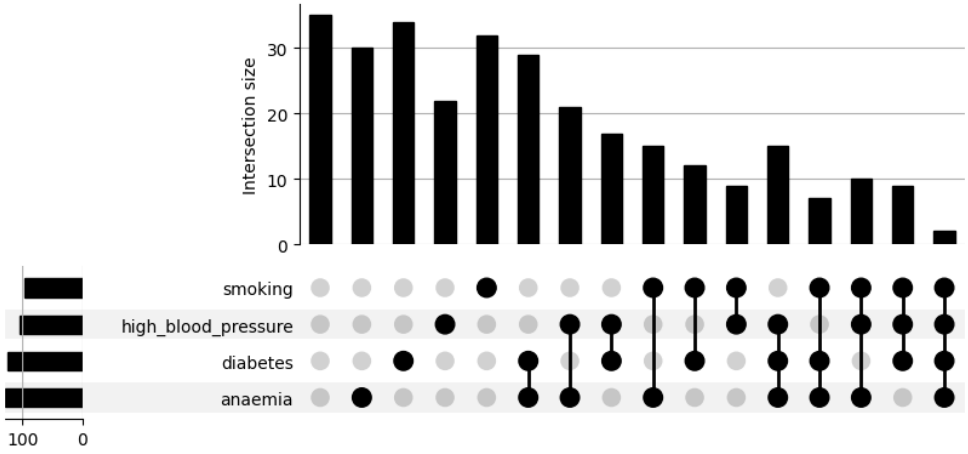


Figure 5: UpSet Plot

3. Mortality Rates

The mortality rate due to heart failure among the study participants is 32.1%.

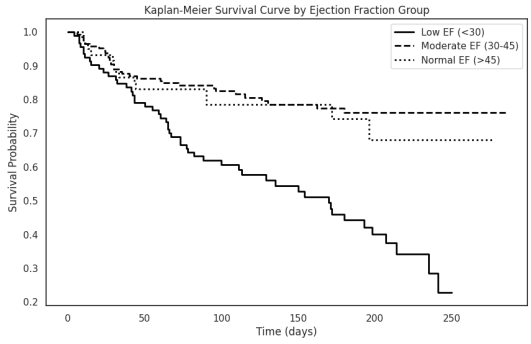
We use *crosstabs* to calculate mortality rates across different sub-groups. For example:

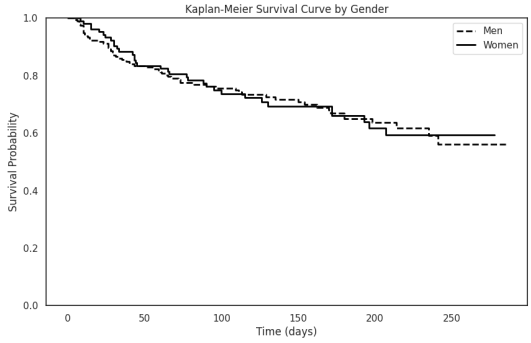
DEATH_EVENT	0	1
sex		
0	67.6	32.4
1	68.0	32.0
All	67.9	32.1

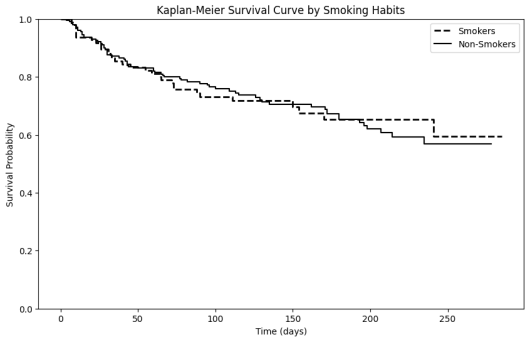
The overall mortality rates for women (32.4%) and for men (32.0%) are quite similar. Moreover, the mortality rates for smokers (31.25%) and for non-smokers (32.51%) are also quite similar. The same non-difference of mortality rates are found for diabetes. However, the mortality rate for people with hypertension (37.14%) is higher than for others (29.38%).

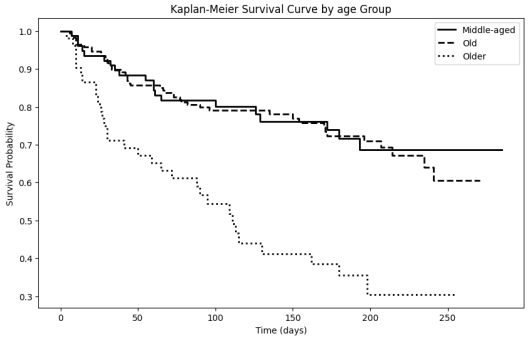
3.1. Survival Patterns

The Kaplan-Meier curves display the probability of the event not occuring (death) as a function of time. Here, we set on visually comparing survival curves for different groups. For instance, at day 60, the probability to be alive is 60% for low EF, while significantly higher for other groups.









Therefore, we can see that those with a low ejection fraction are associated with poorer survival outcomes compared to patients with moderate or average percentages as time passes. In addition, patients aged 70 and more also have poorer survival probabilites over time compared to younger patients. This graph suggests that the effect of age on heart failure is driven by the oldest patients, since we can see that patients in the other two categories have similar trends. Finally, the different categories in gender and smoking behaviors groups are no differences in terms of survival probabilities across time.

These curves are great visual tools, but are univariate analysis used mainly with categorical variables, which is why we need to introduce other models.

4. Identifying Foretelling Factors of Mortality Risk

We start with a *logit model* because the dependent variable is binary (1 if the patient is deceased, 0 otherwise). We include all variables from the dataset except for DEATH_EVENT, the variables we created, and time (as it does not directly explain death outcome; age already captures this factor). Regression tables for logistic regression (Table 3) and marginal effects (Table 4) are provided in the Annex section. Interpreting coefficients from the logistic regression itself does not directly reveal changes in probability, but marginal effects do. This is why we compute them. With marginal effects, we obtain the change in the probability of the outcome associated with a one-unit change in each predictor variable, holding all other variables constant. This allows us to interpret both the sign and magnitude of the effect,

whereas with the initial regression, we could only comment on the sign. The significant variables (at a 5% level) are:

- **Age**: if age increases by 1 year, the probability of dying due to heart failure increase by less than 1%. This seems plausible.
- **Level of the CPK enzyme in the blood**: while the effect is small, it is significant and positive. The higher the level of CPK enzyme in the blood, the higher the mortality risk. It is consistent with the fact that CPK levels rise during acute events like heart attacks
- **Percentage of blood leaving the heart at each contraction**: an increase of 1% in the ejection fraction is associated with a decrease of 1.1% in the probability of dying due to heart failure. It seems plausible since it is an indicator of heart function.
- **Level of serum creatinine in the blood**: Higher levels of creatinine in the blood are associated with higher mortality risk.
- **Level of serum sodium in the blood**: Higher levels of serum sodium in the blood are associated with lower mortality risk.

In a *Cox proportional hazards model*, the focus is on understanding how various factors (covariates) affect the hazard rate— the risk of suffering the event of interest, which is death in our case. The model allows us to see the role a variable plays in the risk of event while controlling for the others.

Upon the p-values evaluations, age, the ejection fraction and the serum creatinine are found to be the most significant variable explaining death by heat failure (see Table 5 in the Annex section). A negative coefficient suggests a protective effect - there is a decreased risk of death for such variables, like the **ejection fraction** : for each unit increase in ejection fraction, the risk of death decreases by 5%.

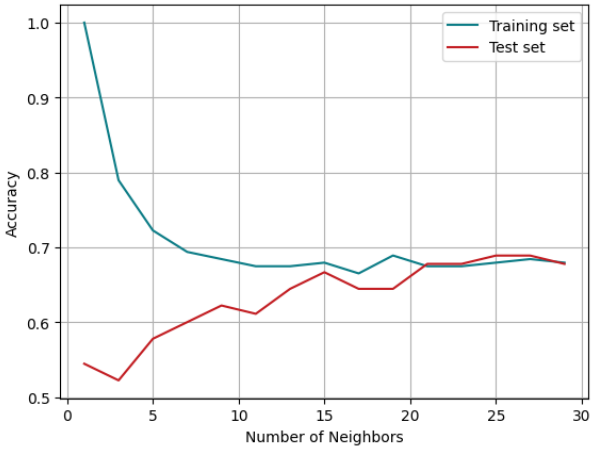
There is a 5% increase in the risk of death for each additional year of **age**, and for each unit increase in **serum creatinine**, the risk of hazard is increased by 38%. However, it is again essential to remember that we are not looking at causality. Therefore, the effect could be that heart failure can impair blood flow to the kidneys, like we previously discussed.

Anemia and **high blood pressure** are both weakly positively significiant (at 5% but not at 1%). Being anemic is associated with a 58% increased in the death risk, while having a high blood pressure increases the risk by 61%.

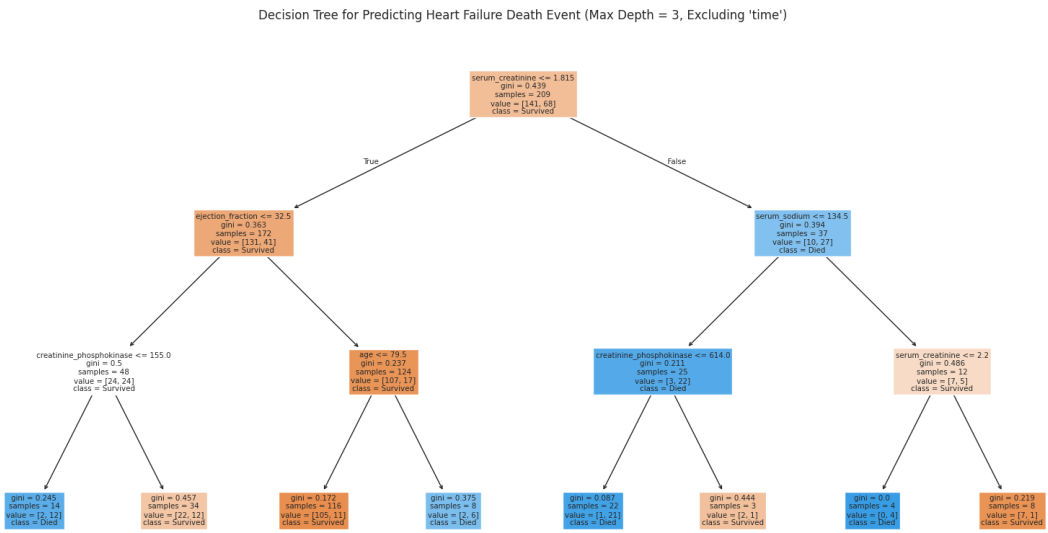
Smoking, the **creatinine phosphokinase**, **diabetes**, **platelets** and **serum sodium** are either not significant, or weakly but without any noticeable effect on the variable of interest. The result concerning smoking is interesting, as the collective saying seems to be that smoking can participate and speed up the clogging of the coronary arteries (atherosclerosis), a cause of heart failure. Some studies suggest that it also affect the ability of the vessel to expand when more oxygen is necessary.

5. Predicting Death Event using Machine Learning

We first built a K-Nearest Neighbour model, although it is considered as a very basic and (too) simplistic model. The KNN algorithm identifies 'k' observations that are similar or nearest to the new record being predicted and then uses the most common class (either deceased or not) among those k observations as the predicted output. We tested the KNN model on our dataset with different values of k. Based on the accuracy level, the optimal model is with k = 15, achieving an accuracy of 0.67 on the test set. The performance of KNN is very sensitive to the choice of k; for example, low values of k tend to overfit.



We then built a decision tree with a maximum depth of 3 nodes. We chose this level to make the results easily interpretable for clinical decision-making. The first higher splits indicate that ejection fraction and serum creatinine are critical features for prediction. Low ejection fraction and high serum creatinine levels are associated with a higher risk of death. Then, in lower nodes, age is also used to split further the data, which provides a robustness check of the cox regressions results. However, CPK is also used in these lower nodes, whereas we did not see any results in our previous regression. As in the cox regressions, gender does not seem to be associated with heart failure. The litterature does not seem to be in agreement, as some study find a significant relationship (concluding that males are more prone to heart disease than females) (Barlera S, Tavazzi L, et al, 2023), while others (Ahmad T, Munir A et al, 2017), like ours, do not. The test set accuracy of this model is 0.69, which could probably be improved with different parameters or alternative tree models like random forests.



Visualisation of the optimal decision tree

6. Conclusion

The findings underscore significant correlations between clinical metrics and heart failure, mainly ejection fraction and serum creatinine. A low ejection fraction is associated with poorer survival rates, while high level of serum creatinine a higher mortality risk. The study also highlights that some conditions like anemia and diabetes coexist in many patients, leading to compounded risk. Finally, the application of machine learning techniques offer a promising future for reseach on heart failure.

7. Bibliography

Ahmad T, Munir A, Bhatti SH, Aftab M, Raza MA (2017) Survival analysis of heart failure patients: A case study. PLoS ONE 12(7): e0181001. https://doi.org/10.1371/journal.pone.0181001

Couissi, A., Haboub, M., Hamady, S., Ettachfini, T., Habbal, R. (2024). Predictors of mortality in heart failure patients with reduced or mildly reduced Ejection Fraction: The CASABLANCA HF Study. The Egyptian heart journal : (EHJ) : official bulletin of the Egyptian Society of Cardiology, 76(1), 5. https://doi.org/10.1186/s43044-024-00436-y

Thomas, S., Rampersad, M. Anaemia in diabetes. Acta Diabetol 41 (Suppl 1), s13–s17 (2004). https://doi.org/10.1007/s00592-004-0132-4

8. Appendix

	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets
count	299.0	299.0	299.0	299.0	299.0	299.0	299.0
mean	60.8	0.4	581.8	0.4	38.1	0.4	263358.0
std	11.9	0.5	970.3	0.5	11.8	0.5	97804.2
min	40.0	0.0	23.0	0.0	14.0	0.0	25100.0
25%	51.0	0.0	116.5	0.0	30.0	0.0	212500.0
50%	60.0	0.0	250.0	0.0	38.0	0.0	262000.0
75%	70.0	1.0	582.0	1.0	45.0	1.0	303500.0
max	95.0	1.0	7861.0	1.0	80.0	1.0	850000.0

	serum_creatinine	serum_sodium	sex	smoking	time	DEATH_EVENT
count	299.0	299.0	299.0	299.0	299.0	299.0
mean	1.4	136.6	0.6	0.3	130.3	0.3
std	1.0	4.4	0.5	0.5	77.6	0.5
min	0.5	113.0	0.0	0.0	4.0	0.0
25%	0.9	134.0	0.0	0.0	73.0	0.0
50%	1.1	137.0	1.0	0.0	115.0	0.0
75%	1.4	140.0	1.0	1.0	203.0	1.0
max	9.4	148.0	1.0	1.0	285.0	1.0

Table 2: Summary Statistics for numerical variables

Model:	Logit	Method:	MLE
Dependent Variable:	DEATH_EVENT	Pseudo R-squared:	0.213
Date:	2024-11-04 21:30	AIC:	317.4783
No. Observations:	299	BIC:	358.1832
Df Model:	10	Log-Likelihood:	-147.74
Df Residuals:	288	LL-Null:	-187.67
Converged:	1.0000	LLR p-value:	5.3226e-13
No. Iterations:	6.0000	Scale:	1.0000

	Coef.	Std.Err.	z	P> |z|	[0.025	0.975]
age	0.0575	0.0130	4.4173	0.0000	0.0320	0.0830
anaemia	0.3955	0.2991	1.3222	0.1861	-0.1908	0.9817
creatinine_phosphokinase	0.0003	0.0001	1.9798	0.0477	0.0000	0.0006
diabetes	0.1966	0.2937	0.6694	0.5032	-0.3791	0.7724
ejection_fraction	-0.0710	0.0149	-4.7660	0.0000	-0.1002	-0.0418
high_blood_pressure	0.4320	0.3052	1.4157	0.1569	-0.1661	1.0302
platelets	-0.0000	0.0000	-0.4132	0.6794	-0.0000	0.0000
serum_creatinine	0.6993	0.1753	3.9890	0.0001	0.3557	1.0430
serum_sodium	-0.0217	0.0076	-2.8577	0.0043	-0.0365	-0.0068
sex	-0.3580	0.3480	-1.0285	0.3037	-1.0401	0.3242
smoking	0.1361	0.3469	0.3922	0.6949	-0.5439	0.8160

Table 3: Results of Logit model

Dep. Variable:	DEATH_EVENT		**Method:**	dydx		
At:	overall					
	dy/dx	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
age	0.0094	0.002	5.000	0.000	0.006	0.013
anaemia	0.0646	0.048	1.335	0.182	-0.030	0.159
creatinine_phosphokinase	4.601e-05	2.27e-05	2.024	0.043	1.46e-06	9.05e-05
diabetes	0.0321	0.048	0.671	0.502	-0.062	0.126
ejection_fraction	-0.0116	0.002	-5.492	0.000	-0.016	-0.007
high_blood_pressure	0.0706	0.049	1.432	0.152	-0.026	0.167
platelets	-1.082e-07	2.62e-07	-0.413	0.679	-6.21e-07	4.05e-07
serum_creatinine	0.1142	0.026	4.426	0.000	0.064	0.165
serum_sodium	-0.0035	0.001	-3.012	0.003	-0.006	-0.001
sex	-0.0585	0.057	-1.034	0.301	-0.169	0.052
smoking	0.0222	0.057	0.393	0.695	-0.089	0.133

Table 4: Marginal effects

covariate	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	cmp to	z	p	-log2(p)
age	0.05	1.05	0.01	0.03	0.06	1.03	1.07	0.00	4.98	0.00	20.56
anaemia	0.46	1.58	0.22	0.04	0.89	1.04	2.42	0.00	2.12	0.03	4.89
creatinine_phosphokinase	0.00	1.00	0.00	0.00	0.00	1.00	1.00	0.00	2.23	0.03	5.26
diabetes	0.14	1.15	0.22	-0.30	0.58	0.74	1.78	0.00	0.63	0.53	0.91
ejection_fraction	-0.05	0.95	0.01	-0.07	-0.03	0.93	0.97	0.00	-4.67	0.00	18.35
high_blood_pressure	0.48	1.61	0.22	0.05	0.90	1.05	2.46	0.00	2.20	0.03	5.17
platelets	-0.00	1.00	0.00	-0.00	0.00	1.00	1.00	0.00	-0.41	0.68	0.56
serum_creatinine	0.32	1.38	0.07	0.18	0.46	1.20	1.58	0.00	4.58	0.00	17.68
serum_sodium	-0.04	0.96	0.02	-0.09	0.00	0.91	1.00	0.00	-1.90	0.06	4.12
sex	-0.24	0.79	0.25	-0.73	0.26	0.48	1.29	0.00	-0.94	0.35	1.53
smoking	0.13	1.14	0.25	-0.36	0.62	0.70	1.86	0.00	0.51	0.61	0.72

Table 5: Results of Cox Regressions