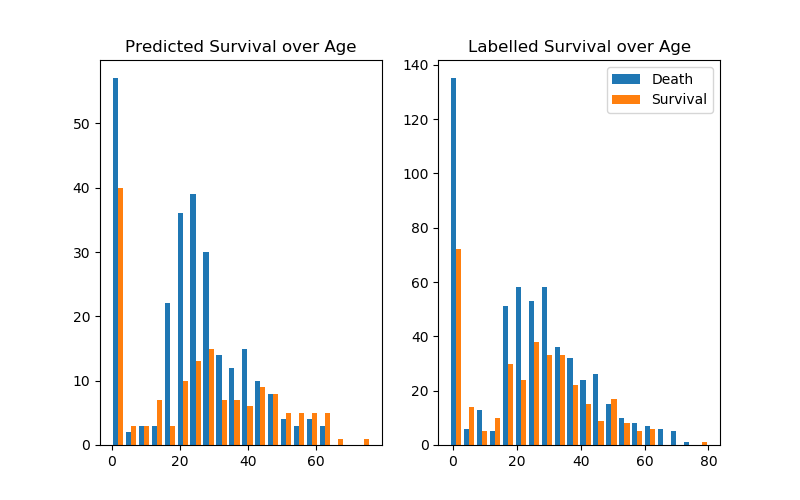
数据结果分析

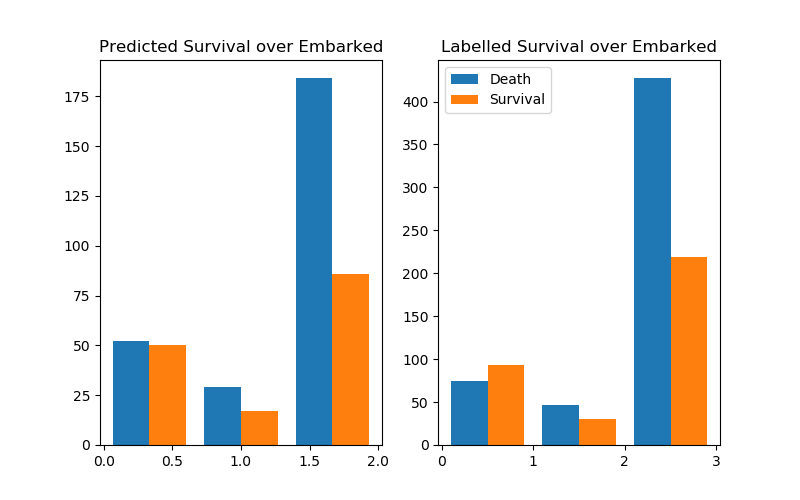
# 1. 分类模型

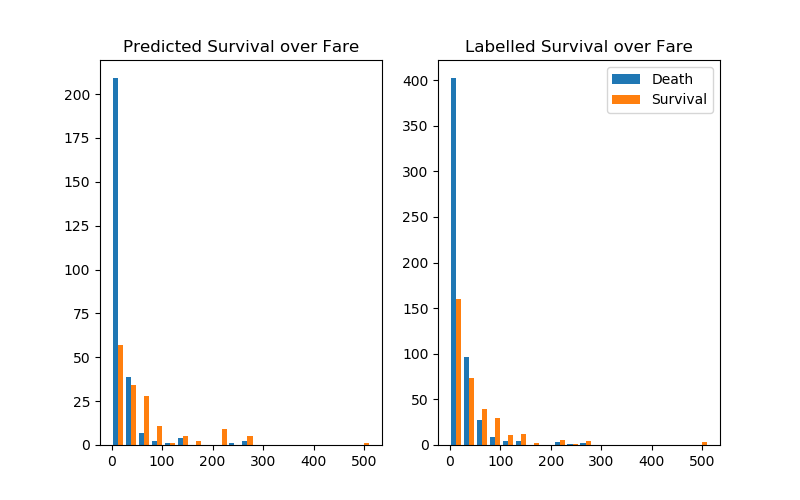
在分类这一部分，我们采用了两种分类算法，一种是决策树，另一种是SVM。我们将筛选出的属性作为输入的特征，通过分类模型预测出测试集的“幸存”属性。

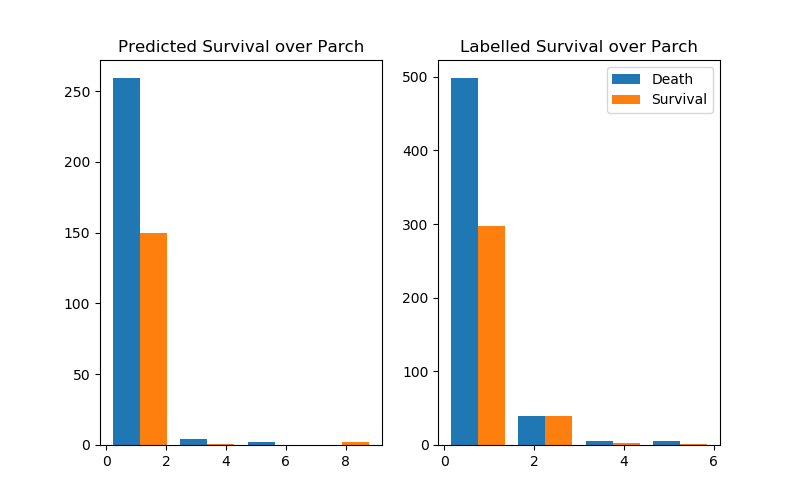
## 1.1 SVM分类

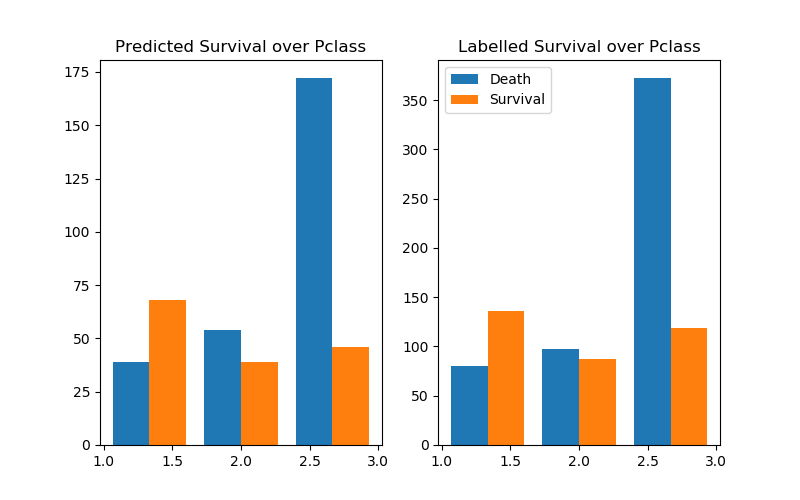
为了分析每个属性对分类结果的影响，在这里我们对每个属性的分类结果进行可视化，并且绘制了相应的训练数据中的直方图进行比较。可视化图像如下图所示。

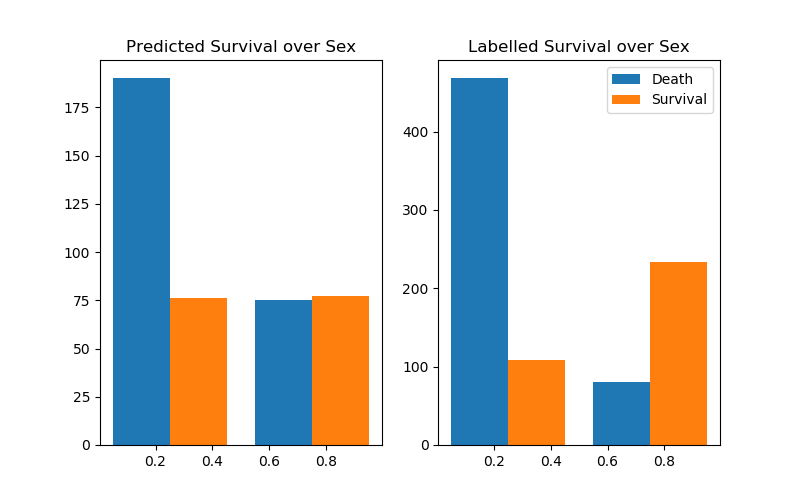


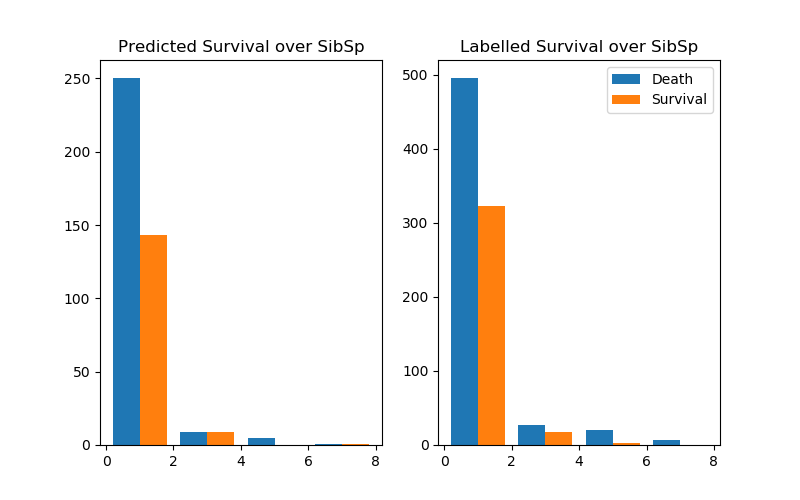












通过预测结果的直方图与实际结果的直方图比较，我们可以看出，我们的多数预测结果的分布与实际结果还是比较相似的，只有Sex的预测有一点误差。

我们分别分析每个属性：

（1）Age: 由图我们可以看出，青壮年的死亡率较高，其余年龄段的死亡率与存活率基本持平。

（2）Embarked: 该属性表示船员登船的位置。可以看出Southampton登船的存活率最低，而Cherbourg登船的存活率最高。

（3）Fare: 可以看出船票越贵的乘客的存活率越高。

（4）Parch: 没有明显关系。

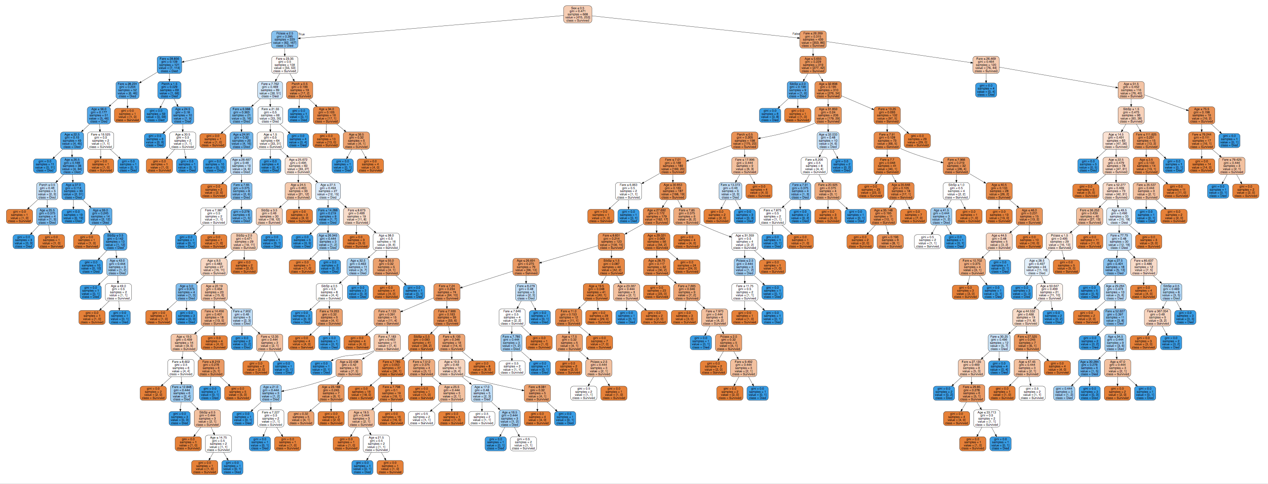
（5）Pclass: 购买一级船票的存活率最高，购买三级船票的死亡率最高，与Fare 属性的结果类似

（6）Sex: 男性的死亡率比女性的死亡率高很多。

（7）SibSp: 与Parch类似，没有明显关系。

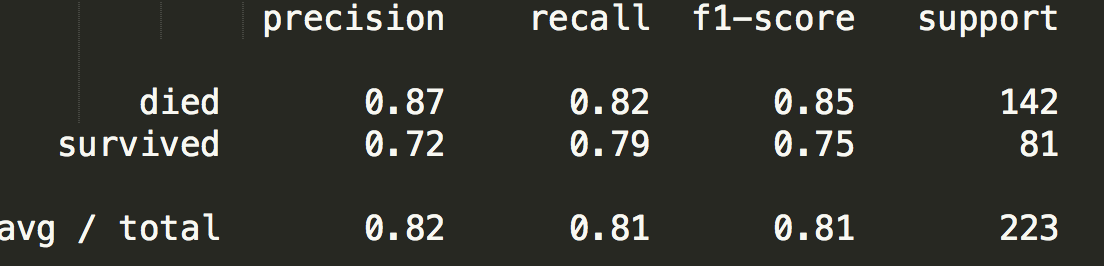
## 1.2 决策树

我们使用决策树进行分类的可视化后的图如下所示。



其中蓝色部分代表该样本未存活，橙色部分代表该样本存活。

下图为在测试集上预测的准确率，平均准确率可达80%以上。



# 2. 聚类

我们使用的两种聚类方法分别为K-Means和BRICH，可视化结果分别如下图所示

