我们将使用数据探索葡萄牙Vinho Verde葡萄酒的质量和理化特性。 这些数据来自Cortez等人发表的论文。 在2009年的决策支持系统中，该论文已包含在项目文件夹中。 开发这些数据是为了根据其理化特性预测葡萄酒的质量。

以下是这些特征/理化特征的简要说明：

质量-根据盲人味觉测试得分在0（最差）到10（最佳）之间

葡萄酒-葡萄酒的类型，白色或红色

fixed\_acidity-酒石酸克/ dm3

volatile\_acidity-乙酸克/ dm3

柠檬酸-柠檬酸克/ dm3

残糖-克/ dm3

氯化物-氯化钠克/ dm3

游离二氧化硫-毫克/ dm3

总量\_二氧化硫-毫克/ dm3

密度-g / cm3

pH-酸度

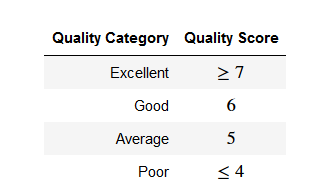
硫酸盐-硫酸钾克/ dm3

酒精-酒精百分比

我们提供了三个数据文件：wine\_qual\_train.csv，wine\_qual\_test.csv和wine\_qual\_holdout.csv，分别包含3000、1000和1000个观测值。 前两个文件可以以您希望用来训练模型的任何方式使用。 与第一个项目一样，保持文件包含与测试文件相同的数据，在您提交作业以获得最终模型的真实样本外评估后，我们将用新数据替换该文件。 有关此项目的其他详细信息，请参阅项目笔记本的“模型验证”部分。

分配

在此项目中，您的任务是开发一种评估这些葡萄酒质量的分类模型。 与其将这些分数视为数值，不使用基于回归的方法，我们将把质量数据离散化为以下几类：优，良，中，差。 必须使用下表进行分类：



因此，目标是开发一种分类模型，以根据其理化特性正确识别葡萄酒的质量类别。与第一个项目一样，目标是生成一个具有高预测准确性的模型，同时也可以进行解释-知道是什么使“优秀”葡萄酒与能够将“好”与“平均”分开一样重要，甚至更多。葡萄酒。

您的模型可以使用所提供的功能的数量少或多，并且您可以通过任何想要生成附加功能的方式来转换和操纵这些功能。唯一的限制是您不得在模型中使用任何其他外部数据。

我们在讲座和讲习班中介绍了许多模型和建模方法，您应该针对此特定任务探索各种不同的方法。但是，您的最终目标是提供一个准确而可解释的模型。这些是相互竞争的利益，您需要在两者之间找到合理的平衡-最终成绩将取决于您的完成程度。

我们提供了一个名为project2.ipynb的Jupyter笔记本模板，其中包含所有必需的部分以及每个部分应包含的内容的简要说明。您的所有工作都必须包含在project2.ipynb笔记本中，我们只会标记该文件中包含的内容。我们已经为每个部分提供了说明，这些说明应在提交前从最终文档中删除。此外，您可以随意添加小节或其他替代格式，但是请保留其中所有5个小节。最终文件的篇幅不得超过10页（包括代码和数字）。

我们还包括了supp\_mat.ipynb文件，用于包含您要添加的任何其他补充材料。这是一个包含替代模型的好地方，您最终拒绝了替代模型，而选择了最终模型。

投稿

该项目的截止日期是4月15日，星期三，下午5点。 您应按以下方式提交完成的工作：

     通过Noteable的分配工具提交笔记本和支持文件。

和

     通过学习提交导出的html或pdf版本的项目笔记本。 若要生成这些文件，请从“文件”菜单中选择“下载为”。 您很多人需要先压缩该文件，然后Learn才能将其接受。 不要在学习中提交补充材料或任何其他文件。

这两个提交步骤对于将您的工作视为已提交都是必需的。 标准逾期罚款将适用于截止日期之前未提交的工作。

强烈建议您在重新启动Jupyter内核后尝试重新运行整个笔记本，以确保在提交作业之前所有内容都可重现。

标记专栏

该项目的评分标准为100，而我们将使用以下标题大致指导评分：

    > 90：代码运行无错误。模型得到正确实施，拟合和评估。最终模型可以实现较高的预测准确性，并且在本文中有充分的文档记录和描述。有大量创新的问题调查。潜在地可以用作模型答案。书面证据表明您对数据和模型有深入的了解。

    80-89：代码运行没有错误。模型得到正确实施，拟合和评估。写的通常是好的，并且对代码进行了适当的注释。最终模型可以达到合理的预测准确性水平，并且在本文中有充分的记录和描述。有适度的其他问题的调查。书面证据证明对数据和模型有很好的理解。

    70-79：代码运行没有错误。仅对次要问题进行了模型的正确实施，拟合和评估。写得还可以，但可能会更好。对该问题可能还有很多其他调查。书面证明充分理解了数据和模型。

    60-69：代码运行没有错误。只有适度的问题才能正确地实施，拟合和评估模型。写得还可以，但可能会更好，包括一些中度的错误或遗漏。书面证明充分理解了数据和模型。

    50-59：代码运行时出现错误。模型已实现，适合和评估，但存在一些严重问题。写的还可以，但可能会更好，包括一些重大的错误或遗漏。书面证据表明对数据和模型的理解不完整。

    <49：代码，模型和/或编写的重大问题。书面证据表明对数据和模型的理解不完整。