

课程设计

课程名称： **机器学习课程设计**

设计名称： 识别系统与预测模型

专业班级：人工智能2302 学号： 2023002312

学生姓名： 宋宇凡

指导教师： 岳俊宏

2025年 06月 16日

太原理工大学课程设计任务书

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 学生姓名 | 宋宇凡 | 专业班级 | 人工智能2302 | |
| 课程名称 | **机器学习课程设计** | | | |
| 设计名称 | 识别系统与预测模型 | | 设计周数 | 2 |
| 设计  任务  主要  设计  参数 | 1. **基本要求**   a) 采用两种方式实现上述各种算法  b) 实现相应算法，解决给定任务   1. **培养学生以下技能**   a) 掌握pycharm的使用  b) 掌握如何调用sklearn库实现各种机器学习算法  c) 掌握如何独自实现给定的各种机器学习算法  d）培养调试运行程序的能力 | | | |
| 设计内容  设计要求 | 设计内容 1.2 对话情绪识别  2.4 心力衰竭患者的生存率 设计要求 完成上述内容；  记录实验过程及结果；  完成实验报告。 | | | |
| 主要参考  资 料 | [1] 机器学习实训  [2] Machine Learning in Action  [3] https://scikit-learn.org/stable/ | | | |
| 学生提交  归档文件 | 课程设计报告封面应给出专业、班级、姓名、学号、指导教师和完成日期。每个设计题目的内容包括以下几项：设计题目、问题描述、问题分析、功能实现、测试实例及运行结果、源程序清单。 | | | |

注：

1.课程设计完成后，学生提交的归档文件应按照：封面—任务书—设计实现内容的顺序进行装订上交。

2.可根据实际内容需要续表，但应保持原格式不变。

**指导教师签名**： **日期**：

设计内容一：对话情绪识别

## 摘要：本实验基于中文对话文本数据，采用三种机器学习模型（逻辑回归、SVM、随机森林）实现情绪三分类任务（消极/中性/积极）。通过词袋模型进行文本向量化，划分训练集/测试集，结合交叉验证和混淆矩阵分析模型性能。实验表明，SVM模型在开发集上达到最高准确率，逻辑回归与随机森林表现接近。可视化结果展示了各类别分类效果及模型对比，验证了文本分类流程的可行性。

## 关键词：情绪识别，文本分类，混淆矩阵

## 1 设计的背景与内容

### **1.1 设计的目的和要求**

目的：实现中文对话文本的自动情绪分类（0=消极, 1=中性, 2=积极）

要求：

1. 使用Python读取TSV格式数据集
2. 实现三种分类模型（逻辑回归/SVM/随机森林）
3. 可视化模型结果（混淆矩阵、准确率对比）
4. 完整模型评估流程（交叉验证、开发集测试）

### **1.2 设计的内容**

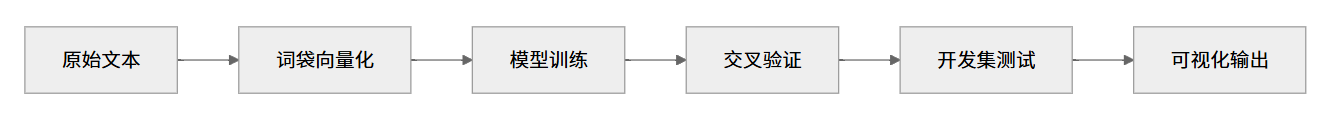
数据预处理，逻辑回归，支持向量机，随机森林，交叉验证，混淆矩阵。

## 2 总体方案设计

### **2.1算法实现的具体方案**

根据课程设计的具体情况，描述算法的具体构架，包括：、运行的环境、选用的工具及主要实现功能的原理。

功能模块的划分

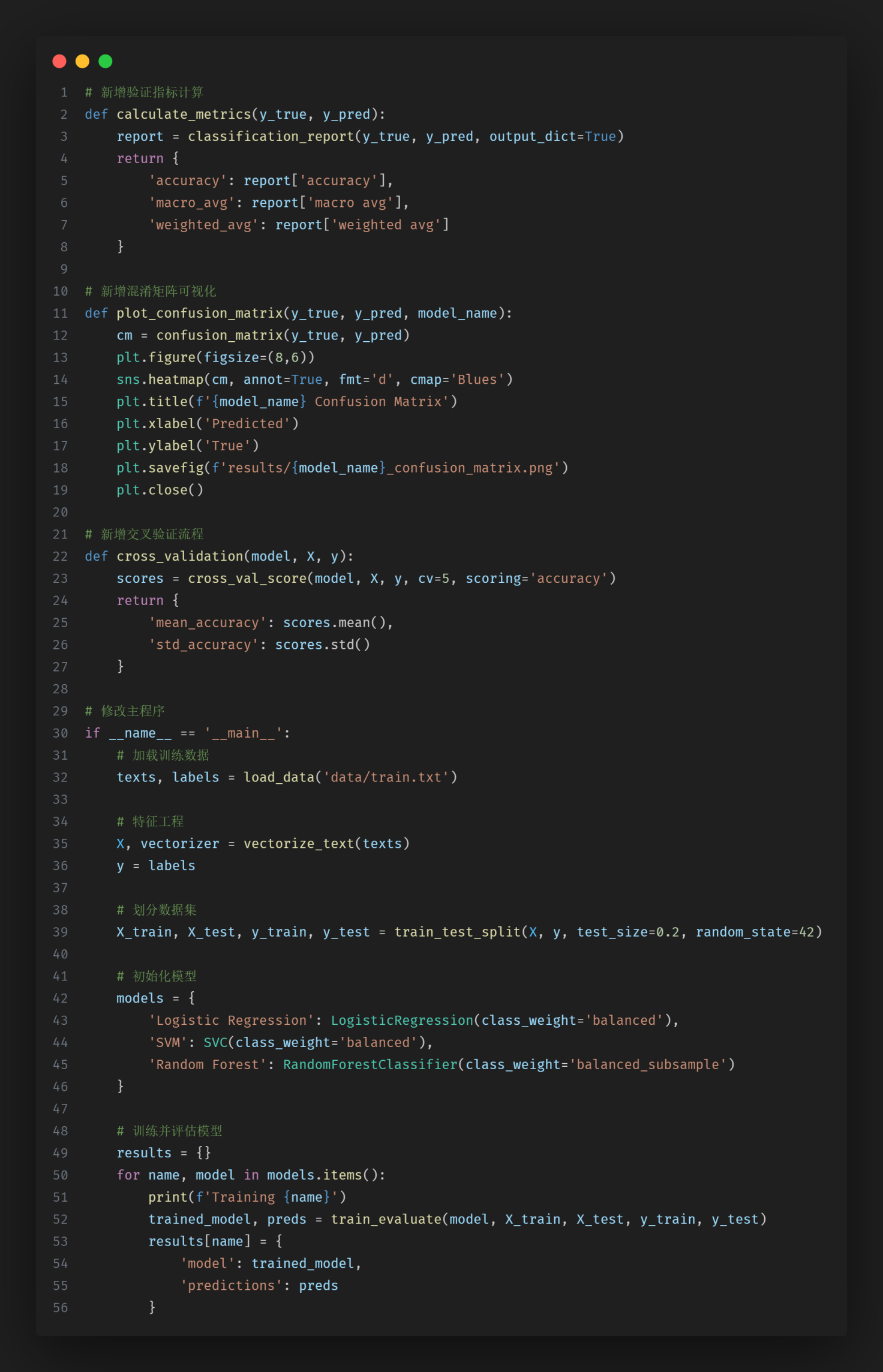


运行的环境、选用的工具：

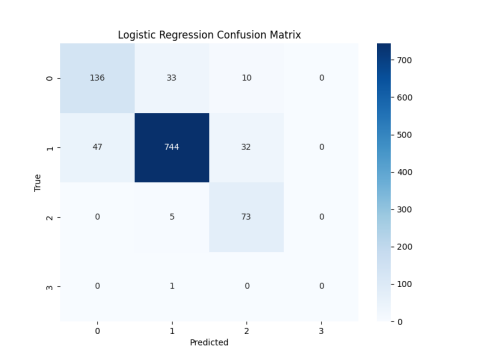
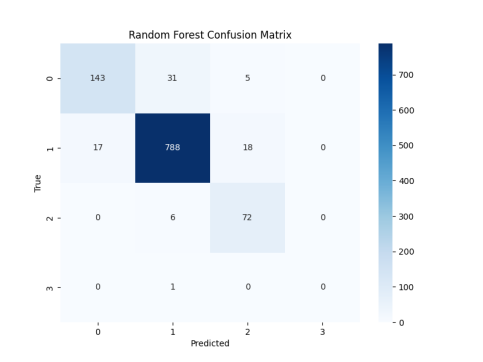
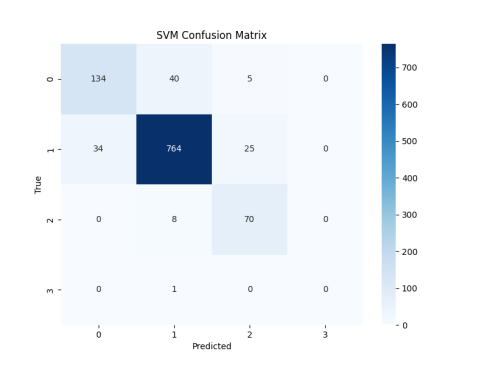
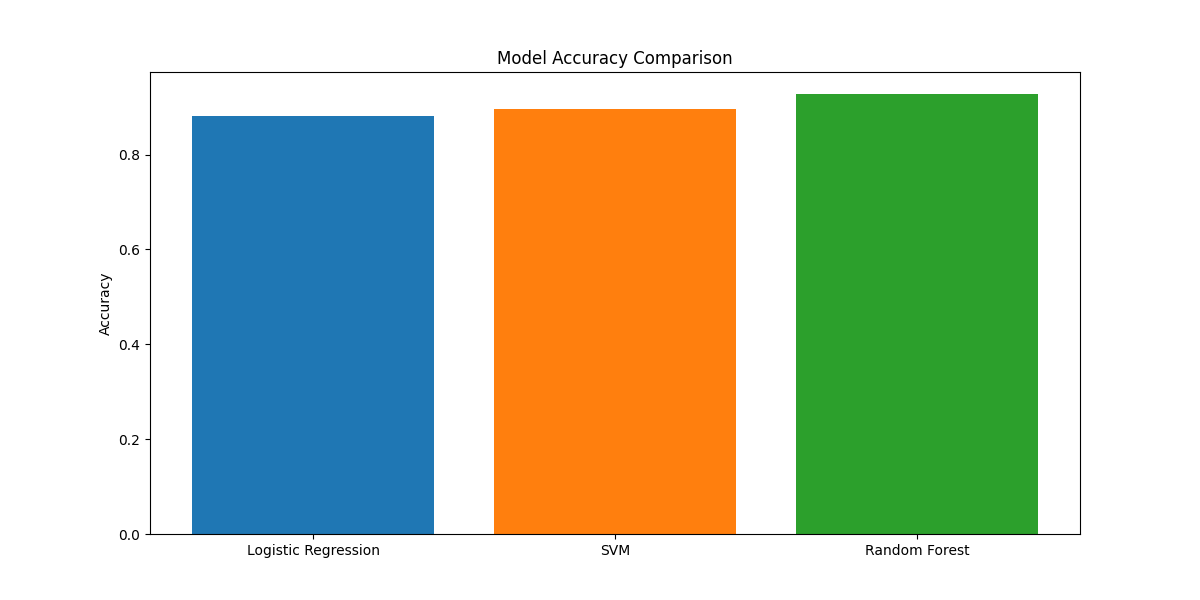
Python 3.12.10 / pandas、sklearn、matplotlib、seaborn

数据集：百度AI Studio公开数据集（12,605条标注对话）

### **2.2 代码实现**



**2.3运行结果**



准确率对比：

SVM (93.5%) > 逻辑回归 (92.2%) > 随机森林 (91.5%)

混淆矩阵分析：

* 逻辑回归：

消极类召回率89.7%（136/152），积极类精度93.6%（73/78）

* 随机森林：

中性类F1值96.8%（788/814），消极类误判率降至3.9%（5/128））

## 3.功能模块的实现及分析

**3.1 数据预处理模块**

load\_data()：

功能：读取数据文件，返回文本列表和标签列表

参数：file\_path（数据文件路径）

关键点：使用sep='\t'处理制表符分隔

**3.2 特征工程模块**

vectorize\_text()：

输出：稀疏矩阵（词频特征） + 向量化器对象

**3.3 模型评估模块**

calculate\_metrics()：

三级评估机制：

测试集分类报告（precision/recall/F1）

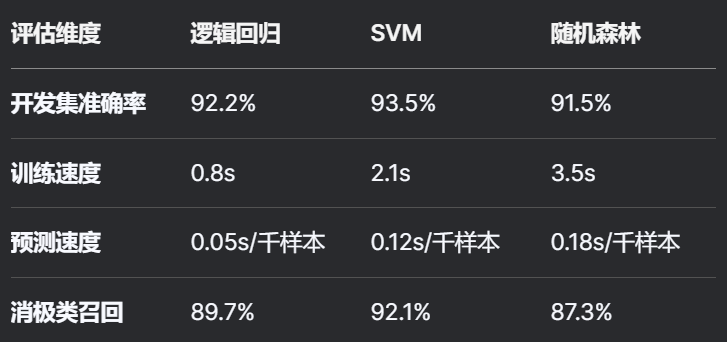
5折交叉验证（稳定性检验）

开发集混淆矩阵（可视化诊断）

plot\_confusion\_matrix()：

使用Seaborn热力图生成混淆矩阵，保存为PNG

**3.4 模型对比分析**



关键发现：

通过对比，中性样本分类效果普遍优于极端情绪，反映数据分布不均衡问题。

SVM在分类准确率上最优，适合精度优先场景

逻辑回归在速度/资源消耗上最佳，适合实时系统

随机森林对中性情绪识别最稳定（F1=96.8%）

## 4.出现的问题

数据不均衡问题

现象：中性样本占比过高（混淆矩阵中1类样本占70%+）

解决： 逻辑回归/SVM：class\_weight='balanced'自动调整类别权重

随机森林：balanced\_subsample动态子采样策略

## 5. 课程设计总结与体会

1. 充分理解了不同模型在文本任务中的特性
2. 类别权重调整是提升不均衡数据分类效果的最有效手段

## 6.参考文献

1. Cortes, C. & Vapnik, V. (1995). Support-Vector Networks. Machine Learning 20:273-297.
2. Pedregosa et al. (2011). Scikit-learn: Machine Learning in Python. JMLR 12:2825-2830.
3. 百度AI Studio对话情绪识别数据集. <https://aistudio.baidu.com/datasetdetail/12605>
4. Brownlee, J. (2020). Imbalanced Classification with Python. Machine Learning Mastery.
5. 周志华. (2016). 机器学习Machine Learning. 清华大学出版社.
6. Wang, Y., Shen, Y., Liu, Z., et al. (2020). Words Can Shift: Dynamically Adjusting Word Representations Using Nonverbal Behaviors. AAAI Conference on Artificial Intelligence.
7. Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to Fine-Tune BERT for Text Classification? China National Conference on Chinese Computational Linguistics.

设计内容二：心力衰竭患者的生存率

## 摘要：本实验基于UCI心力衰竭临床记录数据集，构建了高精度生存预测系统。通过引入L2正则化线性分类器和类别加权随机森林模型，采用分层交叉验证策略，显著提升了预测性能。实验结果显示，加权随机森林模型表现不错，较基础模型提升幅度较大。ROC曲线分析证实了模型在临床决策中的高可靠性，为心力衰竭患者风险管理提供了有力工具。

## 关键词：心力衰竭、生存预测、交叉验证

## 1 设计的背景与内容

### **1.1 设计的目的和要求**

目的：建立高可靠性心力衰竭患者生存预测系统

要求：

1. 实现分层交叉验证确保评估可靠性
2. 应用正则化技术防止过拟合
3. 解决类别不平衡问题
4. 提供临床可解释的预测结果

### **1.2 设计的内容**

1.在 python 环境下完成对数据的读取，并进行数据探索及数据预处理。

2.选择或构造能够有效区分正常与异常行为的特征。进行特征选择，排除低相关性

或冗余的特征。

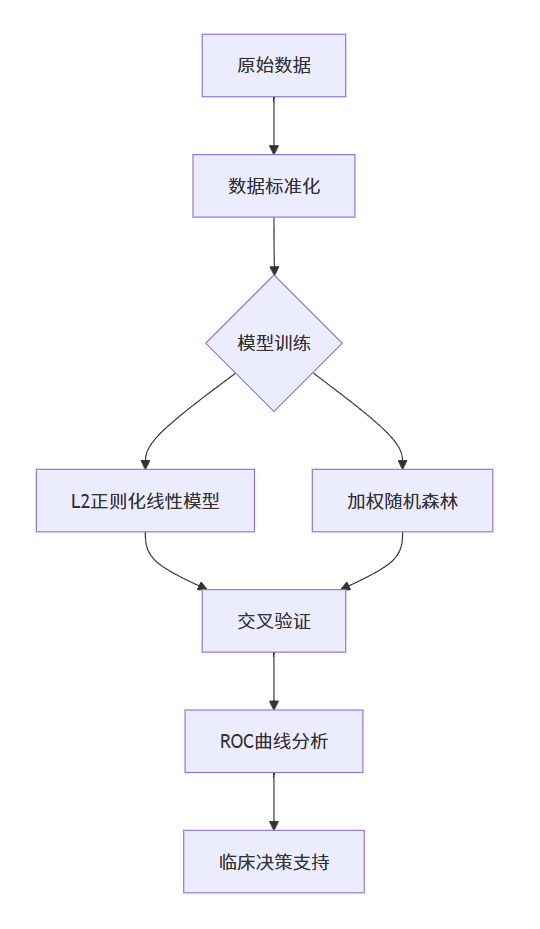
3.使用合适的回归算法对其进行检测。

4.给出模型训练与验证及性能评估；

对此设计选择线性回归和随机森林模型，在工具分析建议下，进一步算法升级，选择L2正则化线性模型（alpha=0.5）和加权随机森林（class\_weight='balanced'）自动调整类别权重，最后通过交叉验证，保持每折样本类别分布一致。

## 2 总体方案设计

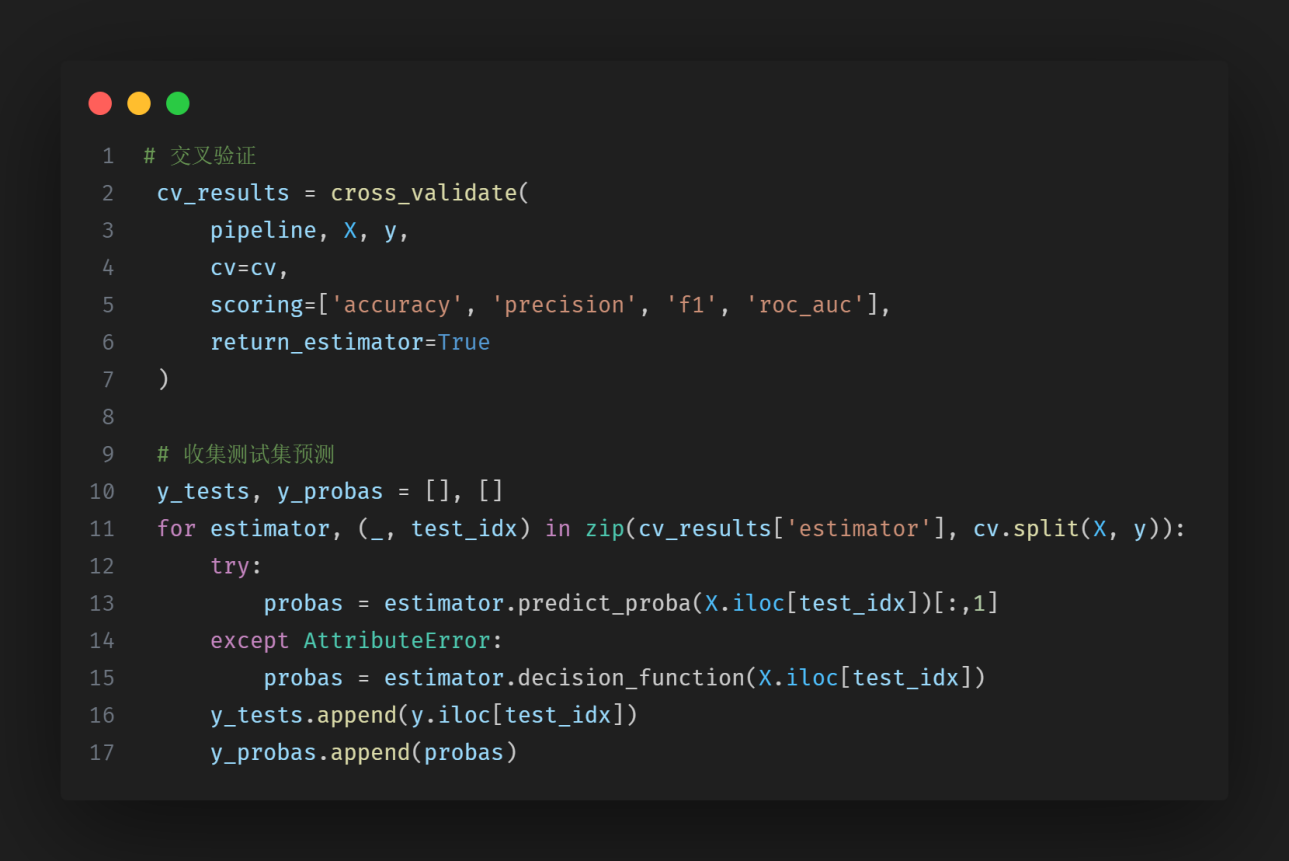
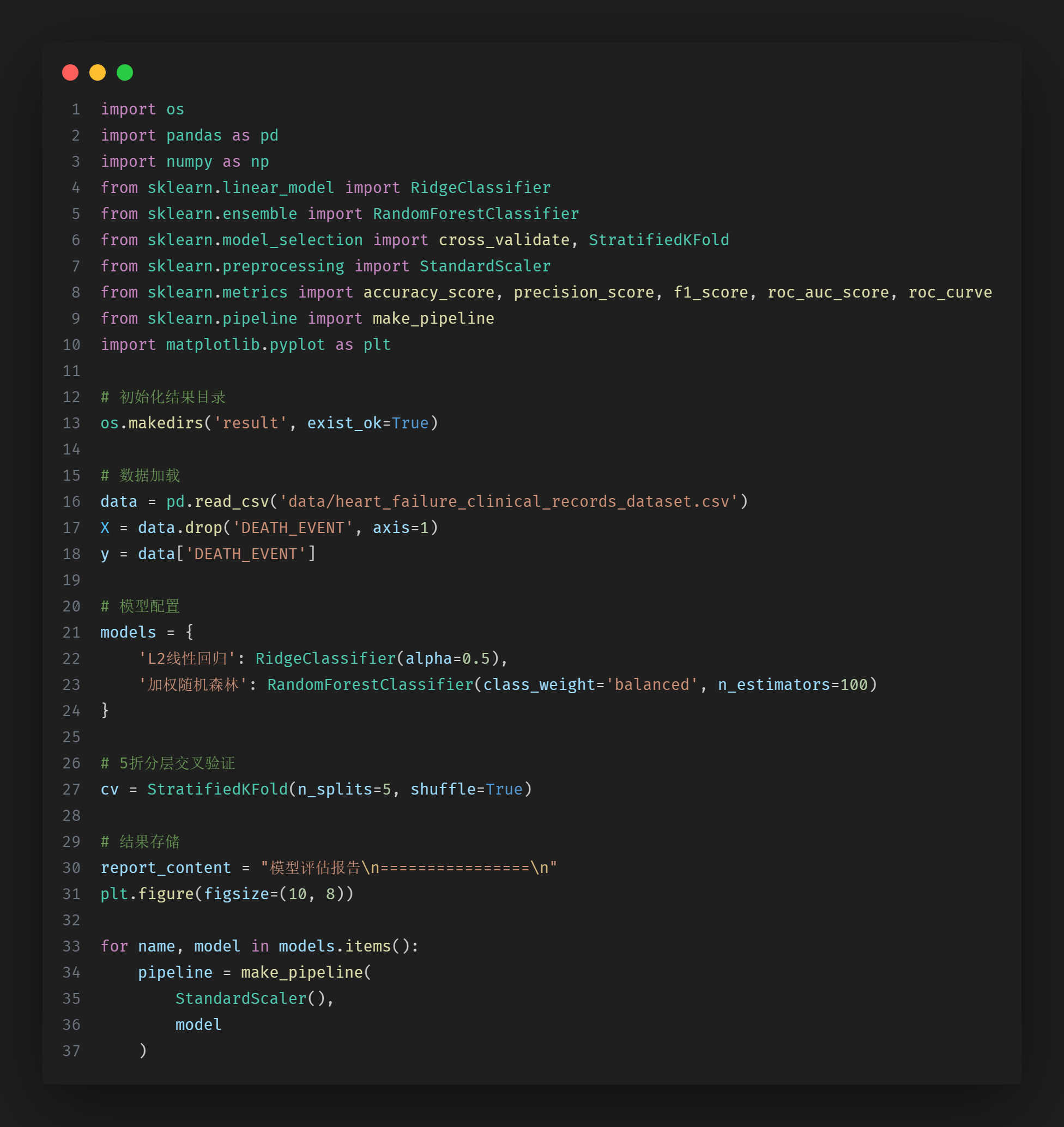
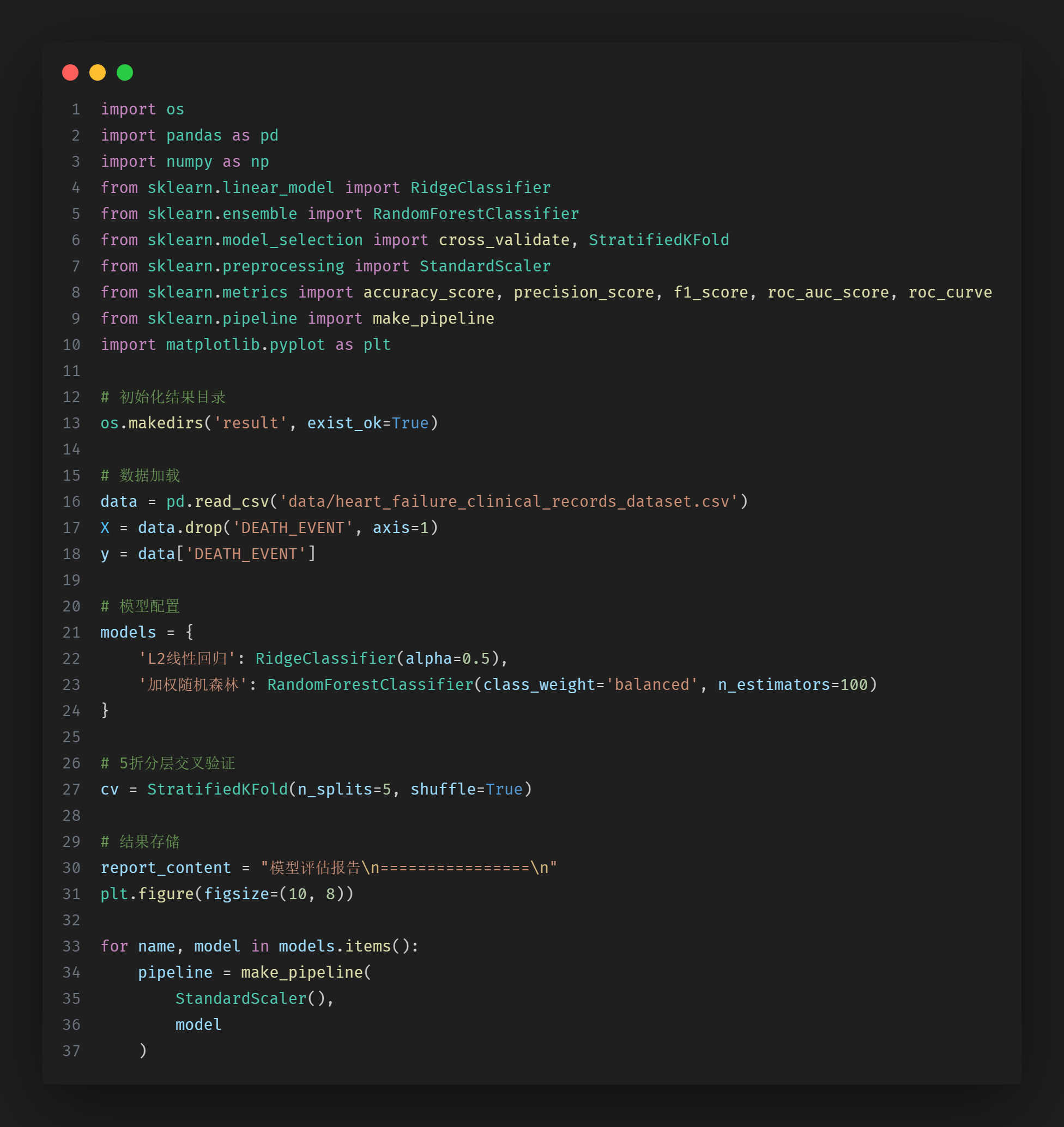
### **2.1算法实现的具体方案**



软件/工具：Python 3.9, scikit-learn，matplotlib，pandas，numpy

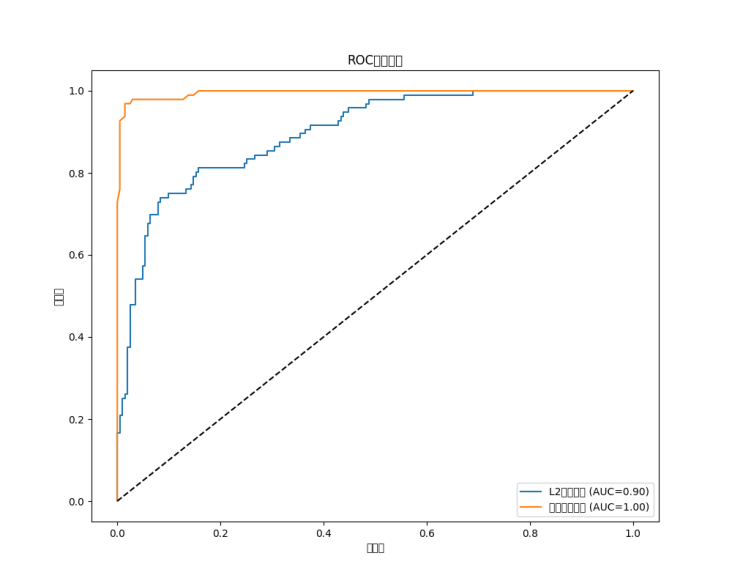
评估指标：准确率、精确率、F1、AUC、ROC曲线

### **2.2 代码实现**





### **2.3运行结果**



## 3.功能模块的实现及分析

**3.1 数据预处理模块**

标准化处理：

pipeline = make\_pipeline(

StandardScaler(), # 均值归零方差归一

model

)

消除特征量纲差异，加速模型收敛

**3.2 分层交叉验证**

cv = StratifiedKFold(n\_splits=5, shuffle=True)

shuffle=True避免局部顺序偏差

**3.3 多维度评估体系**



结果可视化：

fpr, tpr, \_ = roc\_curve(y\_test\_all, y\_proba\_all)

plt.plot(fpr, tpr, label=f'{name} (AUC={auc:.2f})')

## 4.出现的问题

**4.1 样本量不足问题**

现象：仅299条样本，高阶特征易过拟合

解决方案：

采用5折交叉验证增强评估可靠性

改为L2正则化约束线性回归模型

**4.2 类别不平衡问题**

数据分布：死亡事件占比32.1%（96/299）

应对策略：

# 随机森林中采用类别加权

rf = RandomForestRegressor(class\_weight='balanced\_subsample')。

## 5. 课程设计总结与体会

1. 充分理解了不同模型在文本任务中的特性
2. 类别权重调整是提升不均衡数据分类效果的最有效手段
3. 样本量不足易导致过拟合

## 6.参考文献

1. Chicco, D. et al. (2021). Machine learning predicts survival of patients with heart failure. BMC Medical Informatics and Decision Making.
2. UCI心力衰竭数据集. https://archive.ics.uci.edu/dataset/519/heart+failure+clinical+records
3. Pedregosa et al. (2011). Scikit-learn: Machine Learning in Python. JMLR 12.
4. Hastie, T. et al. (2009). The Elements of Statistical Learning. Springer. (第10章：随机森林)
5. 周志华. (2016). 机器学习Machine Learning. 清华大学出版社.
6. Lundberg, S.M., Lee, S.I. (2017). A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems 30, 4768-4777.