# 软件设计说明书

1. 软件名称

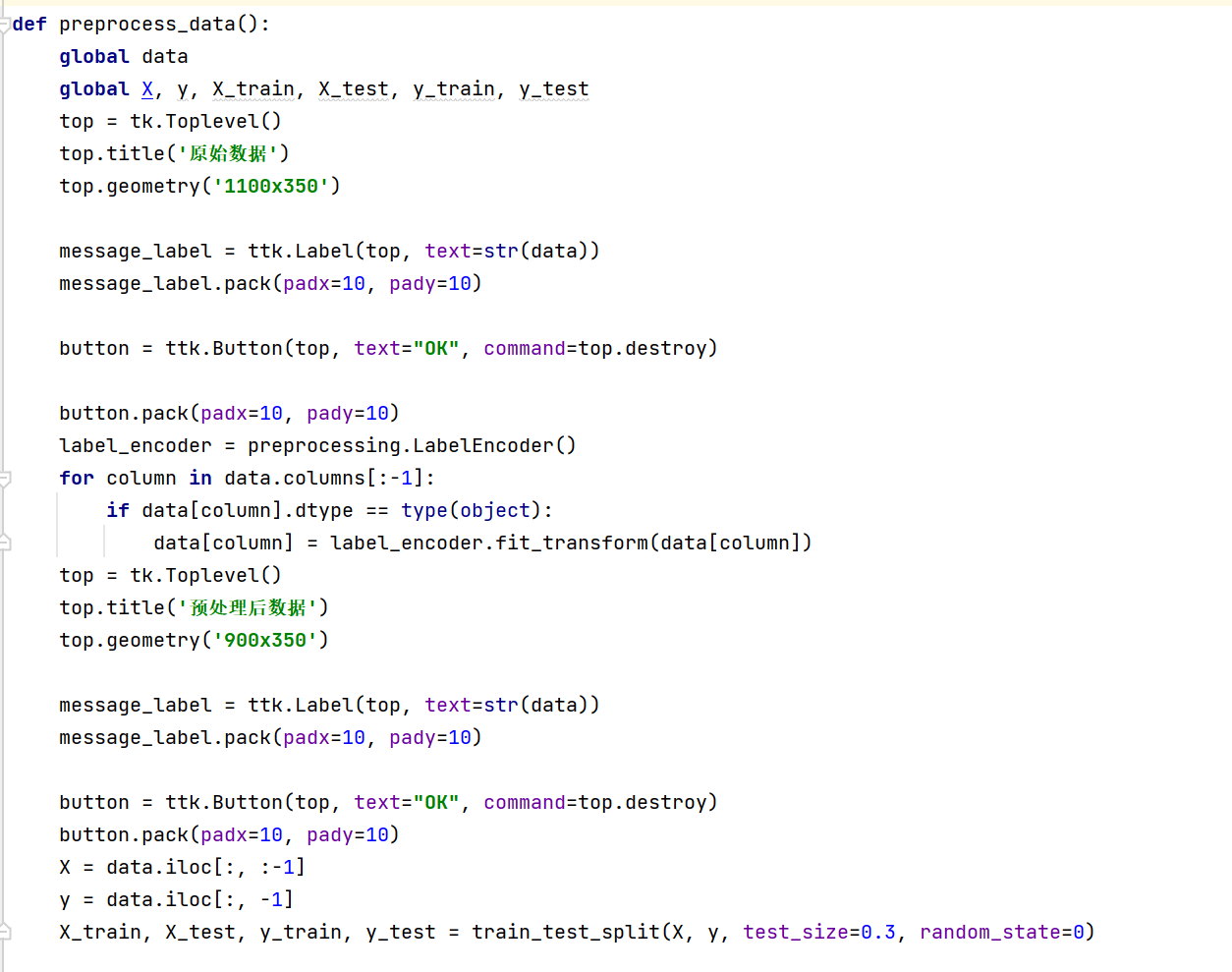
基于支持向量机的收入预测软件

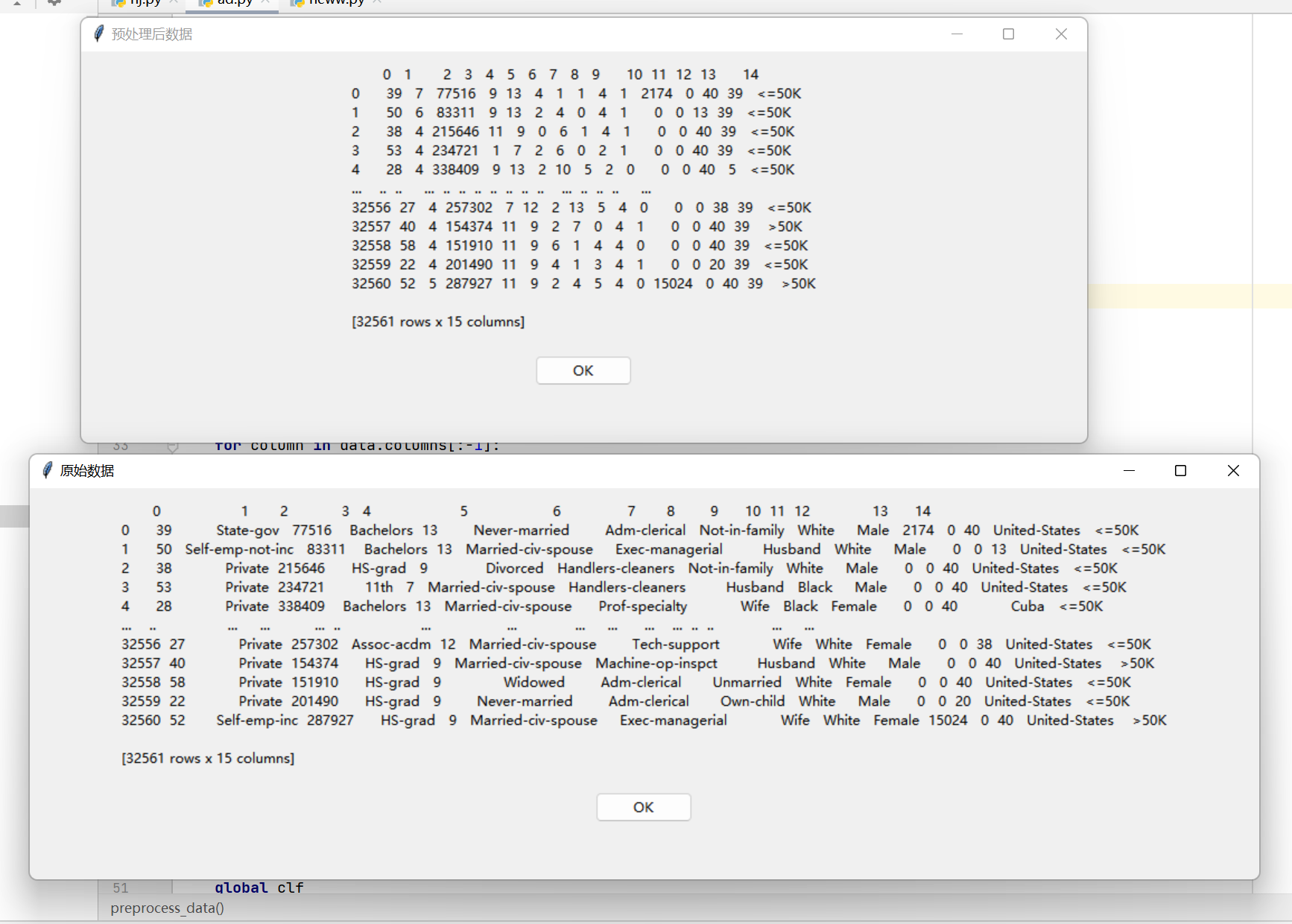
2. 功能介绍

本软件基于Python开发，使用Tkinter库构建图形用户界面，用于对UCI的Adult数据集进行数据预处理、模型训练、模型预测和模型评估。数据集40000+条数据，使用支持向量机人口普查数据预测一个人的收入是否超过50K。

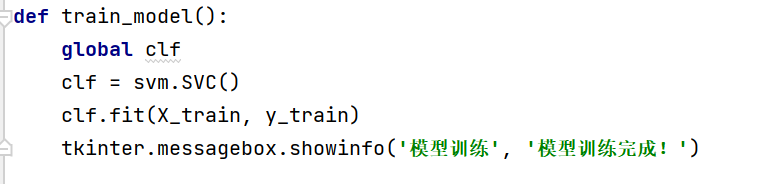
软件包含以下四个主要功能：

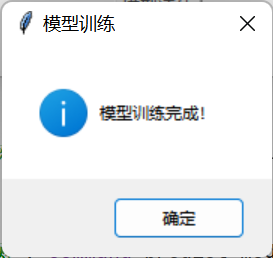
数据预处理：对Adult数据集进行预处理，包括将非数值型数据编码为数值型数据，展示处理前和处理后的数据。然后将数据集通过随机划分分为训练集和测试集。



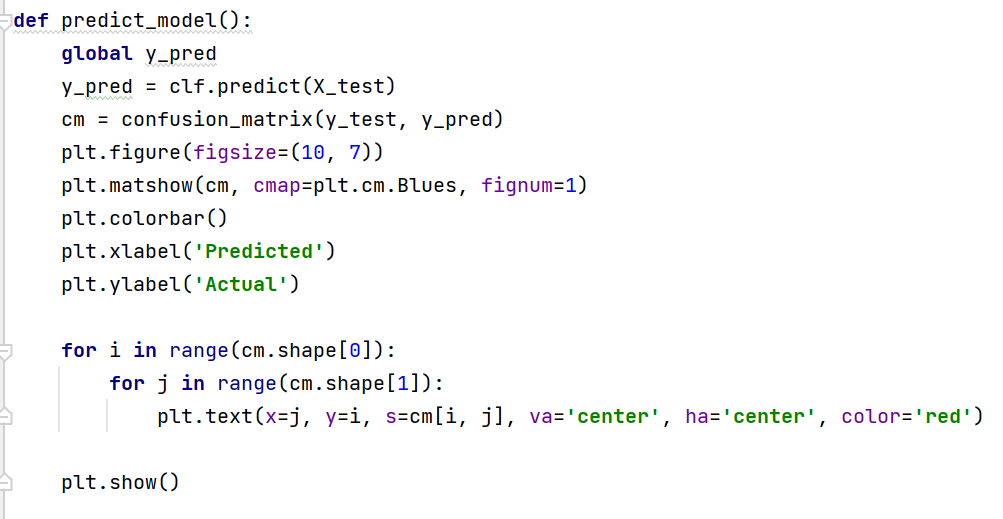


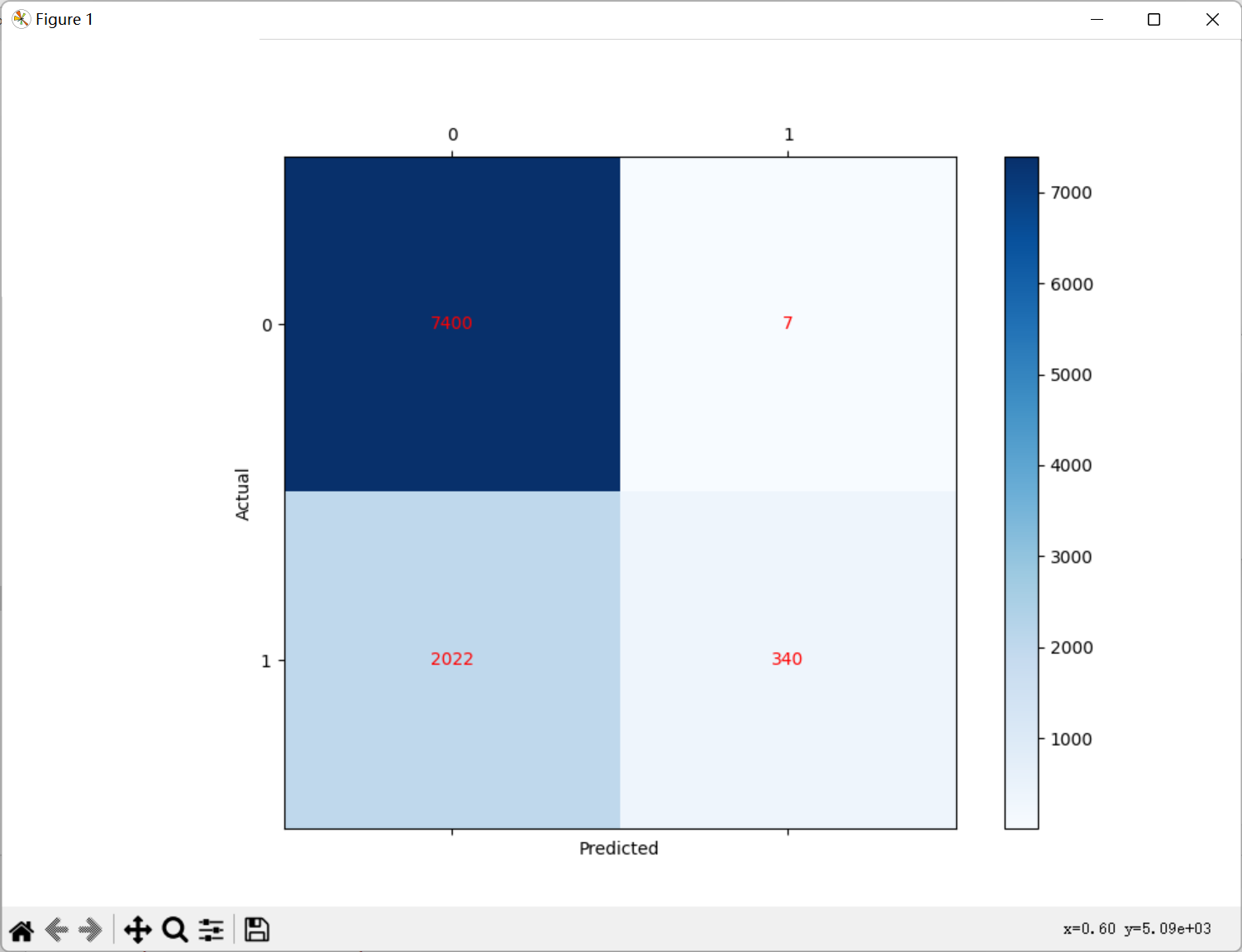
模型训练：使用支持向量机（SVM）算法在训练集上训练模型。



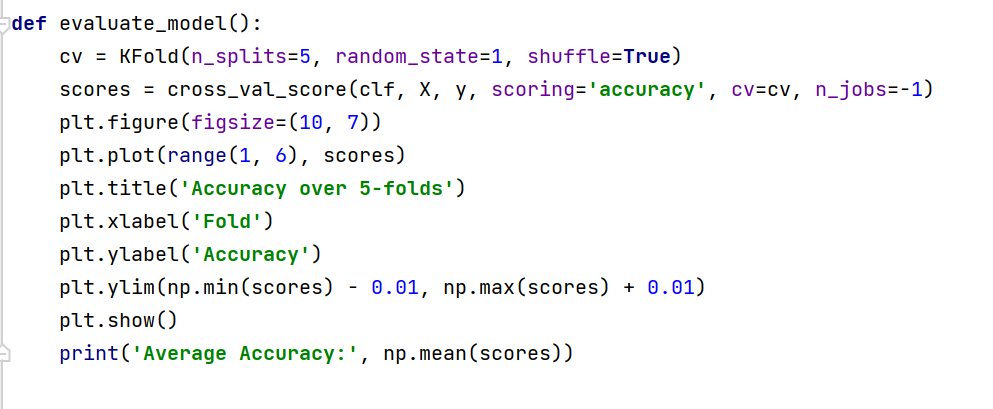


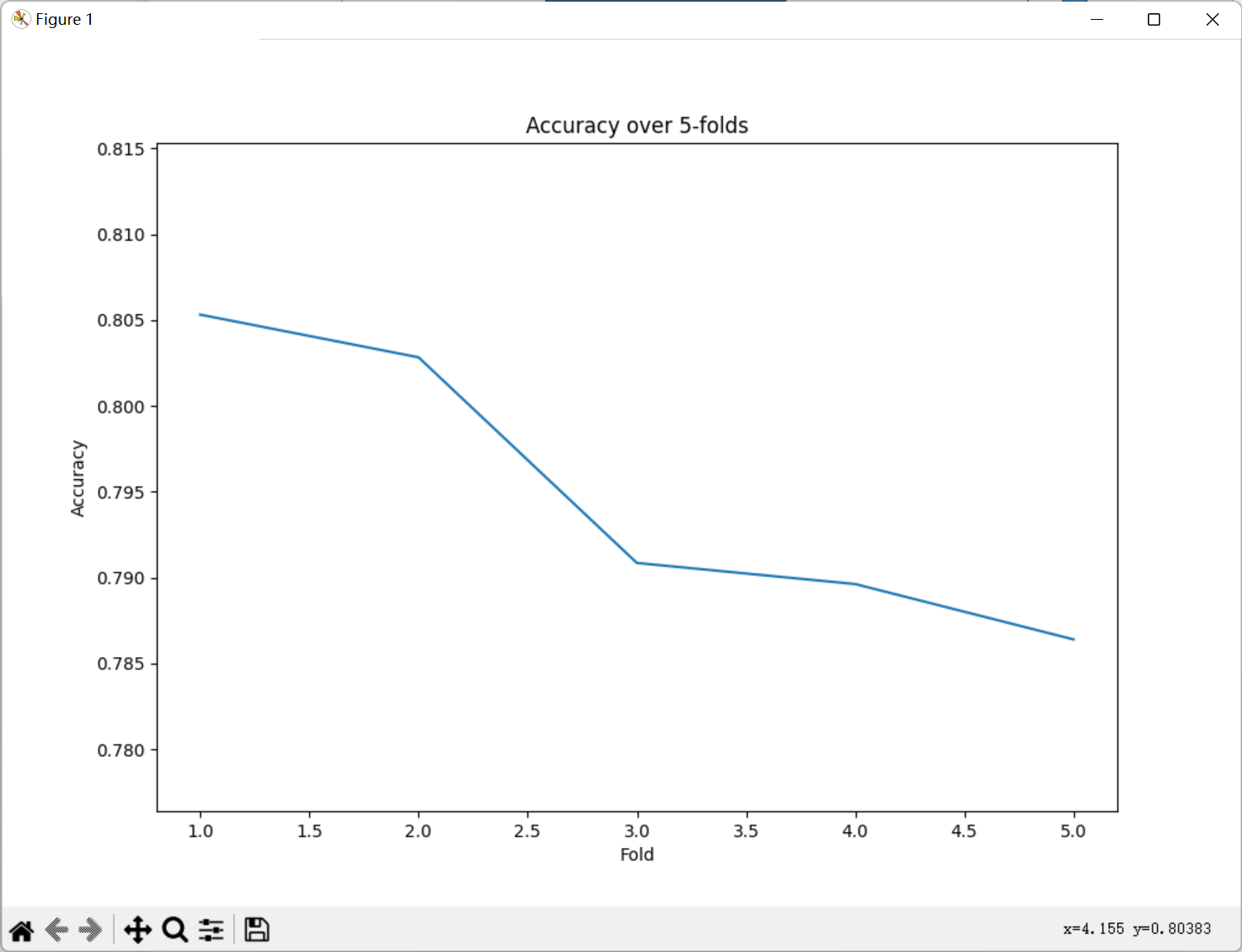
模型预测：在测试集上进行预测，并展示混淆矩阵。





模型评估：对模型进行5折交叉验证，并绘制平均准确率折线图。







3. 使用说明

点击“数据预处理”按钮，软件会自动读取并编码Adult数据集，同时将数据分割为训练集和测试集。

点击“模型训练”按钮，软件将使用SVM算法在训练集上训练模型。

点击“模型预测”按钮，软件会在测试集上进行预测，并展示混淆矩阵。

点击“模型评估”按钮，软件会进行5折交叉验证，并绘制平均准确率折线图。

1. 算法原理

支持向量机（SVM）：

支持向量机（SVM）是一种非常强大且灵活的有监督学习模型，既可以用于线性分类任务，也可以用于非线性分类任务。它基于统计学习理论中的结构风险最小化原则，其目标是找到一个超平面以便使得正例和反例尽可能地被正确分类，而且两类样本间隔尽可能地大。

混淆矩阵：

混淆矩阵是一种特定的表格布局，允许可视化算法的性能，通常用于监督学习（分类）。在混淆矩阵中，每一列代表预测类别，每一行代表实际类别。矩阵的维数等于分类问题中的类别数。主对角线元素表示正确分类的数量，非主对角线元素表示错误分类的数量。

K折交叉验证：

K折交叉验证是一种评估模型性能的方法，用于避免过拟合。在K折交叉验证中，原始样本被随机分为K组或称为"折"（folds）。然后，对模型进行K次训练和验证，每次选择不同的折作为验证集，其余的K-1个折作为训练集。得到K个模型性能度量值，通常取这K个值的平均作为模型性能的估计。一般情况下，选择K=5或10。