Divya Thomas

Class: CS 677

Date: 4/10/2023

Non-code answers to assignment.

**Question 1-1.** Pandas dataframes are printed in question1.py file.

**Question 1-2.** Matrices printed in question1.py output and plots saved in datasets directory (deceased\_patients.png and surviving\_patients.png)

**Question 1-3.**

1. The top features with the highest correlation for surviving patients are sex and smoking (0.49) and serum sodium and serum creatinine (-0.22).
2. The features with the lowest correlation for surviving patients are serum creatinine and time (0).
3. The features with the highest correlation for deceased patients are sex and smoking (0.36), sex and diabetes (-0.26), as well as creatinine phosphokinase and anemia (-0.26).
4. The features with the lowest correlation for deceased patients are serum sodium and diabetes (-0.008).
5. The results were very different between the surviving and deceased patients. The only consistency was that sex and smoking had one of the highest correlation values for both datasets, being the closest to a value of 1 in both cases.

**Question 2.** Group 3: X: serum sodium, Y : serum creatinine

See outputs from question2.py

**Question 3.** See summarized results from question 2 below.

|  |  |  |
| --- | --- | --- |
| **Model** | **SSE (death\_event=0)** | **SSE (death\_event=1)** |
| **y = ax + b** | 37.757 | 38.578 |
| **y = ax2 + bx + c** | 41.02 | 61.599 |
| **y = ax3 + bx2 + cx + d** | 36.379 | 2608.423 |
| **y = a log x + b** | 37.616 | 38.657 |
| **log y = a log x + b** | 38.905 | 22.567 |

**Question 3-1.**

The model with the smallest (best) SSE for surviving patients was the cubic spline model.

The model with the smallest (best) SSE for deceased patients was the second GLM model.

**Question 3-2.**

The model with the largest (worst) SSE for surviving patients was the quadratic model.

The model with the largest (worst) SSE for deceased patients was the cubic spline model.