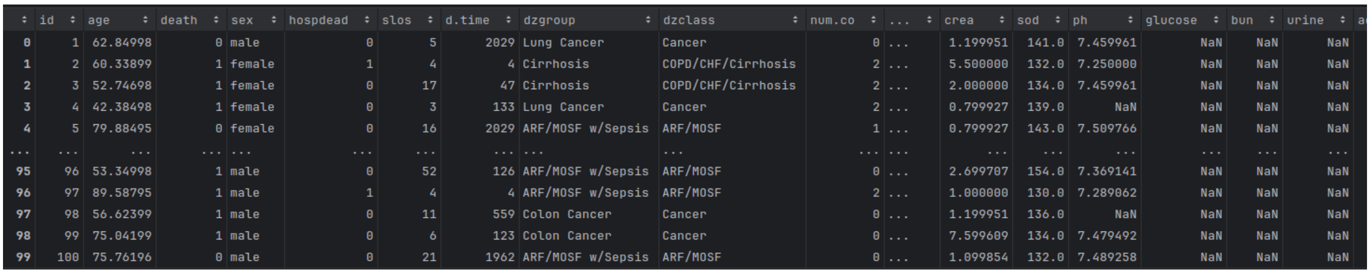
AA1 PROJECT - Dataset Declaration

GCED - UPC

2023-2024 (Group 12)



	id	age	death	sex	hospdead	slos	d.time	dzgroup	dzclass	num.co	...	crea	sod	ph	glucose	bun	urine	a
0	1	62.84998	0	male	0	5	2029	Lung Cancer	Cancer	0	...	1.199951	141.0	7.459961	NaN	NaN	NaN	
1	2	60.33899	1	female	1	4	4	Cirrhosis	COPD/CHF/Cirrhosis	2	...	5.500000	132.0	7.250000	NaN	NaN	NaN	
2	3	52.74698	1	female	0	17	47	Cirrhosis	COPD/CHF/Cirrhosis	2	...	2.000000	134.0	7.459961	NaN	NaN	NaN	
3	4	42.38498	1	female	0	3	133	Lung Cancer	Cancer	2	...	0.799927	139.0	NaN	NaN	NaN	NaN	
4	5	79.88495	0	female	0	16	2029	ARF/MOSF w/Sepsis	ARF/MOSF	1	...	0.799927	143.0	7.509766	NaN	NaN	NaN	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
95	96	53.34998	1	male	0	52	126	ARF/MOSF w/Sepsis	ARF/MOSF	0	...	2.699707	154.0	7.369141	NaN	NaN	NaN	
96	97	89.58795	1	male	1	4	4	ARF/MOSF w/Sepsis	ARF/MOSF	2	...	1.000000	130.0	7.289062	NaN	NaN	NaN	
97	98	56.62399	1	male	0	11	559	Colon Cancer	Cancer	0	...	1.199951	136.0	NaN	NaN	NaN	NaN	
98	99	75.04199	1	male	0	6	123	Colon Cancer	Cancer	0	...	7.599609	134.0	7.479492	NaN	NaN	NaN	
99	100	75.76196	0	male	0	21	1962	ARF/MOSF w/Sepsis	ARF/MOSF	0	...	1.099854	132.0	7.489258	NaN	NaN	NaN	

Group

This project will be carried out by Anna Esteve Gallifa and Biel Altimira Tarter, both enrolled in group 12.

Dataset

The chosen dataset is **SUPPORT2**, obtained from **UCI Machine Learning Repository**. The original source of the data is the **Vanderbilt University Department of Biostatistics**, which was made publicly available in 2023.

About

The dataset provides 9105 instances representing critically ill patients from 5 US medical centers. Each instance concerns a patient suffering from one or more of 9 categorized diseases. The 42 features, regarding each row, provide information about the patient's psychological state, disease severity and demographics.

No prior data processing has been performed to the dataset, nor any cleanup of null values. The features have categorical, real, integer and binary types.

Goal

Our goal will be to use the data to perform a classification task. It is of the upmost importance to assess critically ill patients to prevent painful and prolonged dying processes. Therefore, the gathered data should provide means to determine, based on their health status, the patients' survival rates from two to six months sight.

Link

https://archive.ics.uci.edu/dataset/880/support2