**MATH2269/2305 Applied Bayesian Statistics**

**Final Project Report**

Declaration of contributions:

|  |  |  |
| --- | --- | --- |
| **No** | **Name of Team Member** | **Contribution to the project** |
| 1 | Jason Thomas | 1 |
|  | Sum: | must be 1 |

RMIT University,

School of Science

2022

Table of contents

[1. Introduction 3](#_Toc180942462)

[1.1 Report objectives 3](#_Toc180942463)

[1.2 Data 3](#_Toc180942464)

# Introduction

## 1.1 Report objectives

This report will address whether publicly-accessible data can be used to predict heart failure.

## 1.2 Data

The data for this project is available on Kaggle, and is called “Heart Failure Clinical data”

**https://www.kaggle.com/datasets/andrewmvd/heart-failure-clinical-data**

The data includes 299 observations. There are 12 independent variables, and one dependent variable.

The dependent variable and many independent variables are binary.

|  |  |
| --- | --- |
| Variable | Type |
| age | Numerical |
| anaemia | Binary |
| creatinine\_phosphokinase | Numerical |
| diabetes | Binary |
| ejection\_fraction | Numerical |
| high\_blood\_pressure | Binary |
| platelets | Numerical |
| serum\_creatinine | Numerical |
| serum\_sodium | Numerical |
| sex | Binary |
| smoking | Binary |
| time | Numerical |
| DEATH\_EVENT | Binary (dependent) |

## 1.3 Expert information

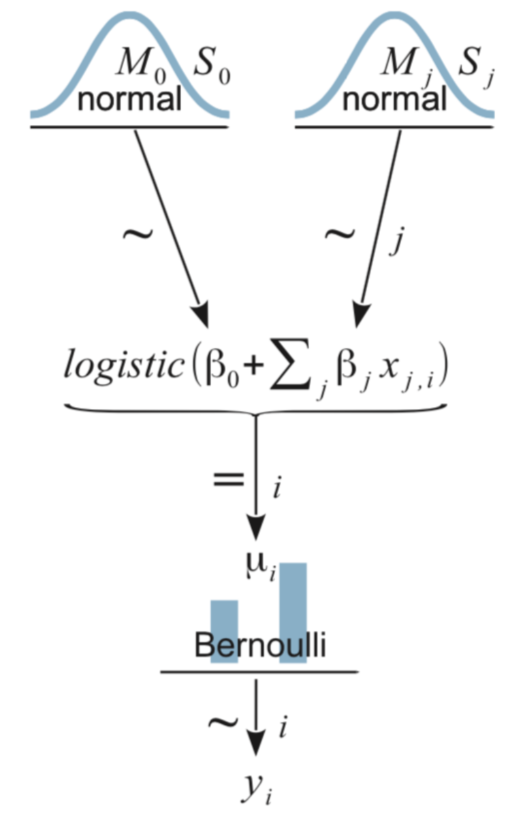
This report uses the assumption that there is no expert information available.

# Analysis

## 2.1 Model specification

Given that the dependent variable is binary, then logistic regression is an appropriate technique to create a model.

This is the model that this analysis will use.



***Fig 1***

The normal priors are appropriate since regression parameters are defined to be between negative to positive infinity and are continuous. The Bernoulli likelihood will allow for representing each prediction as a distribution.

## 2.2 Data summary

For the 12 independent variables mentioned in section 1.1, the first natural question was: do those variables correlate? If so then this would lead to removing some of those variables from the analysis.

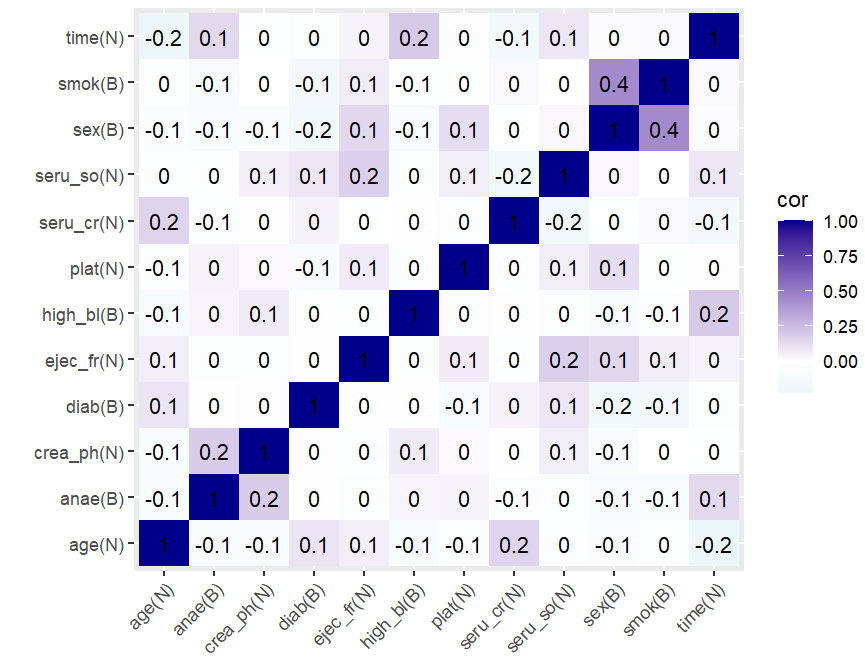
This analysis used the following correlation techniques:

Numerical ~ numerical: Pearson Correlation Coefficient

Binary ~ binary: Matthews Correlation Coefficient

Binary ~ numerical: Biserial Correlation Coefficient

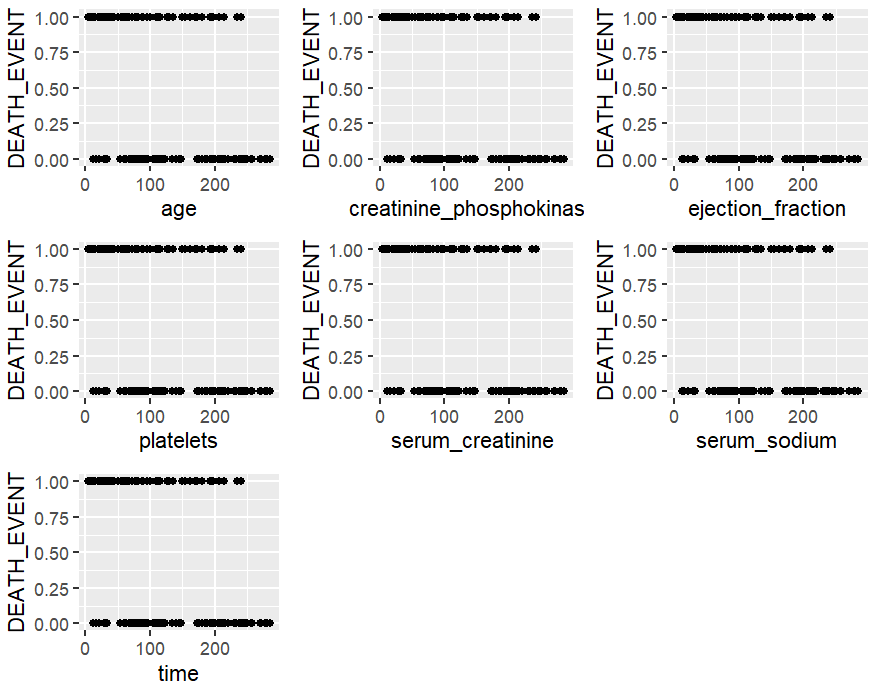
These are displayed on the same plot below:



***Fig 2***

All independent variables have correlations less than 0.5, so there is no basis to remove any for being too much like another variable.

Another important question is: how do the independent variables correlate with the dependent variable? This plot shows the results of numerical independent variables:



***Fig 3***

The correlations between independent variables and the dependent variable are:

The relationships