数据挖掘大作业三：分类与聚类

——周宇田-2120171106

本次作业选择的数据集是第三个，关于泰坦尼克号的数据。数据包含了三个部分：train.csv、test.csv、gender\_submission.csv。数据记录了泰坦尼克号上乘客的信息，包括年龄，性别，消费情况等等，既包含数值属性也包含标称属性。

在数据预处理部分，在分析了数据之后我删除了部分属性，例如，Name、Ticket、Cabin。因为Name对数据挖掘不能提供有用信息，Ticket编号同样不能提供有用信息，Cabin表示船舱号，在给定的数据中大量缺失，因为当时之后上等仓的人才有船舱号。同事我也增加了一部分属性，例如，Title、Deck、Family\_Size、Fare\_Per\_Person。其中Title来自于对Name属性的分析，获取名字里面的敬语称谓来标定人物别，Deck来自于对Cabin的分析，Fare\_Per\_Person表示船上团体中个人在船上的消费，这是一种家庭实力的象征，也许和最后能够活下来有关系。需要注意到的是数据中Fare出现了空缺，可能是有人逃票上船，因此费用应该设置为0 。另外，Age的缺省值用平均年龄代替，登船港口用频率最高的港口代替。最后，所有的标称属性用数值编码代替。

对数据进行分析的过程中，采用了4中分类模型和2种聚类模型。分类模型分别是：朴素贝叶斯分类模型、KNN分类模型、支持向量机分类模型、决策树分类模型。这四种模型在训练集上训练之后在测试集上开始测试，并生成了对应的CSV文件记录结果。聚类模型是层次聚类和KMEANS聚类，聚类模型直接在测试集上展开聚类，将测试集分成了两个簇，考虑到泰坦尼克号的实际情况，将数目较多的一类设置为0，即死亡。

在所有的模型训练并预测之后，对所有结果进行了横向对比，评估出了各种模型的准确度，结果见summary.csv 。

所有的实验结果均见文件夹output中的图片和文件。