Imagine it is 1912 and you work for a well-known British insurance company. One of your clients, the White Star Line, has just lost a rather large ship at sea and after having collected information on their passengers, they wish to know which passengers had the lowest survival probability. In case you were wondering, the rather large ship was called HMS Titanic. Knowing which factors predicted the survival of each passenger will change the shipping company's future insurance purchases. In particular, they wish to not insure the highest-risk passengers themselves but offer the option to those passengers to optionally buy insurance privately upon purchasing the ticket.

The data contain the following information on each passenger:

表格

描述已自动生成

The test and training files were prepared for you and are attached below. What the company wants from you are the following pieces of information / answers:

* provide an explanation for which variables you selected (including any data transformations) for your prediction based on **economic intuition**
* a description of the main features of the data
* a classification tree which
  + estimates the probability of death in the Titanic data
  + provides an overview of which features were the strongest predictors of death
* a comparison with a simple probit classifier and a recommendation as to which method will perform better

The company pays a fictitious bonus salary for those who develop a more sophisticated tree model that can dominate both the probit classifier and the chosen baseline classification tree based on a given error metric.

you need to provide

* your python code
* a 1-page non-technical summary of how you reached the answers to the above questions, methods used, and results which can be understood by a non-expert