zhejiang-yige

智能软件综合研究

实训报告



学期 2023-2024-1

班级 计算机科学与技术留学生21（1）

姓名 LOW REN HONG

学号 2021529620004

教师 马明泽、吴婷婷

|  |  |  |
| --- | --- | --- |
| 学号： 2021529620004 |  | 姓名： LOW REN HONG |

1. **实验目的和要求**
2. 掌握数据预处理相关操作；
3. 掌握深度学习模型构建、训练、评估、优化等过程；
4. 掌握智能软件全周期开发、部署过程。
5. **实验内容**
6. **环境配置。**
7. **数据加载及预处理**
8. **数据集分割。**
9. **模型构建、训练、评估**
10. **实验任务**
11. **环境配置。**报告中需说明实验环境相关信息。
12. **数据加载及预处理**

* 数据加载及数据确认。要求展示头部数据、尾部数据、数据类型，确认数据完整性。
* 数据预处理。

（1）检查是否有重复数据并删除；

（2）检查是否存在缺失数据以及各特征缺失数据的比例，对缺失数据进行填充操作；

（3）将字符串和布尔字符转换成整数，完成编码操作；

（4）尝试分析特征之间的关系，将其可视化。

* 预处理数据保存。

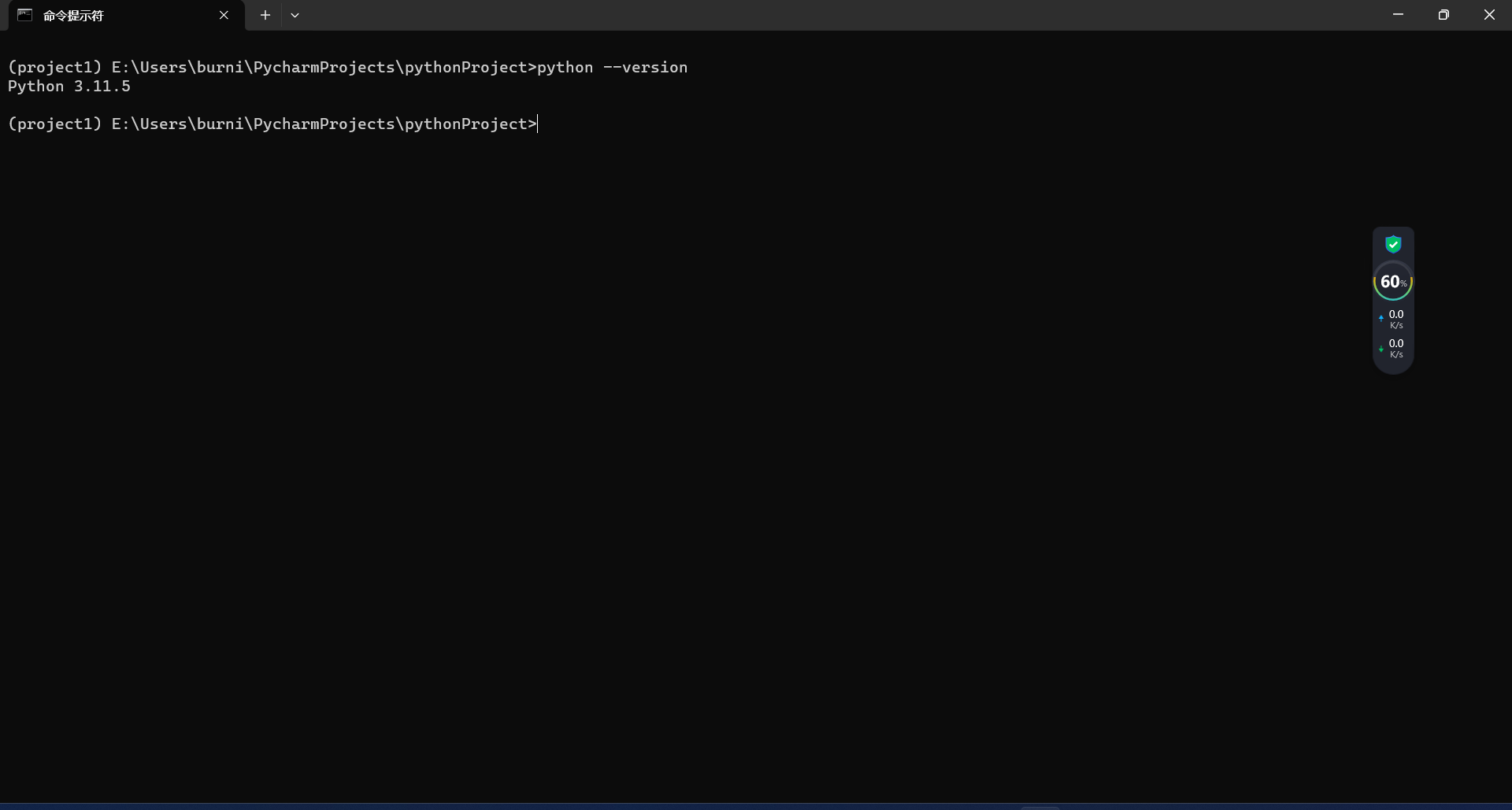
1. **数据集分割。**将预处理后的数据集分割成训练集和测试集，其中训练集70%，测试集30%。
2. **模型构建、训练、评估**

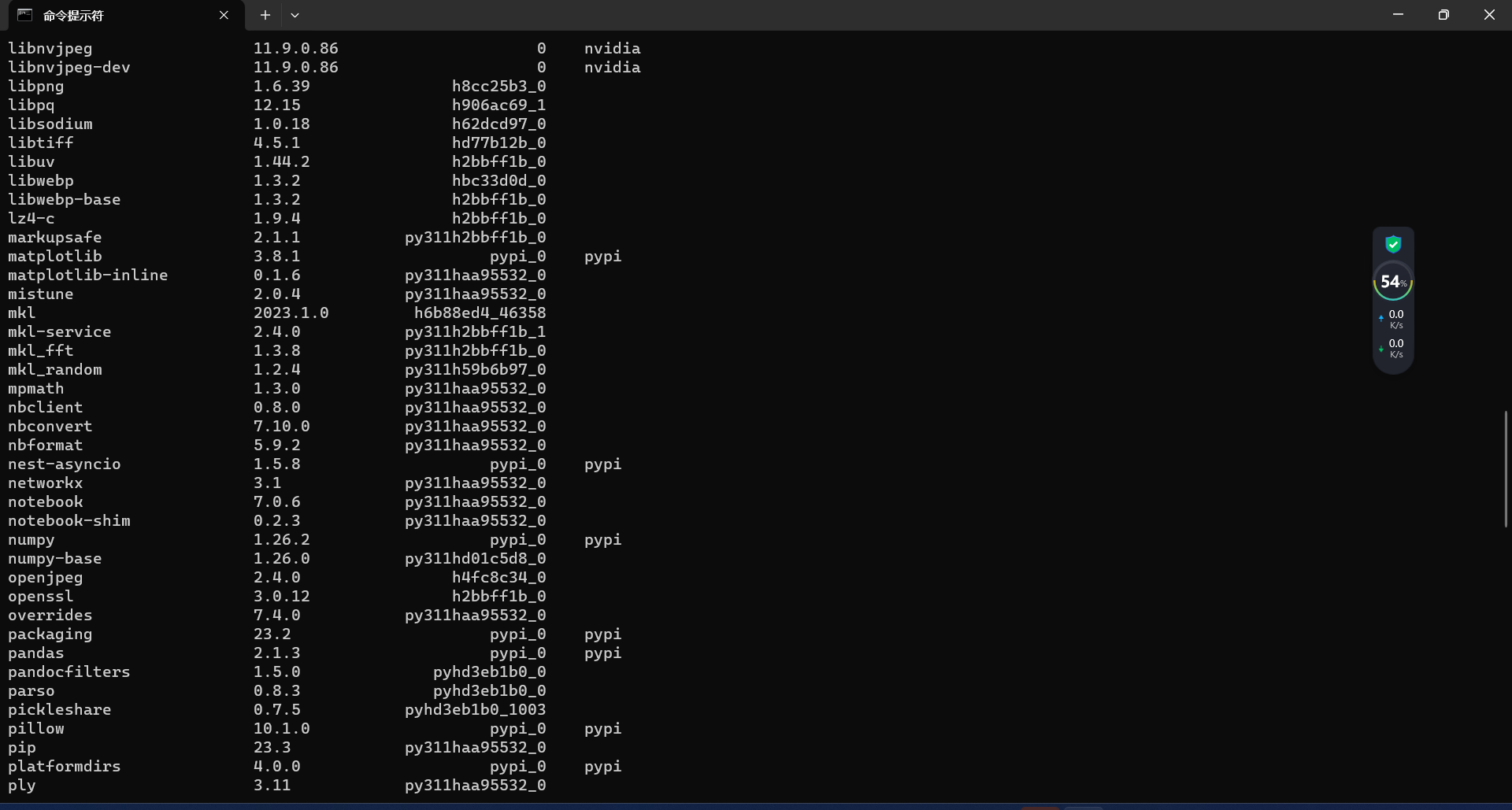
分别利用**决策树、支持向量机、随机森林**三种模型对以上预处理后的乳腺癌数据进行训练、准确度评估、模型性能对比。

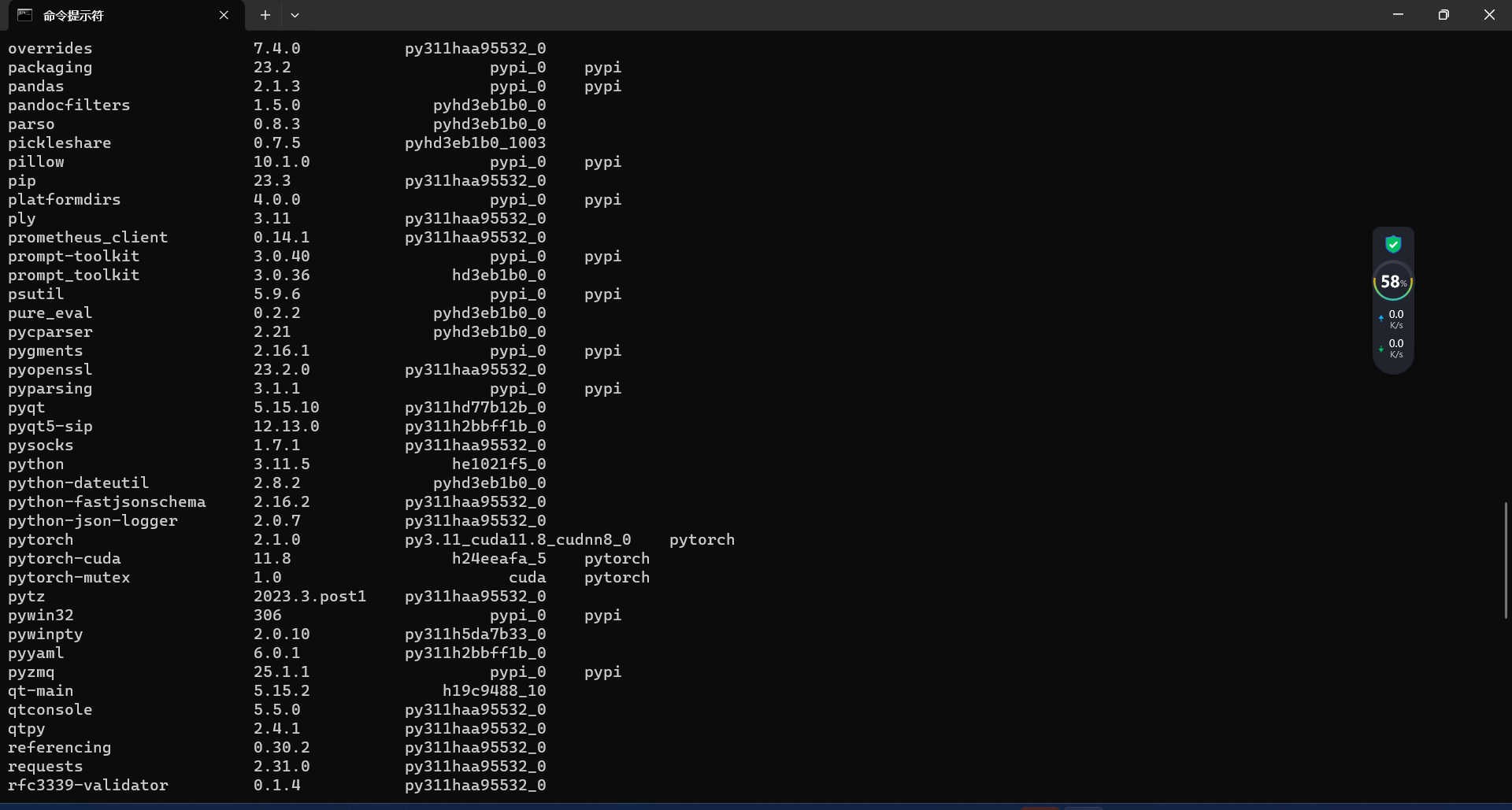
1. **实验过程和结果分析**
2. **环境配置。**

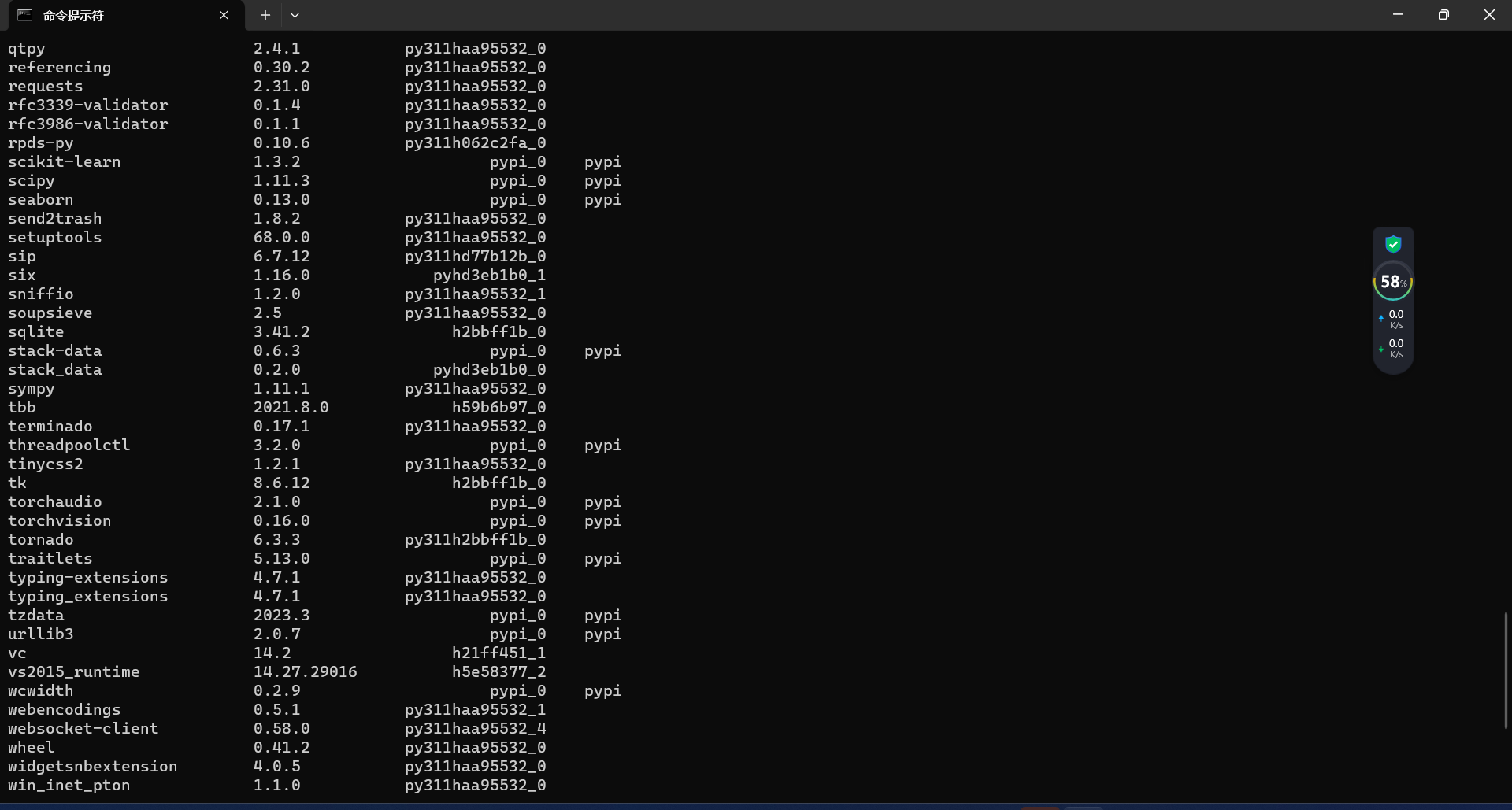
**本次实验使用的是JupyterNotebook，配置了matplotlib还有pandas和seaborn以及scikit-learn。**

**Python是3.11.5**

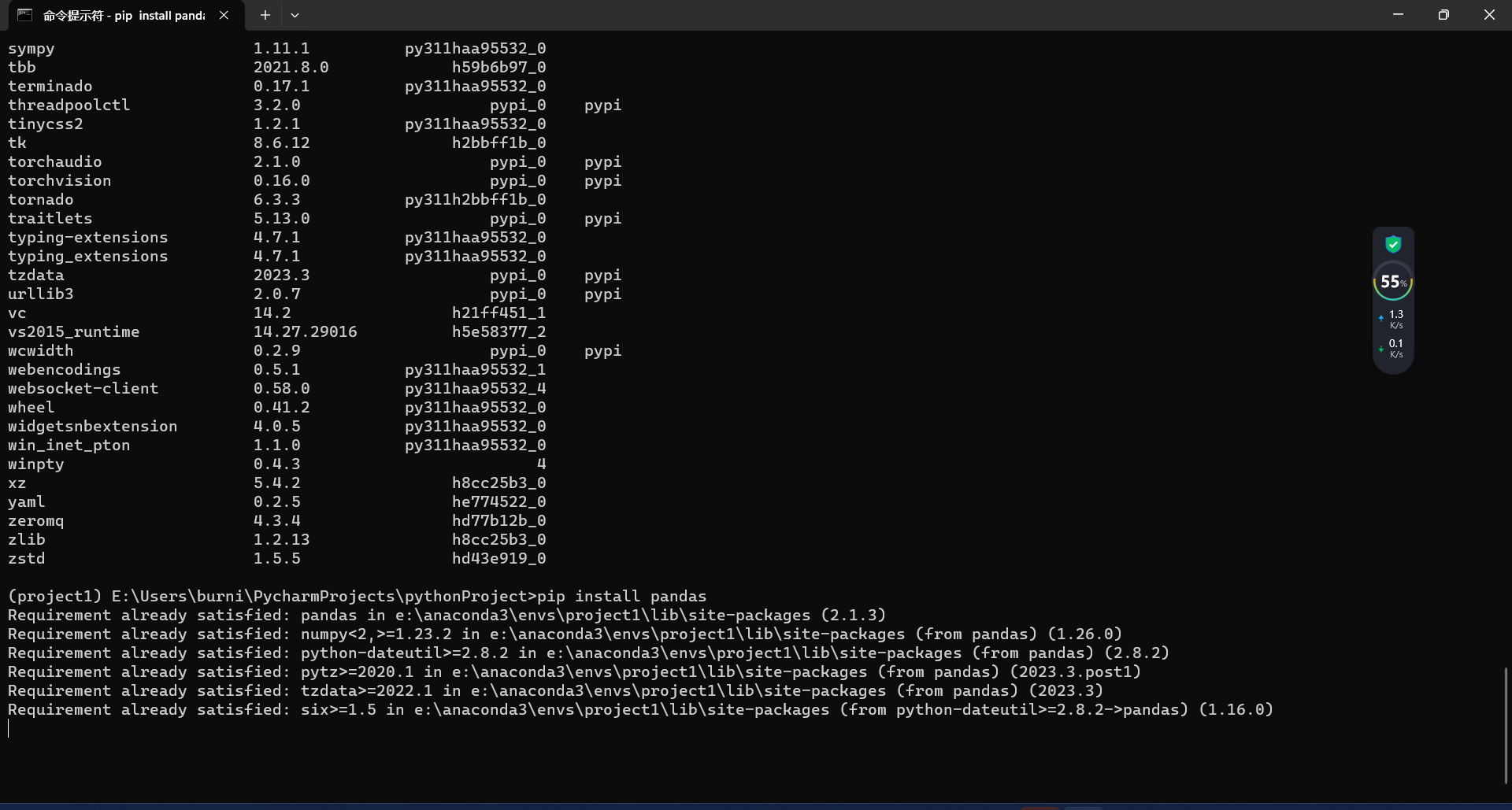




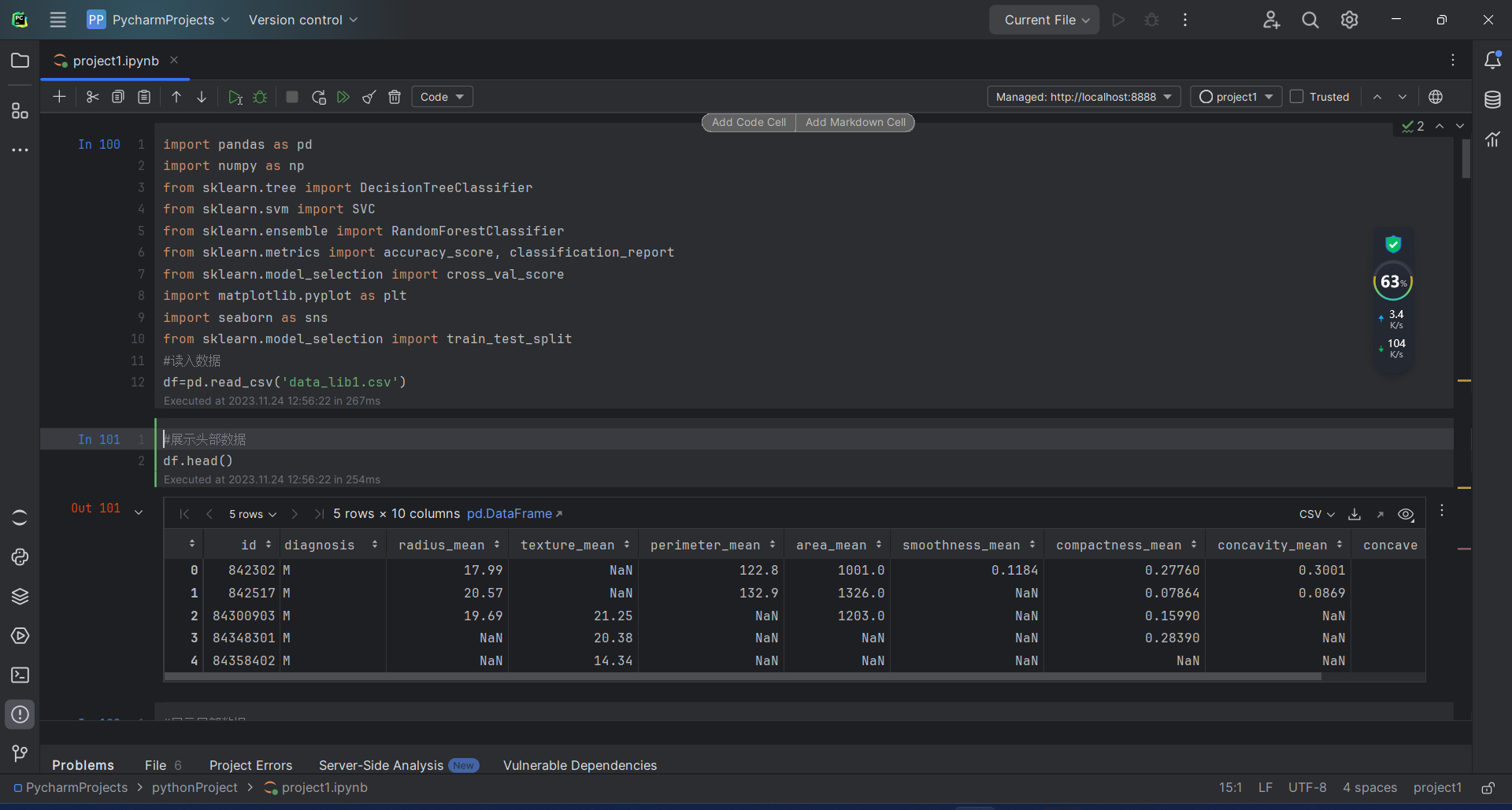


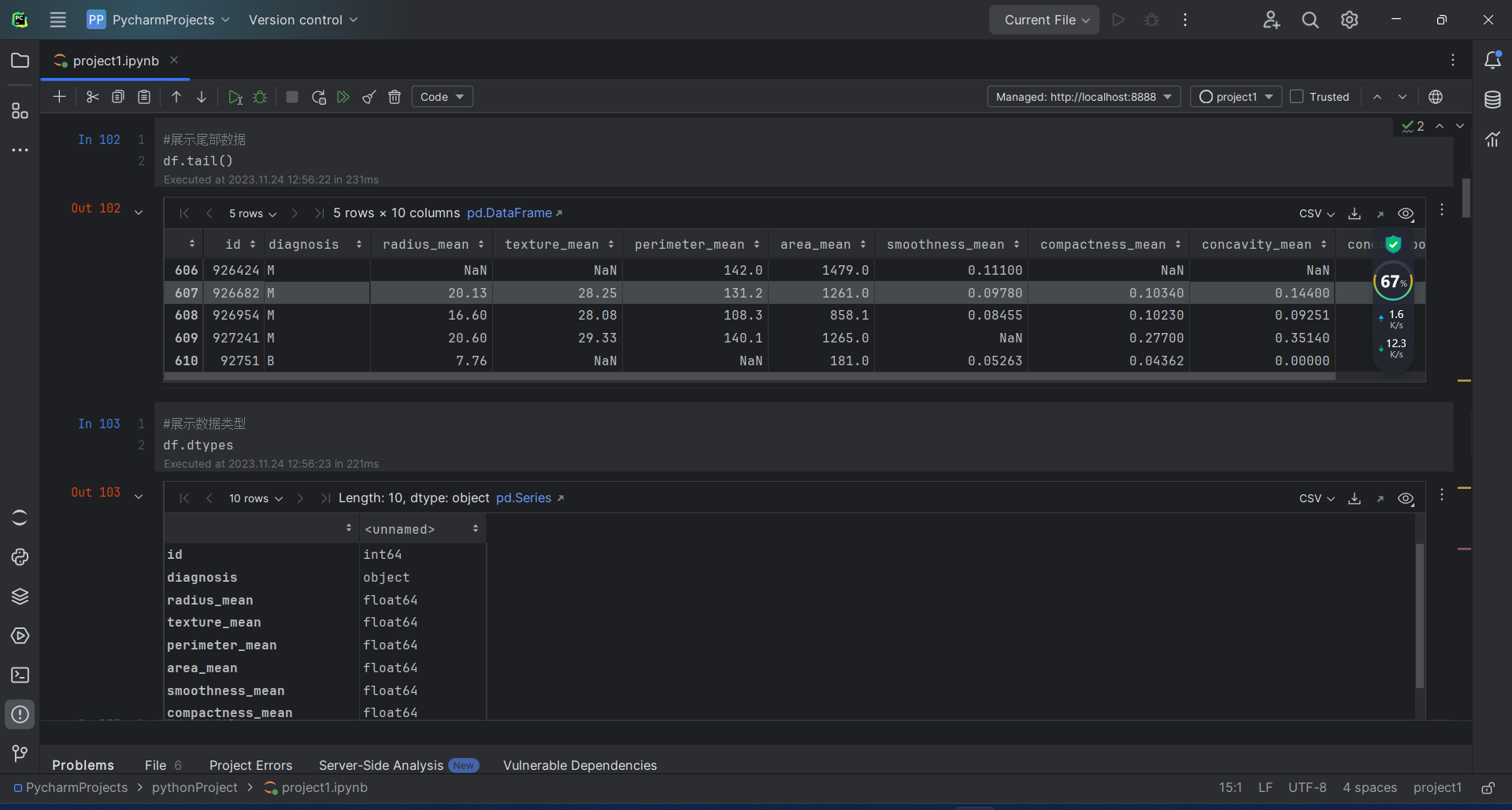


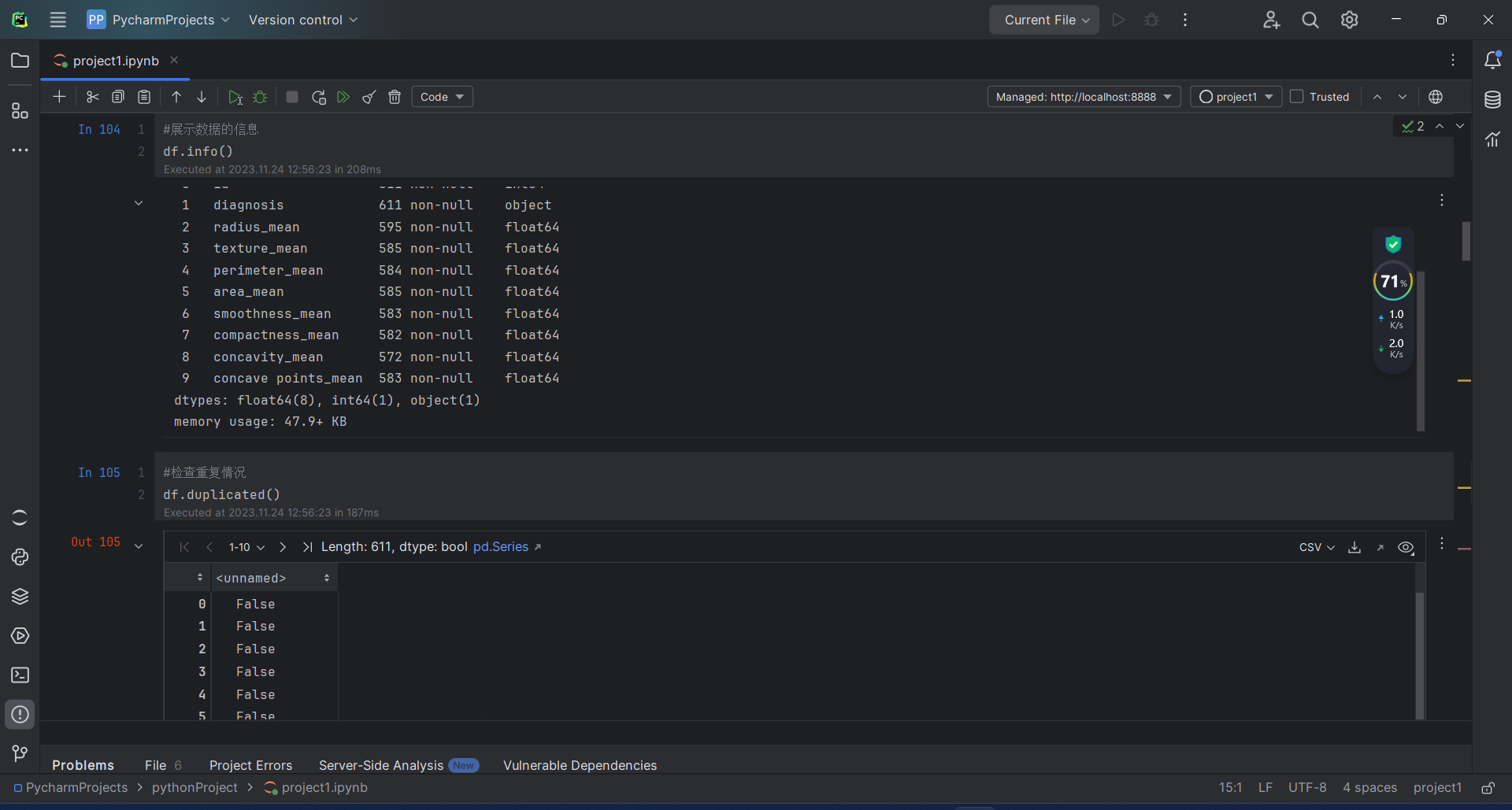
可以使用pip install 需要用到的库就可以下载相应的库

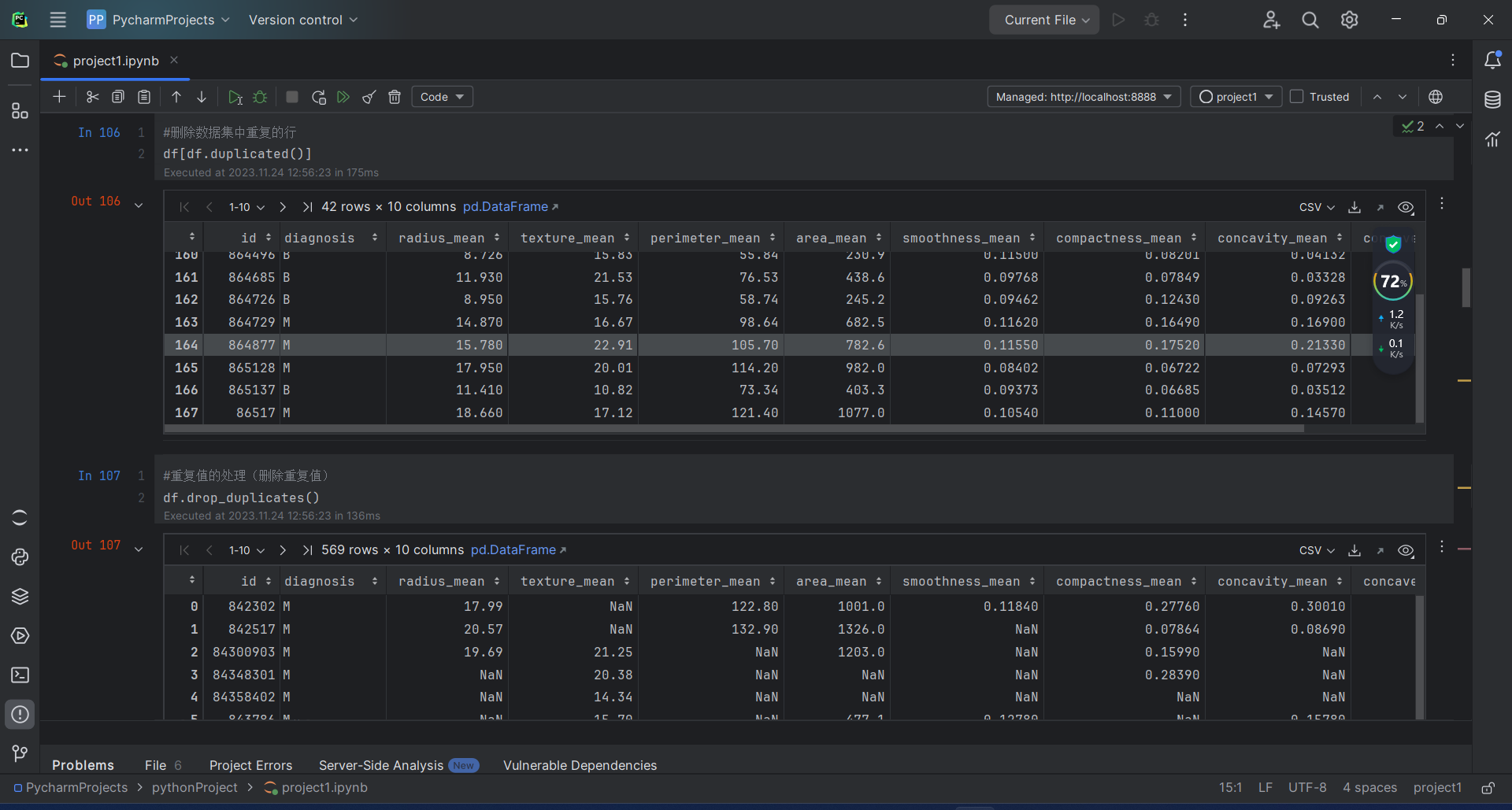


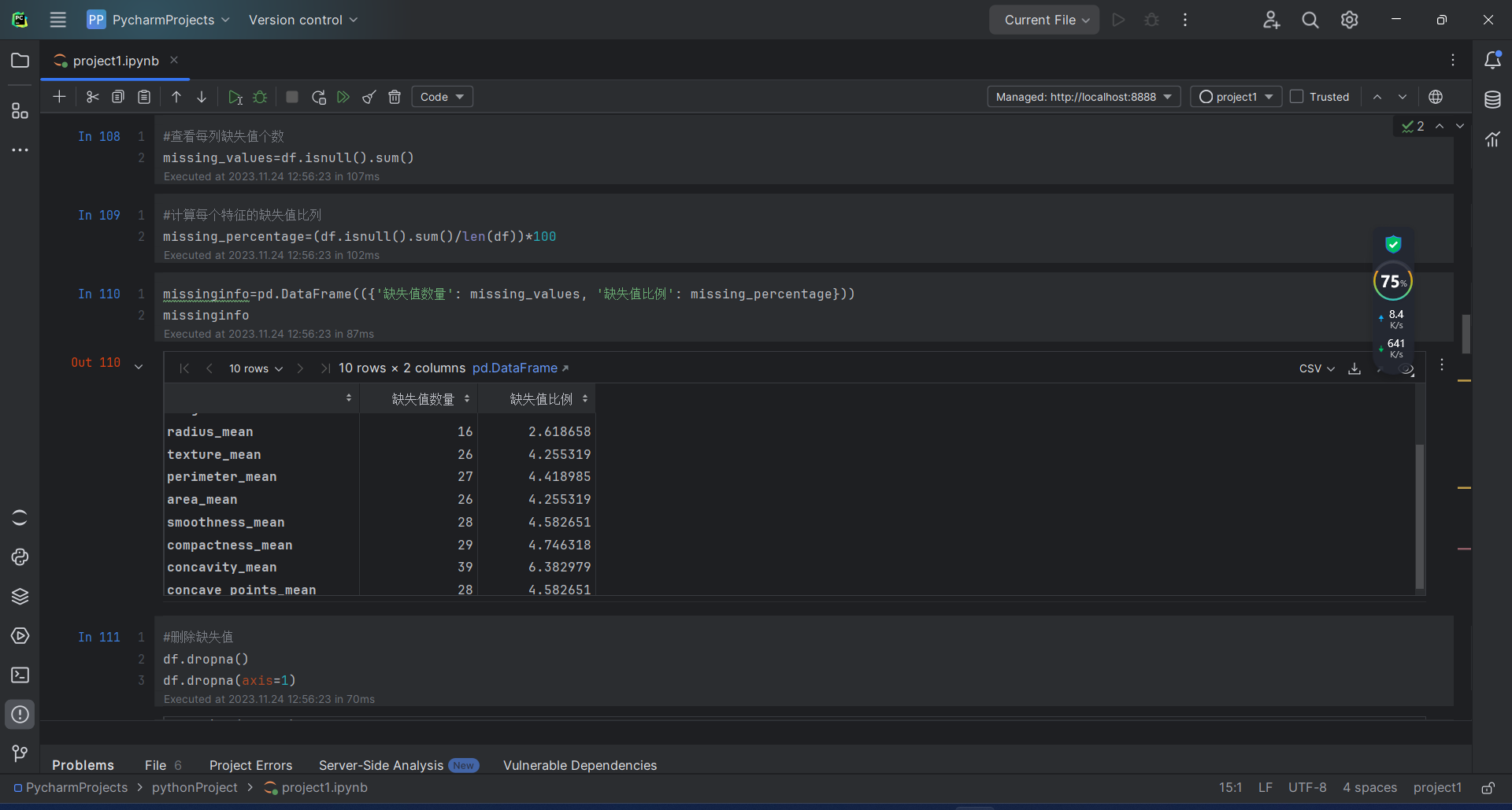
1. **数据加载及预处理**

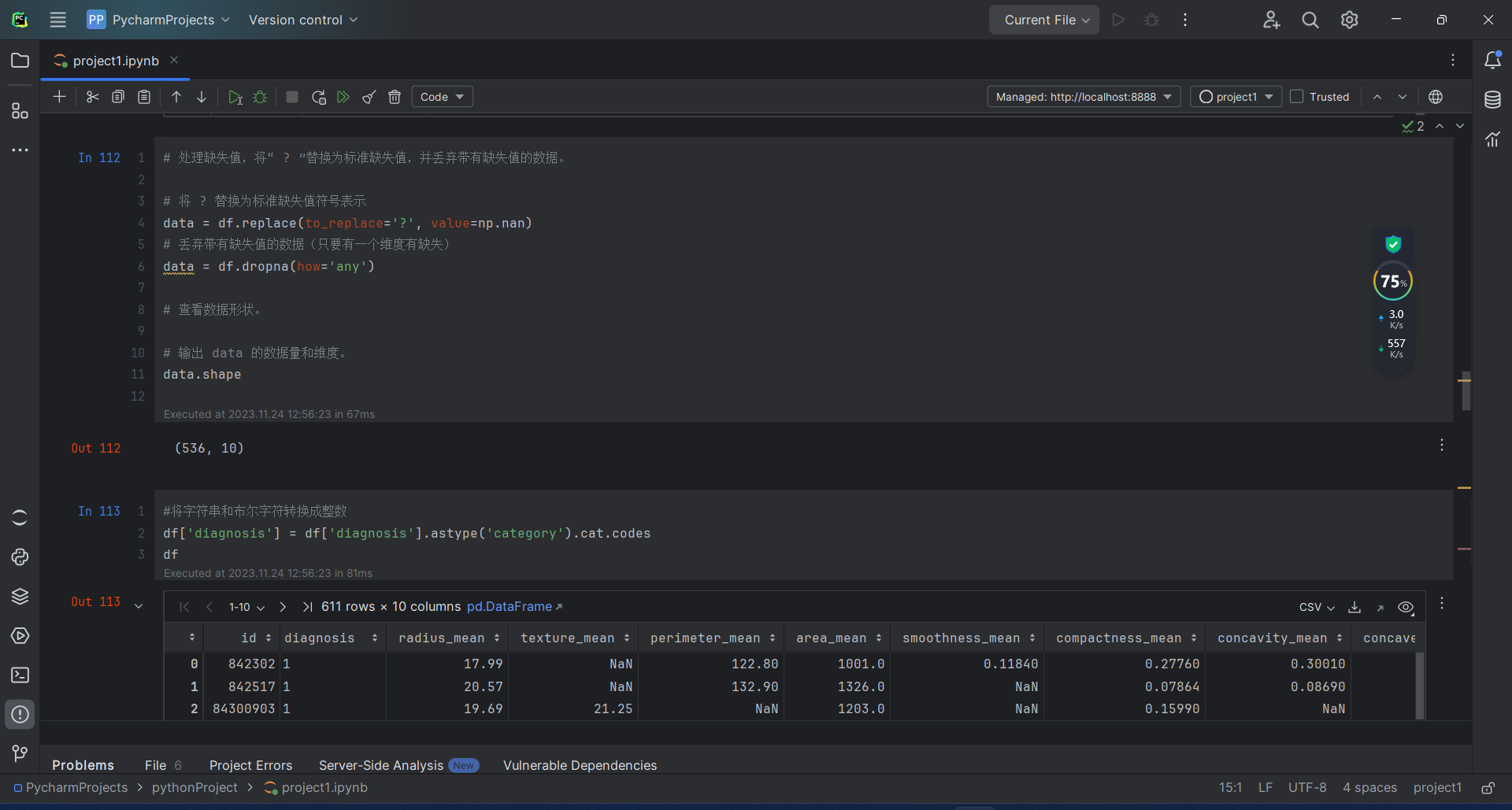


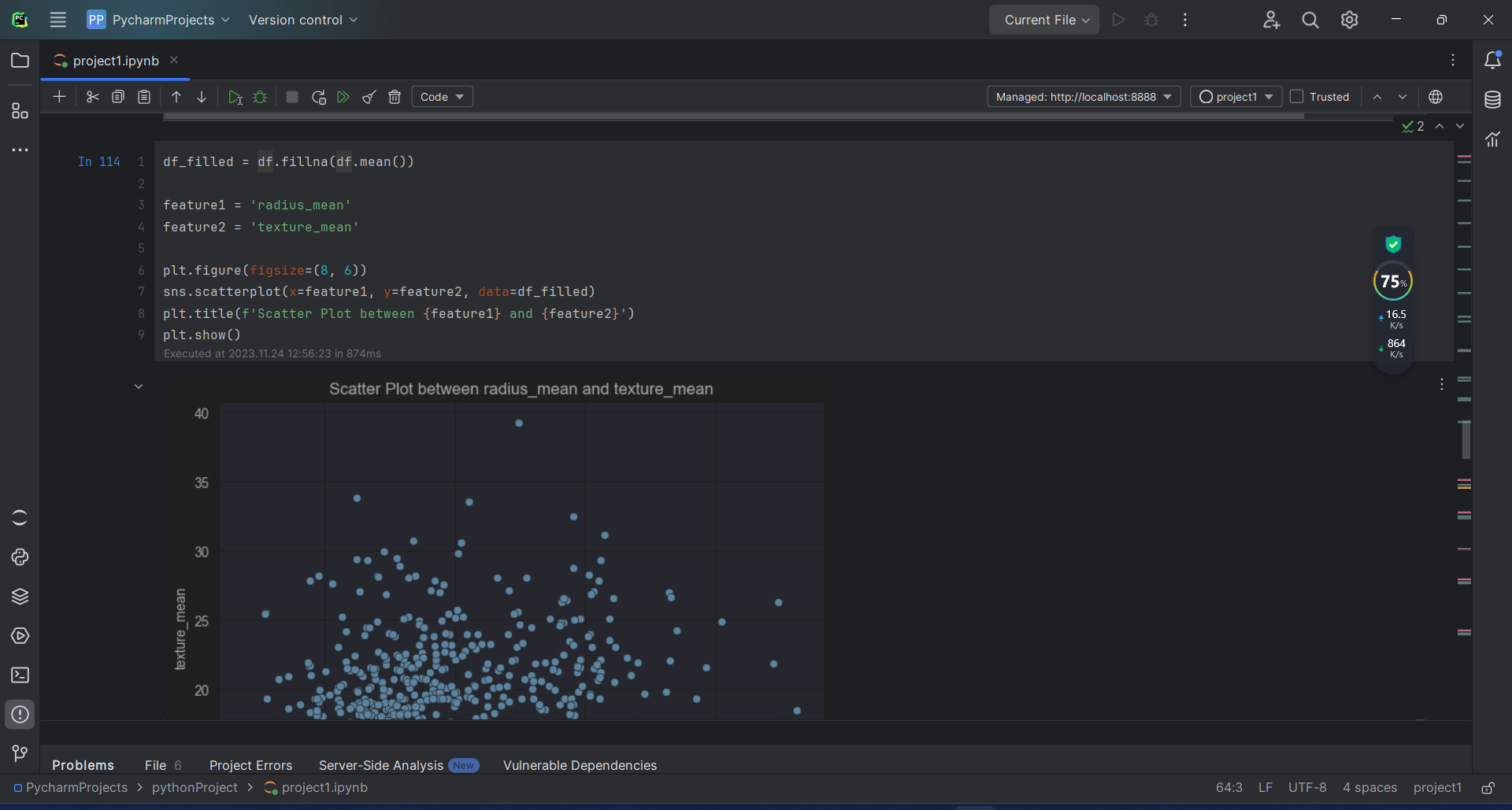




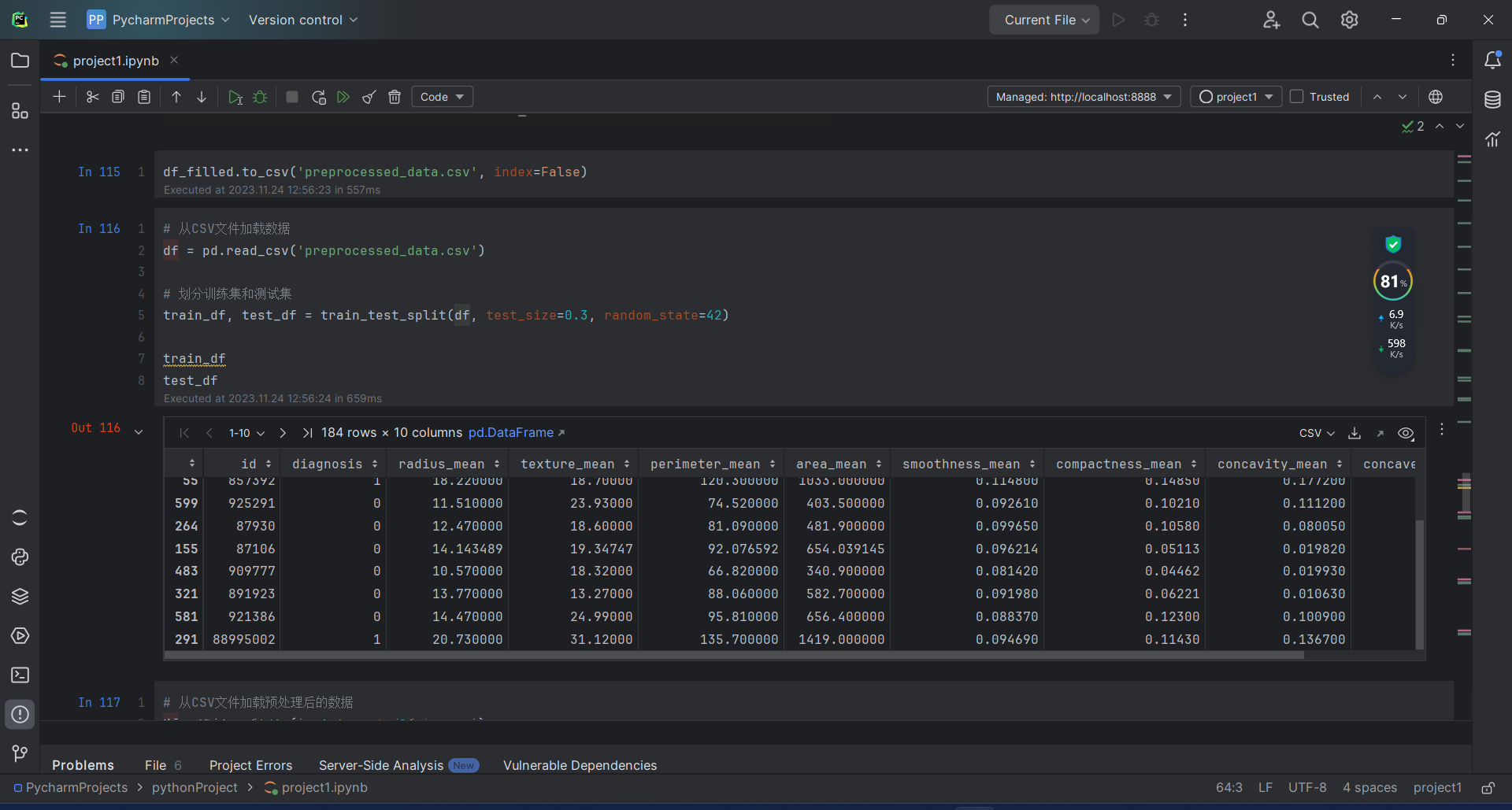




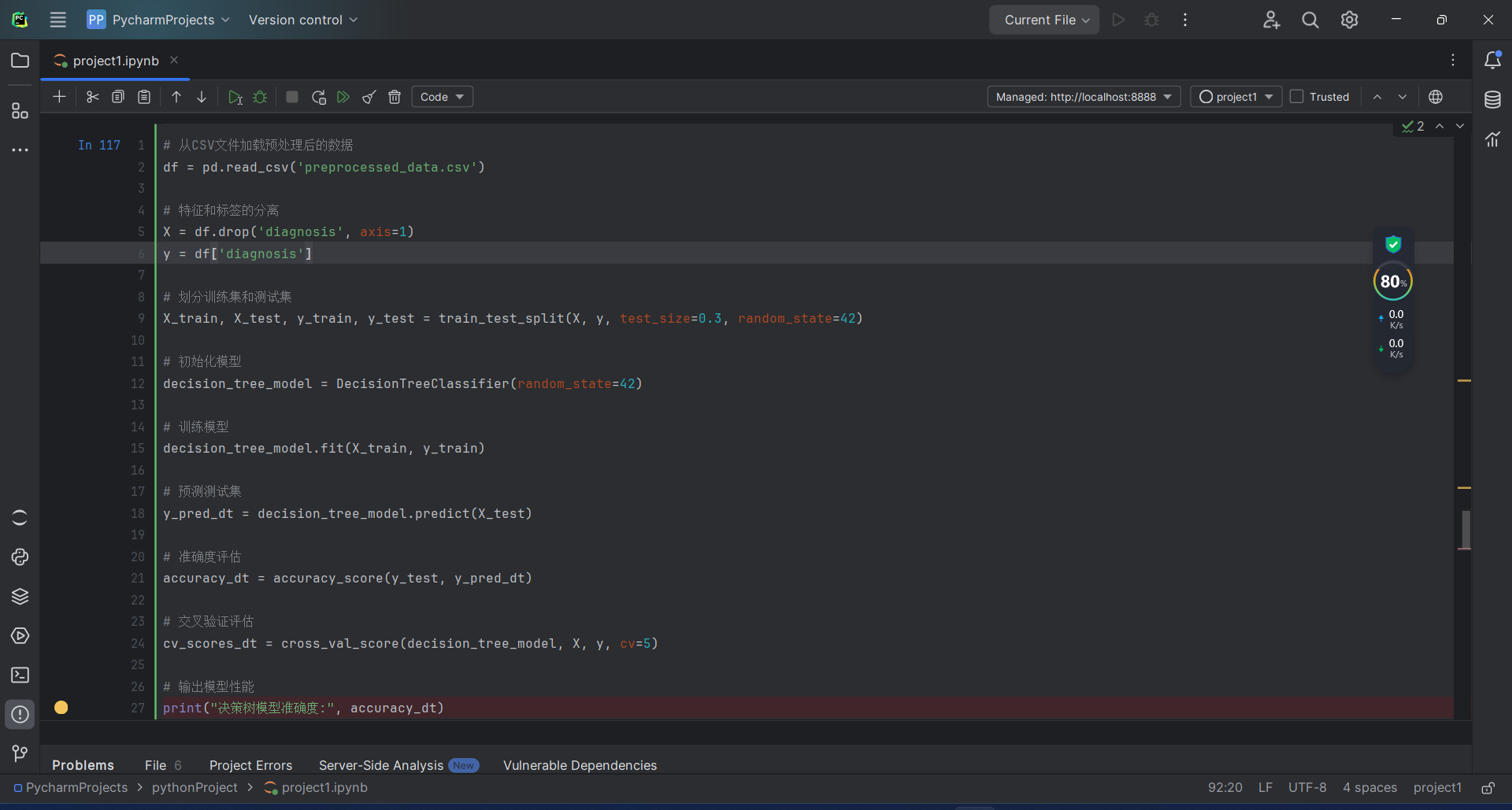


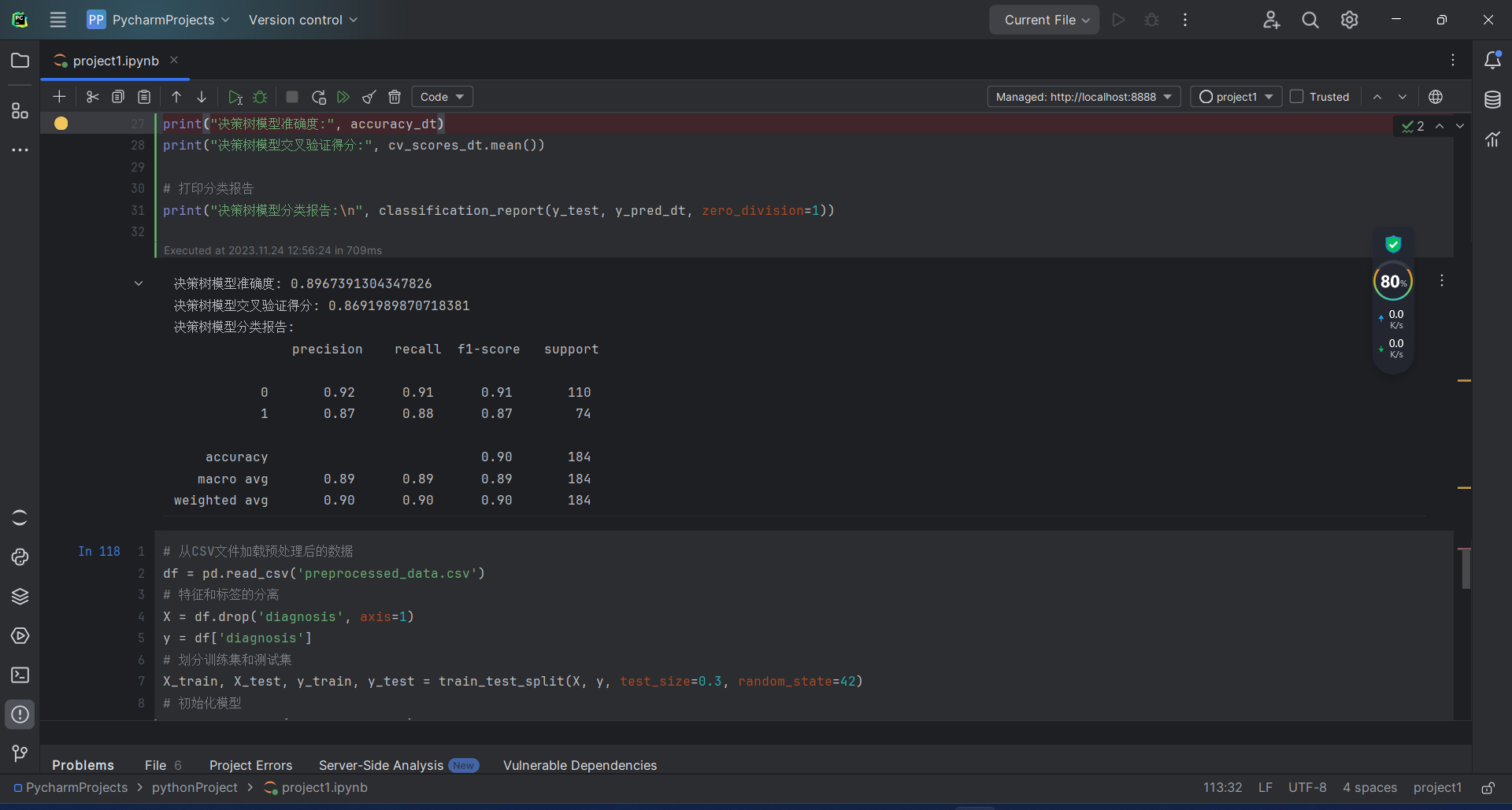


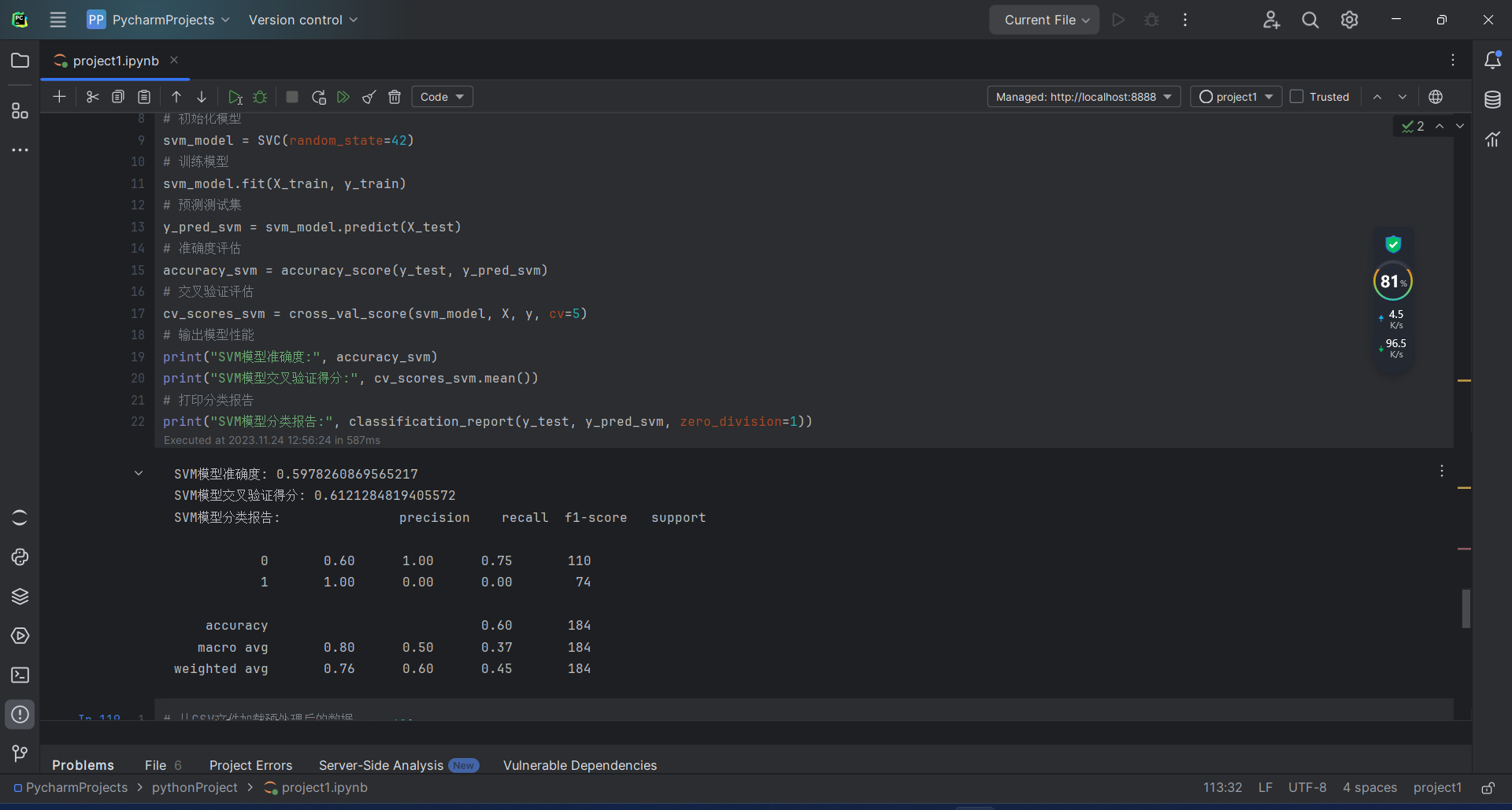
1. **数据集分割。**

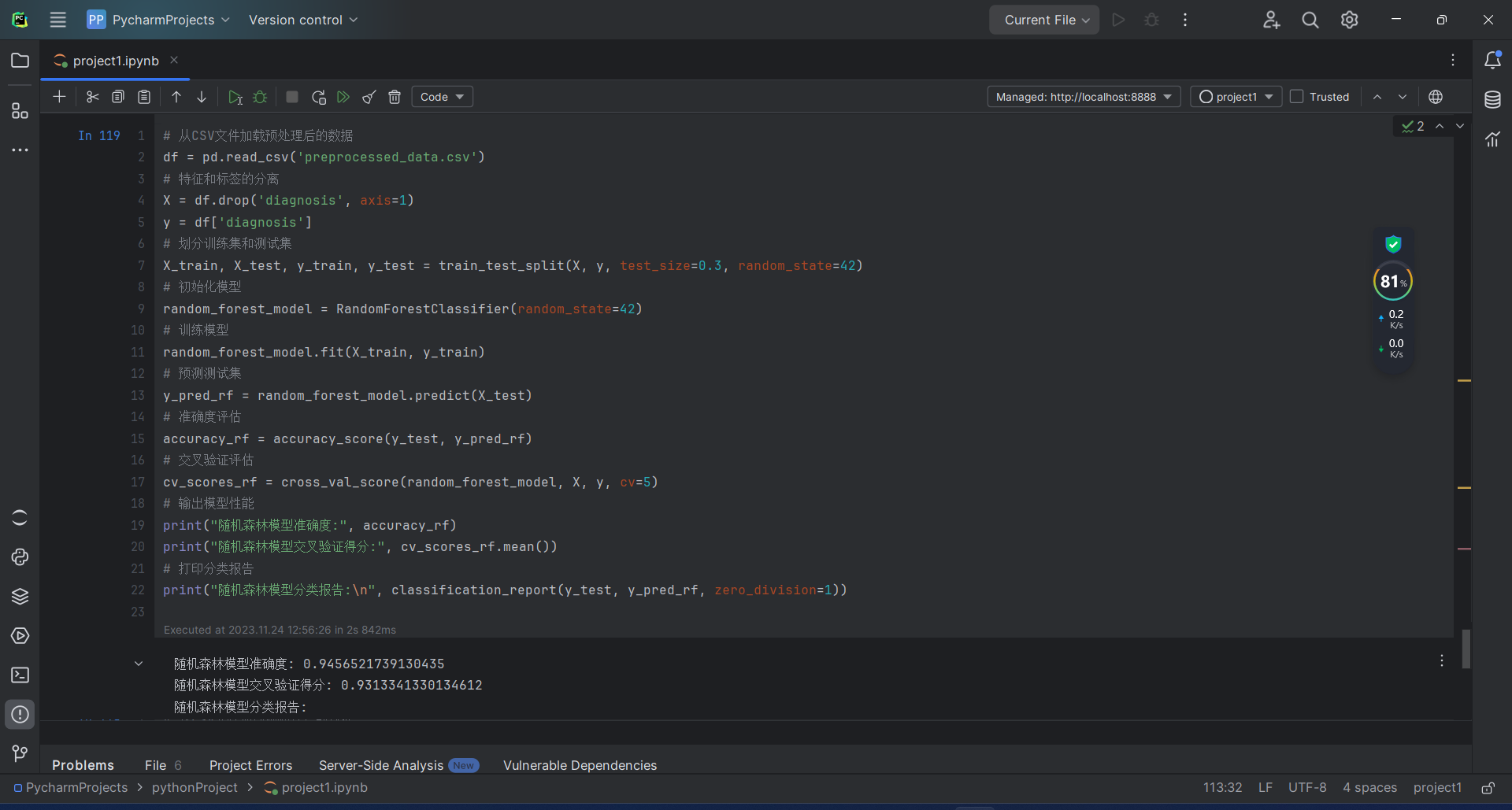


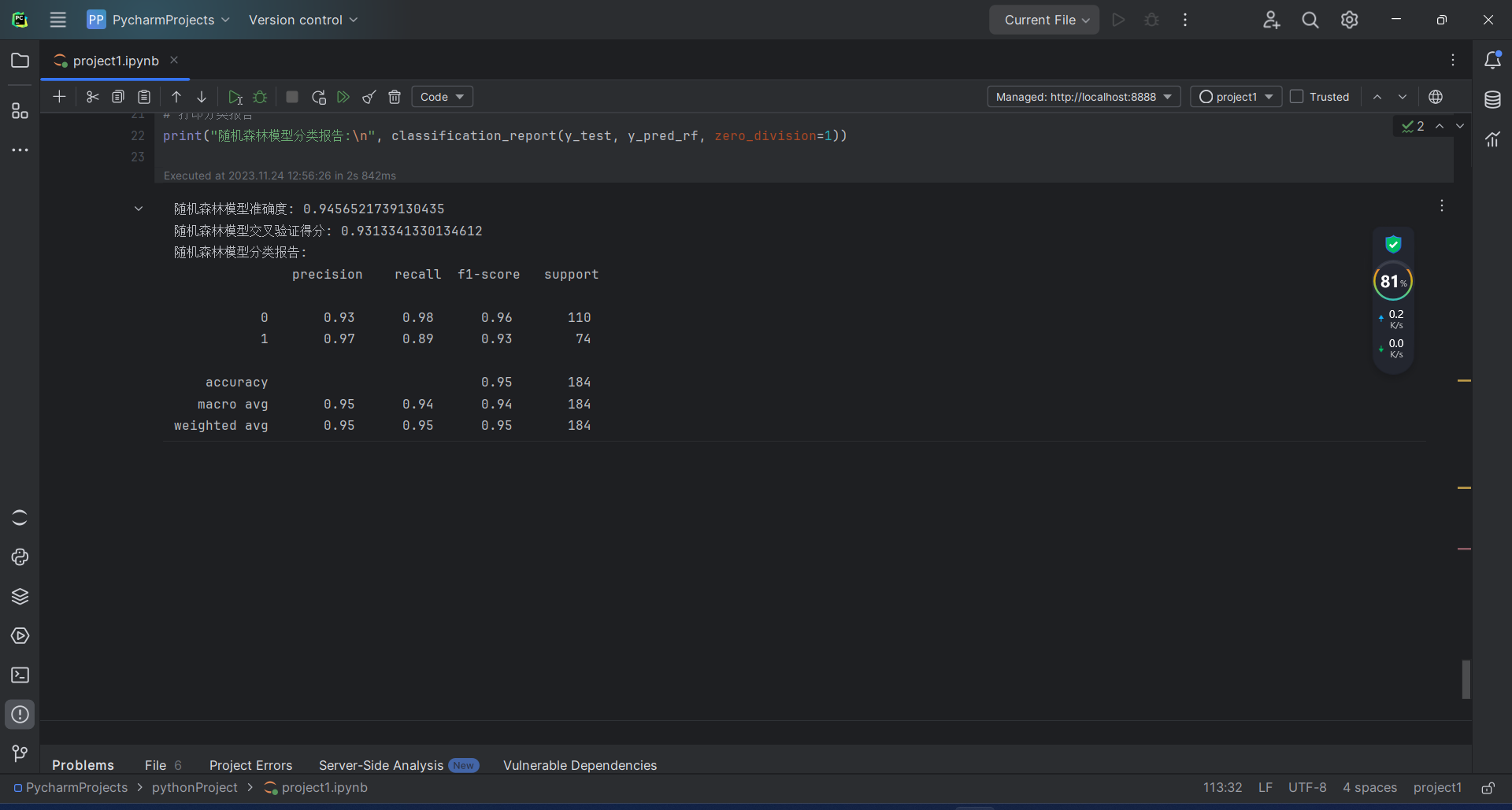
1. **模型构建、训练、评估**











1. **实验总结**

本次实验在整个过程中，对数据的深入了解、处理缺失值、进行可视化以及构建不同模型并进行比较等步骤都是非常重要的。每一步都可以影响最终模型的性能。定期检查模型的性能并进行调整，可以帮助你更好地理解问题并改进你的解决方案。不同的模型可能在不同的数据集上表现不同，

* 1. **心得体会**

在完成乳腺癌预测问题的任务中，我经历了整个机器学习项目的流程，包括环境配置、数据加载及预处理、数据集分割、模型构建、训练和评估等步骤。以下是一些我的心得体会：

环境配置在开始项目之前，正确配置和管理实验环境是非常重要的。确保使用的Python和相关库的版本是兼容的，可以避免一些潜在的问题。数据加载及预处理了解数据的基本信息、查看数据的头部和尾部、检查数据类型是开始项目的关键步骤。在预处理阶段，处理重复数据、缺失数据和进行特征编码是确保数据质量的重要步骤。通过绘制特征之间的关系图，比如热图，可以更好地理解特征之间的相关性。这对于选择特征、调整模型以及理解数据结构都是非常有帮助的。将数据集分割成训练集和测试集是为了更好地评估模型的泛化能力。确保使用合适的比例，例如，一般使用70%的数据作为训练集，30%的数据作为测试集。选择不同的机器学习模型，并训练、评估它们是项目中最核心的步骤。在这个任务中，我使用了决策树、支持向量机和随机森林这三种模型。对于每个模型，我评估了准确度、分类报告以及通过交叉验证得到的分数。通过比较不同模型的性能，可以找到最适合解决问题的模型。在这个任务中，决策树、支持向量机和随机森林可能在性能上有所不同，具体取决于数据的特性。机器学习项目是一个迭代的过程。在了解模型性能后，可以尝试进行参数调优、特征工程等操作，以进一步提高模型的性能。总的来说，通过这个任务，我深刻理解了机器学习项目的步骤和流程，学到了如何正确处理数据、选择模型并进行评估。在解决实际问题时，综合考虑数据的特性和模型的性能是非常重要的。这个实践经验对我今后处理类似问题时将会大有裨益。