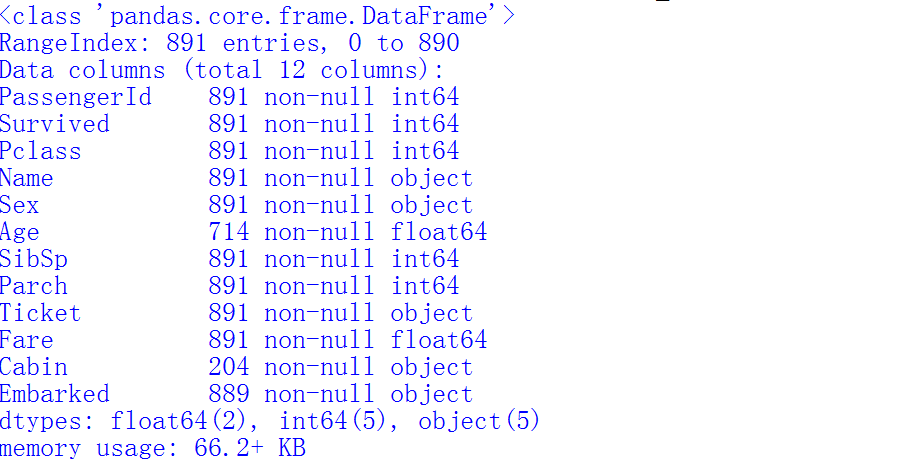
**作业三分析过程报告**

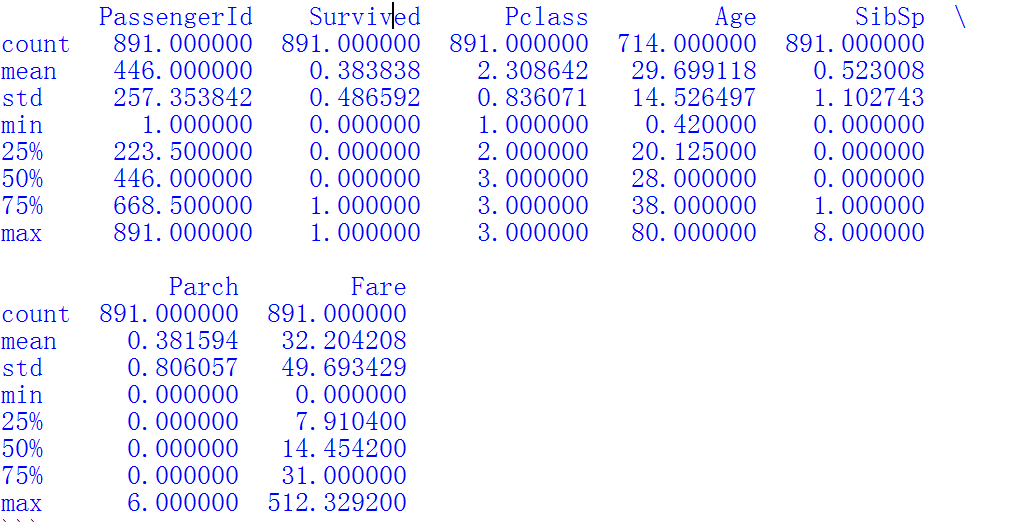
（第三个数据集—泰坦尼克号遇难预测）

数据集包括训练集、测试集和参考结果。首先，要了解数据集的组成情况，对数据集进行相应的预处理。然后利用聚类方法和分类方法进行预测。

1. 数据整体预览
2. 查看数据整体情况，结果如下：

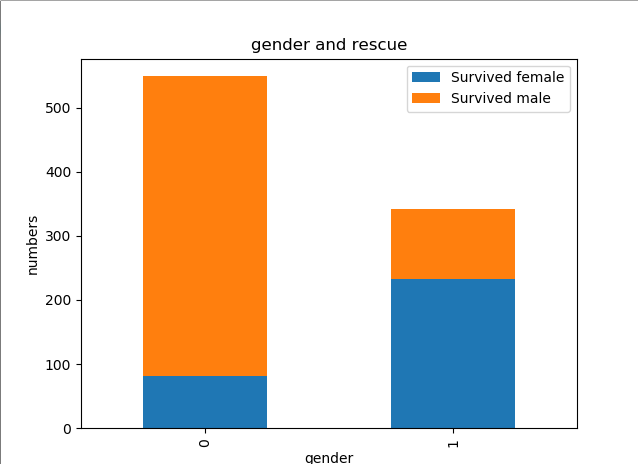


1. 查看各属性的数值特征



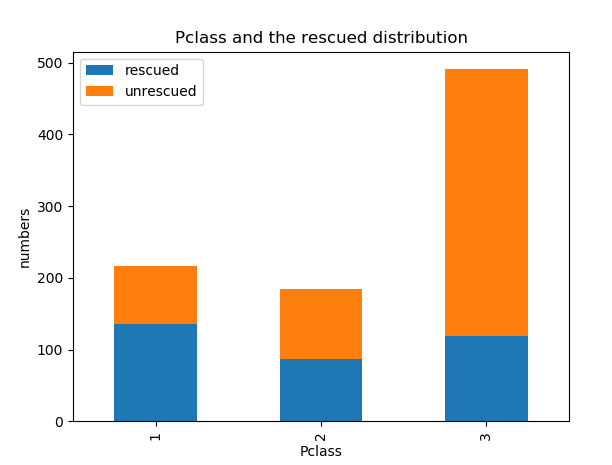
3.查看属性之间的关系

（1）观察性别与是否存活之间的关系



由上图可以看出，女性存活的概率远大于那行存活的概率。

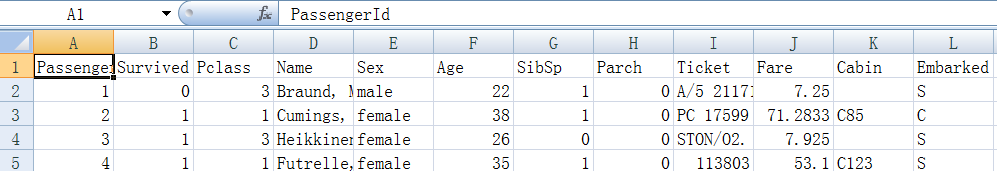
1. 查看乘客等级与是否存活之间的关系

通过上图可以看出，顾客等级越高，存活的比率越大。

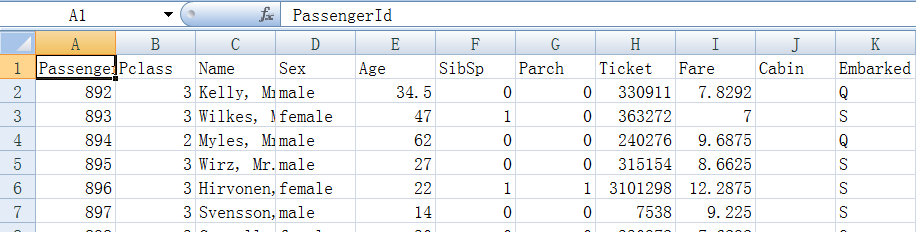
二、数据预处理

1. 通读数据，对于训练集和测试集都要做去除与预测无关的属性。

训练集数据属性如下图所示：



测试集数据属性如下图所示：

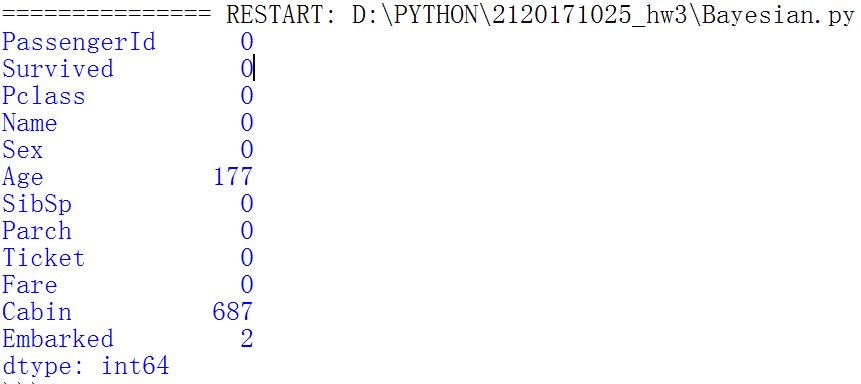


|  |  |  |
| --- | --- | --- |
| 属性 | 描述 | 是否保留 |
| PassengerId | 乘客编号 | 是 |
| Pclass | 票类（共有三种票别） | 是 |
| Name | 乘客姓名 | 否 |
| Sex | 乘客性别 | 是 |
| Age | 乘客的年龄 | 是 |
| Sibsp | 泰坦尼克号上该乘客的兄弟姐妹/配偶数量 | 是 |
| Parch | 泰坦尼克号上该乘客的直系亲属的数量 | 是 |
| Ticket | 票编号 | 否 |
| Fare | 乘客票价 | 是 |
| Cabin | 客舱号码 | 否 |
| Embarbation | 登船港口（C、Q 、S分别表示不同的港口名） | 是 |
| Survived | 是否生存（1表示是，0表示否） | 是 |

1. 缺失值处理
2. 查看缺失值情况

print(datatrain.isnull().sum())

结果如下图所示：

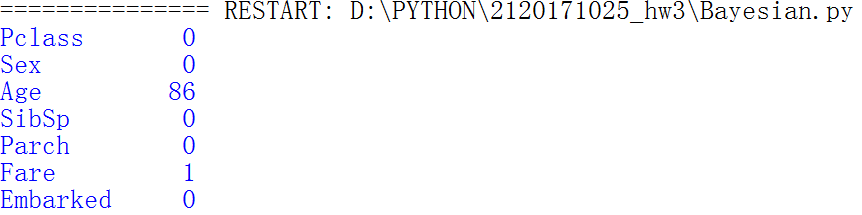


对于训练集，发现cabin属性缺失值过多，因此将其删除。而对缺失值少的属性，可以将有缺失值的记录删除掉，不作为训练集数据。

datatrain = datatrain.dropna()

对于测试集，也查看其缺失值：

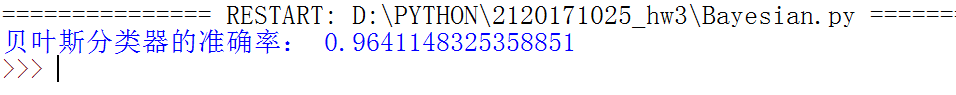
print(datatest.isnull().sum())



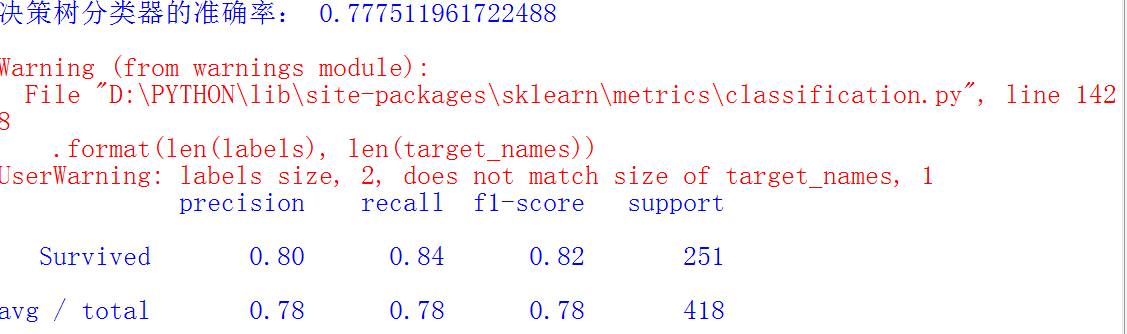
只有Age和Fare有缺失值，且这两个属性都为数值属性，因此可用两个属性的均值填补缺失值。

datatest = datatest.fillna(datatest.mean()['Age':'Fare'])

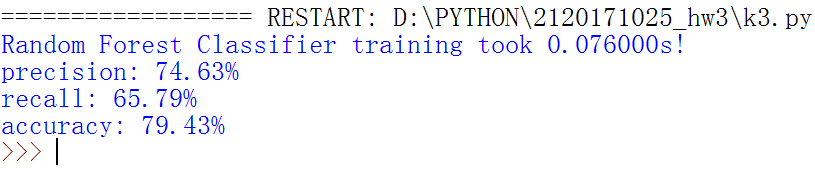
1. 贝叶斯分类结果



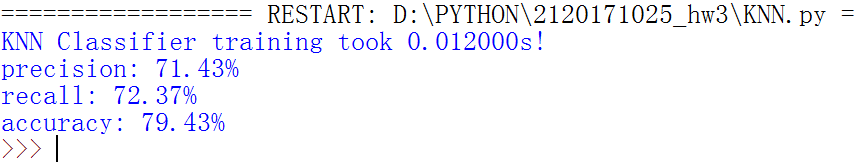
1. 决策树分类结果



1. 随机森林聚类算法



1. KNN聚类算法



1. 逻辑回归

