2023年 次世代交通团队技术部考核

需求内容：

写出一个具有界面的软件，可以导入excel、csv等表格数据文件。

数据文件demo

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.925		S

导入数据文件后，可以执行三种操作：拆分数据集，训练，预测

- 拆分数据集：将导入的数据文件按照选定的比例进行拆分（常用的比例为7：3，即导入的数据文件的70%的行是训练集，30%的行是测试集），拆分为训练集和测试集

- 训练：根据训练集的列名，以Survived列作为标签，其他列作为可选择的特征，进行拟合，最终得到一个可以对Survived进行预测的模型。
- 预测：使用训练得到的模型，输入测试集中的特征，对标签进行预测，并输出其性能（性能指标可以使用accuracy）。

技术栈：

1. 图形界面：**pyqt**
2. 模型拟合：线性回归/支持向量机/决策树等**分类模型**
3. 导入数据：**pandas、文件IO**

需求技术细节：

1. 导入excel、csv等表格数据文件可以使用pandas库。如果使用自己python自带的io会额外加分。
2. 拆分数据集的比例可以是写死的，如果可以自己选择比例会额外加分。
3. 模型训练和预测可以使用sklearn库。如果使用自己numpy、tensorflow或pytorch自己写模型会额外加分，完全使用python自带的函数会加更多的分。
4. 模型训练的特征和标签是哪一些可以是写死的，如果可以自己选择哪一些列作为特征 哪一些列作为标签会额外加分。

5. 在模型训练前往往对数据进行预处理，预处理可以是写死的，如果可以自己选择对哪些列进行哪些预处理会额外加分。
6. 在模型训练和预测的时候可以只输出结果，如果可以输出图像（如损失值的下降等）会额外加分

参考材料：

1. pyqt官方文档 https://maicss.gitbook.io/pyqt-chinese-tutoral/pyqt6/firstprograms
2. Sklearn 机器学习官方文档 https://scikit-learn.org/stable/
3. 机器学习算法讲解视频 https://www.bilibili.com/video/BV1m24y1h7bL/?spm_id_from=333.337.search-card.all.click
4. Pyqt开发教程 https://www.bilibili.com/video/BV1tV41167k1/?spm_id_from=333.337.search-card.all.click
5. pandas 开发教程 https://www.bilibili.com/video/BV1UJ411A7Fs/?spm_id_from=333.337.search-card.all.click&vd_source=b76cb9f3bba1341731bf8848c7a94ce6