

泰坦尼克号生存率预测项目

1、背景介绍

背景介绍，提出问题

2、数据准备

数据来源及字段解释

创建分析环境

数据处理包

numpy/pandas

可视化包

matplotlib/seaborn

机器学习包

sklearn

其他包

warnings

导入数据

3、理解数据（观察数据）

查看数据情况，记录缺失/异常数据

可视化理解特征与标签间的关系

4、数据预处理（操作数据）

数据清洗（缺失值处理）

特征工程（构造新特征）

同组识别（修正数据）

筛选子集（数据降维）

5、构建模型

模型比较

比较不同算法性能

模型评分

准确率/ROC/混淆矩阵

模型选择

选择表现较好的构建模型

6、模型预测

解决问题

7、总结