GeoSOS-AI 用户手册

1 软件介绍

地理模拟与优化系统（Geographical Simulation and Optimization System，GeoSOS）理论，是根据黎夏教授、叶嘉安院士、刘小平教授及其团队多年来在地理元胞自动机、多智能体建模和空间优化研究的基础上，由黎夏教授提出的[1]。GeoSOS 作为目前唯一耦合了地理模拟和空间优化的模型，提供一整套理论、方法和工具用于模拟、预测和优化复杂地理格局和过程。可以用于全球土地利用变化、城市扩张模拟、公共设施选址选线、生态红线和城市增长边界划定等地理模拟和空间优化工作[2]，能够帮助学术界、规划业界和政府部门针对以上问题进行更科学、更智能的研究和决策。

GeoSOS-AI 是为了在当前人工智能/地理人工智能（AI/GeoAI）飞速发展的背景下，为用户提供跨平台的机器学习、GeoAI 等功能而设计和研发的 GeoSOS 系列软件的最新成员。由黎夏教授进行理论指导，李丹总负责其软件的设计和开发。目前 GeoSOS-AI 的最新版本为 1.0.1a，已经完成了主要的机器学习和部分深度学习等功能，更多的 GeoAI 功能将在后续不断更新。

2 软件设计及数据准备

2.1 软件设计

2. 1. 1 软件总体设计

GeoSOS-AI 是支持机器学习、GeoAI 多种算法的 GeoSOS 系列软件的最新成员。利用基于 Python 的机器学习库 scikit-learn、Keras 和 TensorFlow 等相关软件组件库，通过 wxPython 界面组件，构建支持多种机器学习、深度学习算法的跨平台桌面端软件。目前 GeoSOS-AI 内置了多元线性回归、逻辑回归、支持向量机、决策树、随机森林、人工神经网络、深度神经网络、主成分分析等机器学习和深度学习算法，后续还将继续融合更多的 GeoAI 方法。

GeoSOS-AI 主界面

2.1.2机器学习及深度学习功能设计

通常来讲，一般的机器学习模型的运行包括以下步骤：

（1）问题定义：明确问题的目标，比如分类、回归或聚类。定义成功标准和关键绩效指标（KPIs）。

（2）数据收集：获取数据源，可以是公开数据集、内部数据、API 抓取或调查问卷。确保数据质量和相关性。

（3）数据预处理：

- 清洗：删除重复值、处理缺失值（填补或删除）。
- 转换：标准化或归一化数据，处理类别变量（如独热编码）。

（4）特征选择与工程：

- 选择：使用统计方法或模型评估特征的重要性，去掉冗余特征。
- 构造：创建新特征，可能包括组合现有特征或提取时间、文本等特征。

（5）模型选择：根据问题性质和数据特征选择合适的机器学习算法，如决策树、支持向量机、神经网络等。

（6）模型训练：将处理后的数据输入选择的模型，利用训练数据进行学习，调整模型内部参数。

（7）模型评估：使用交叉验证、混淆矩阵、ROC 曲线等工具评估模型性能，确保模型具备泛化能力。

（8）模型优化：通过网格搜索、随机搜索等方法优化超参数，改进模型表现。

（9）模型部署：将经过验证的模型部署到生产环境，通常包括构建 API 或集成到现有系统中。

（10）监控与维护：持续监控模型性能，定期更新数据和重新训练模型，以应对数据漂移或新趋势。

对于深度学习模型来说，步骤有所不同：

（1）问题定义：

- 明确要解决的任务，例如图像分类、自然语言处理或生成模型等。
- 确定成功标准，如准确率、F1 分数、损失值等。

（2）数据收集：

- 收集大量的相关数据，深度学习模型通常需要大量样本。
- 数据来源可以是公开数据集、网络爬虫、传感器数据等。

（3）数据预处理：

- 清洗：处理缺失值、重复数据和异常值。
- 转换：根据模型要求对数据进行标准化、归一化或正则化。
- 增强：在图像任务中，使用数据增强（如旋转、翻转）增加数据多样性。

（4）特征选择与工程：

- 深度学习模型通常能够自动学习特征，但在某些情况下，可以进行手动特征选择。
- 对于文本数据，可以使用词嵌入（如 Word2Vec、GloVe）将文本转化为向量。

（5）模型选择：

- 根据任务类型选择合适的深度学习架构，例如卷积神经网络（CNN）用于图像处理，递归神经网络（RNN）或变压器（Transformer）用于序列数据。

（6）模型构建：

- 使用深度学习框架（如 TensorFlow、PyTorch）构建模型架构。
- 定义层数、每层的节点数、激活函数等。

（7）模型训练：

- 将训练数据输入模型，使用反向传播算法优化参数，通常使用梯度下降法。
- 设定批次大小、学习率和训练轮数（epoch）。

（8）模型评估：

- 使用验证集评估模型性能，调整超参数以避免过拟合。
- 评估指标包括准确率、召回率、AUC 等。

（9）模型优化：

- 应用正则化（如 L2 正则化、Dropout）以防止过拟合。

- 通过调整学习率、使用学习率衰减或更复杂的优化算法（如 Adam）来优化训练过程。

（10）模型部署：

- 将训练好的模型导出，并部署到生产环境，通常需要构建 API 或集成到应用中。
- 确保部署环境具备适当的计算资源。

（11）监控与维护：

- 持续监控模型在实际应用中的表现，收集反馈。
- 定期更新模型以适应新的数据，可能需要重新训练或微调。

因此，GeoSOS-AI 软件利用 Python 机器学习库 scikit-learn、Keras 和 TensorFlow 等软件包内置的机器学习、深度学习能力，根据用户提供的数据文件，由用户选择相关模型（如人工神经网络、决策树、逻辑回归、随机森林、深度神经网络等）实现对应问题模型的构建/训练和预测等功能。

2.2 软件数据准备

2.2.1 数据准备

根据所研究的具体问题，收集相关数据，如涉及的各类特征（Feature），即自变量或影响因素；以及目标（Target），即因变量或影响结果。

2.2.2 数据处理

获取的数据需要满足一定的要求，才能符合 GeoSOS-AI 软件运行的要求，但其处理步骤较为简单和方便。

（1）数据格式要求

软件目前支持的数据格式包括 CSV、XLS、XLSX 等格式。

（2）数据处理要求

（1）数据的第一行为各类特征（Feature）数据列和目标（Target）列的列名。

（2）各类特征（Feature）所属的数据列需要放在目标（Target）列的前面，

目标列为最后一列。


示例的数据如下图，第一行为列名，前面的数据列为特征（Feature），最后一列为目标（Target）。

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
2	0.00632	18	2.31	0	0.538	6.575	65.2	4.09	1	296	15.3	396.9	4.98	24
3	0.02731	0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	396.9	9.14	21.6
4	0.02729	0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7
5	0.03237	0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4
6	0.06905	0	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	396.9	5.33	36.2


如果第一行为数据，则第一行被认为是列名，不参与计算。

3 软件运行设置

3.1 软件运行条件

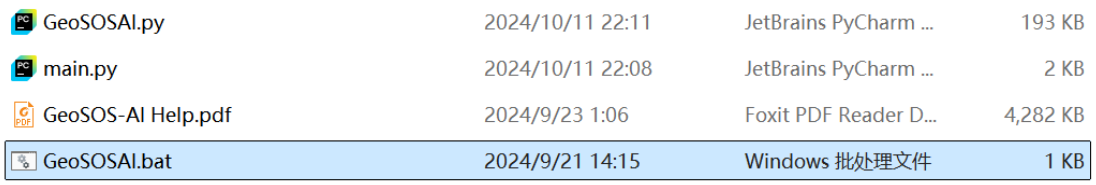
本软件作为基于 Python 构建的软件，可以跨操作系统平台运行，但需要在 Python 3.11.9 及以上版本的 Python 环境中运行。

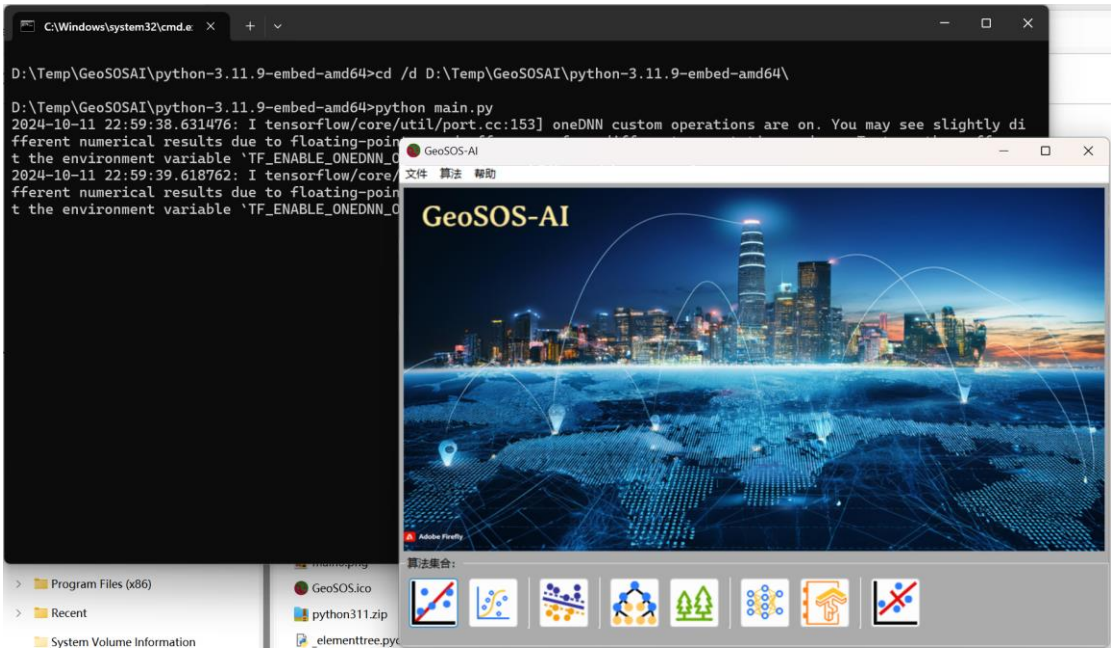
软件提供了包括 Python 运行环境的运行包，操作系统未安装 Python 环境即可运行。如安装了 Python 3.11.9 及以上版本的 Python 环境，则可以执行所提供的 Python 源代码文件运行。

3.2 软件执行方式

3. 2. 1 运行包执行方式

解压“GeoSOSAI.zip”文件，在解压后的“GeoSOSAI\python-3.11.9-embed-amd64”路径中，点击“GeoSOSAI.bat”文件，即可执行 GeoSOS-AI 软件。





3.2.2 源代码文件执行方式

3.2.2.1 使用 PyCharm 执行

可以使用 PyCharm 软件打开解压后的源代码文件 main.py 和 GeoSOSAI.py，运行 main.py 文件，即可执行 GeoSOS-AI 软件。

ontrol

main.py | GeoSOSAI.py

```
1  ################################################################################
2  ## Copyright [2024] [GeoSOS and Prof. Xia Li]
3  ##
4  ## Licensed under the Apache License, Version 2.0 (the "License");
5  ## you may not use this file except in compliance with the License.
6  ## You may obtain a copy of the License at
7  ##
8  ##     http://www.apache.org/licenses/LICENSE-2.0
9  ##
10 ## Unless required by applicable law or agreed to in writing, software
11 ## distributed under the License is distributed on an "AS IS" BASIS,
12 ## WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
13 ## See the License for the specific language governing permissions and
14 ## limitations under the License.
15 ################################################################################
16
17
18 from GeoSOSAI import MainFrame  # 从noname.py文件中导入Frame类
19 import wx  # 导入wx的模块
20
21 app = wx.App()  # 运行wx.App()方法
22 frame = MainFrame(None)  # 调用Frame类，并且不指定父类，当前就成为父类
23 icon = wx.Icon( *args: 'geosos.ico', wx.BITMAP_TYPE_ICO)
24 frame.SetIcon(icon)
25 frame.Show()  # 运行展示界面的方法Show()
26 app.MainLoop()  # 进入程序wx.App()循环
```

使用该方式还可以浏览、修改和调试 GeoSOS-AI 软件的源代码。本软件采用 Apache License Version 2.0 软件协议，因此后续的修改及发布请遵循该软件协议进行。

3.2.2.2 使用 Python 命令行执行

如果本机安装了 Python 环境，可以打开命令行工具，在源代码文件所在的目录下，执行“python main.py”命令，即可执行 GeoSOS-AI 软件。

3.3 软件界面

GeoSOS-AI 软件界面主要包括菜单和按钮，菜单具有全部的软件功能，也可以通过按钮执行所有的机器学习、深度学习算法功能。

目前 GeoSOS-AI 主要包括以下模型：

：多元线性回归

 ：逻辑回归

 ：支持向量机

 ：决策树

 ：随机森林

 ：人工神经网络

 ：深度神经网络

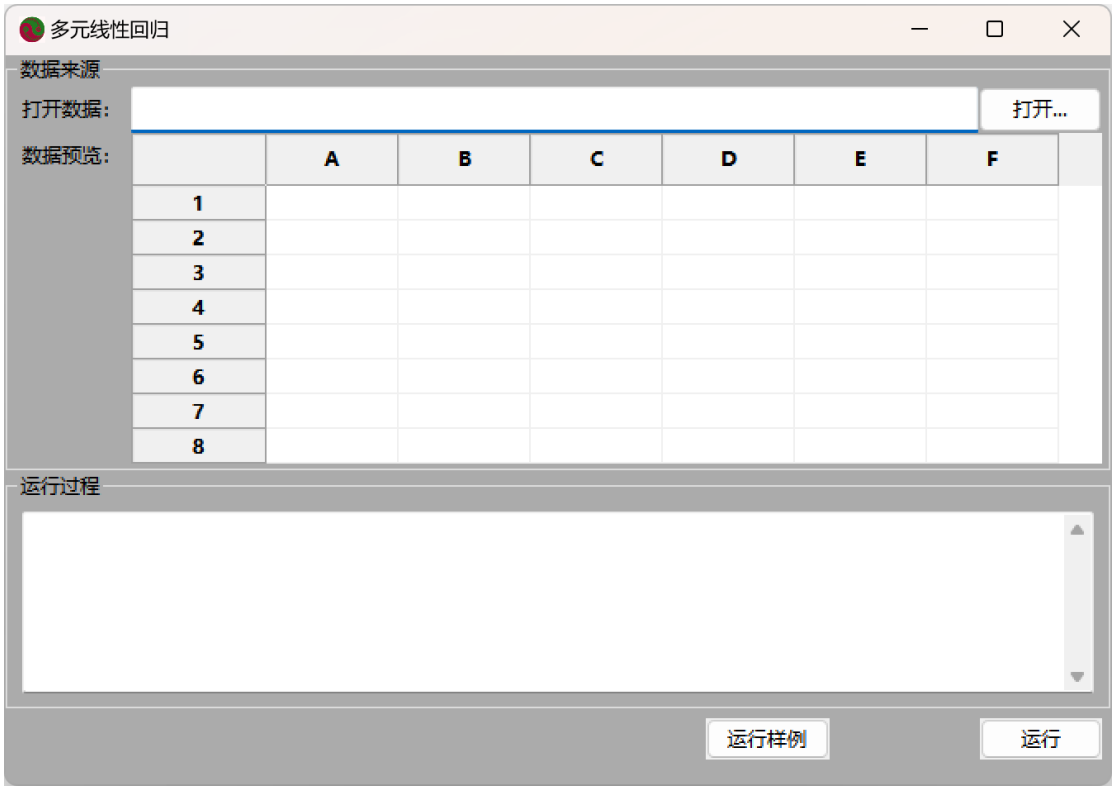
 ：主成分分析

4 系统功能

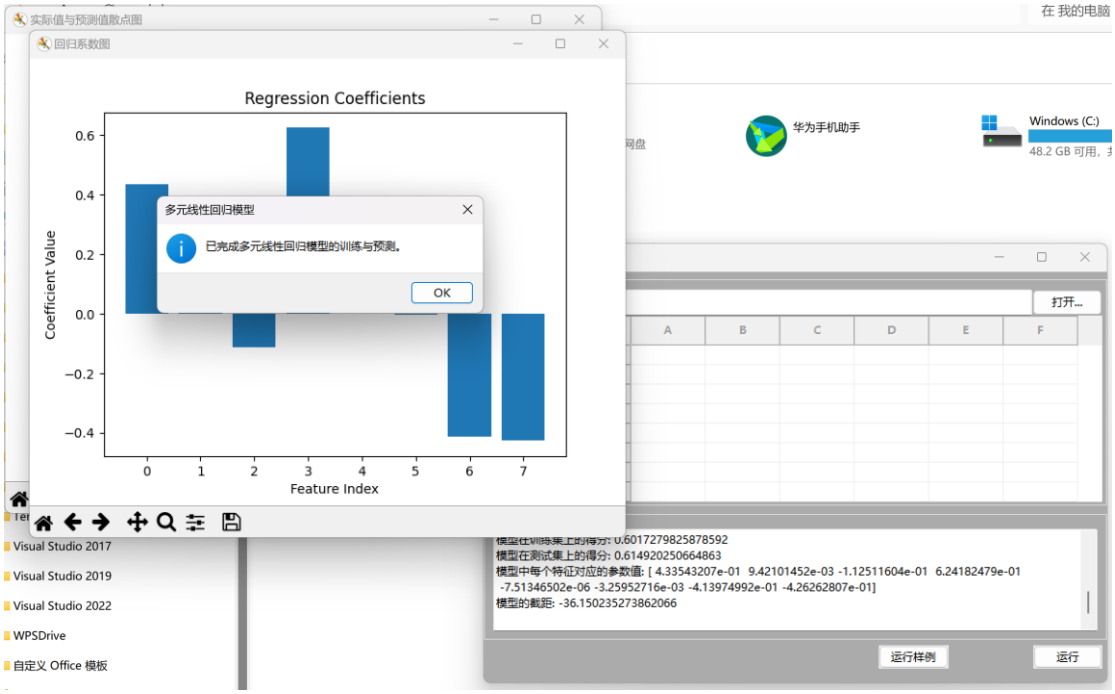
4.1 多元线性回归

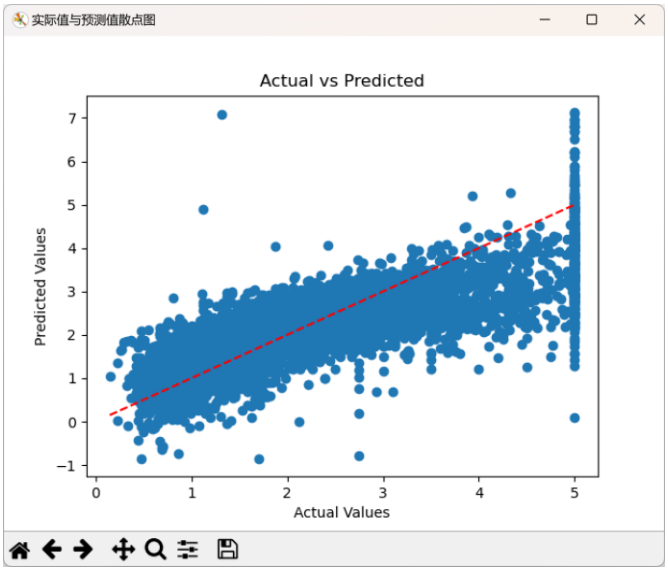
4. 1. 1 运行样例

使用内置的 california housing 样例数据，进行多元线性回归模型计算的示例。点击“算法”菜单中的“多元线性回归”子菜单，或者点击主界面中的  按钮，打开多元线性回归模型界面，点击“运行样例”按钮，将自动进行基于样例数据的模型计算。



运行结果将在多元线性回归窗体的文本框中输出模型训练等过程和结果信息，以及相关图件，并提示已完成模型的计算。相关结果图主要包括实际值与预测值散点图、回归系数图。





实际值与预测值散点图可以用于：

（1）模型拟合效果：可以直观地查看模型的预测结果与实际值之间的关系。如果点大致沿着对角线分布，说明模型拟合得很好；如果点分布较散，说明模型可能没有很好地捕捉到数据的趋势。

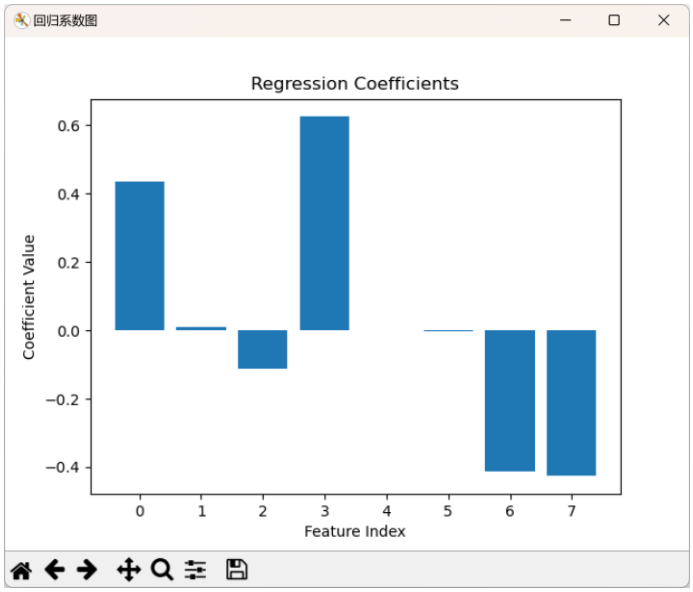
（2）识别异常值：散点图可以帮助识别异常值或离群点。这些点可能会影响模型的表现，了解它们的存在有助于进一步分析。

（3）检测线性关系：对于线性回归模型，散点图可以帮助判断是否存在线性关系。如果散点呈现某种曲线形状，可能意味着数据不符合线性假设，可能需要

考虑非线性模型或数据转换。

（4）模型偏差：如果散点图显示出系统性的偏差，例如预测值总是高于或低于实际值，这可能表明模型存在偏差，需要重新考虑模型的设计或特征选择。

（5）验证假设：通过观察实际值与预测值的散点图，可以验证模型假设是否成立，例如线性关系、同方差性等。



回归系数图的作用主要包括：

（1）特征重要性：回归系数反映了每个特征对目标变量的影响程度，正值表示正向影响，负值表示负向影响。通过可视化这些系数，可以快速识别最重要的特征。

（2）比较特征：回归系数图使得不同特征之间的影响力易于比较，帮助分析各特征在模型中的相对重要性。

（3）解释模型：可视化回归系数可以为模型提供直观的解释，帮助理解模型如何基于输入特征做出预测。

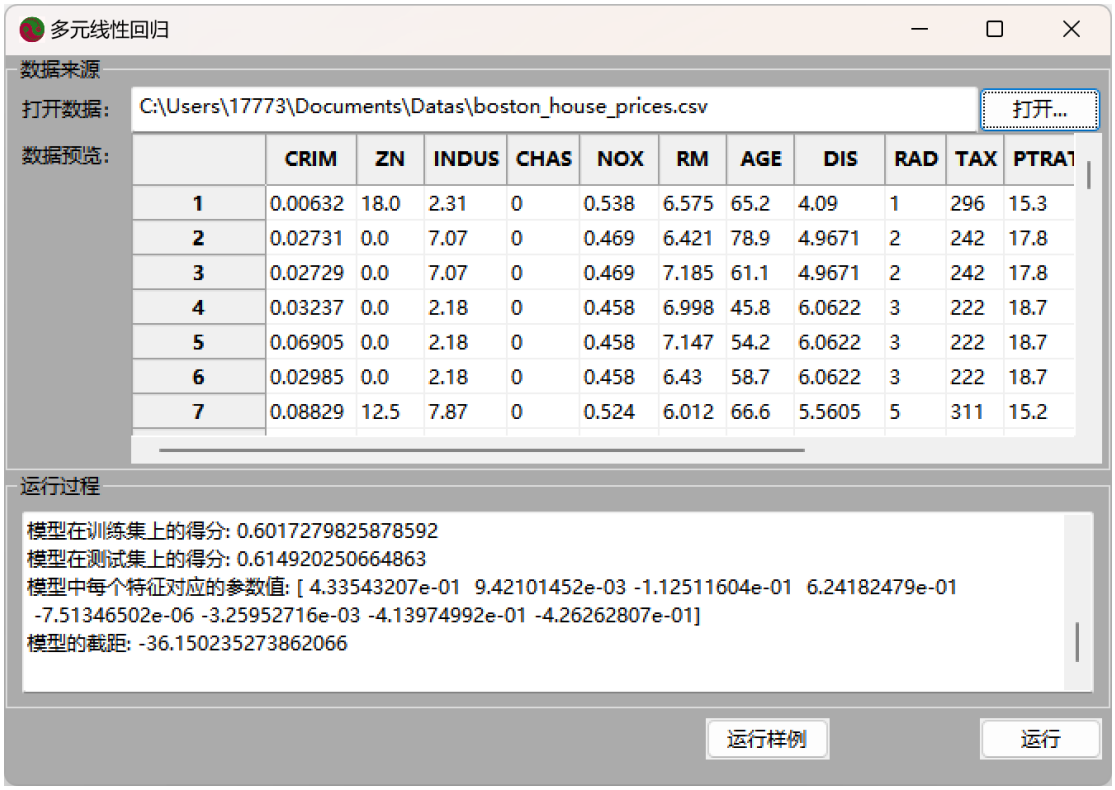
（4）识别多重共线性：如果某些特征的系数接近于零或符号不一致，可能指示多重共线性问题，进一步分析可帮助改进模型。

（5）模型调整：通过观察回归系数，可以判断是否需要对特征进行选择、变换或添加交互项，以提升模型的表现。

文本框中输出的信息可以进行复制，图表窗口中的图片可以进行浏览、保存等操作。

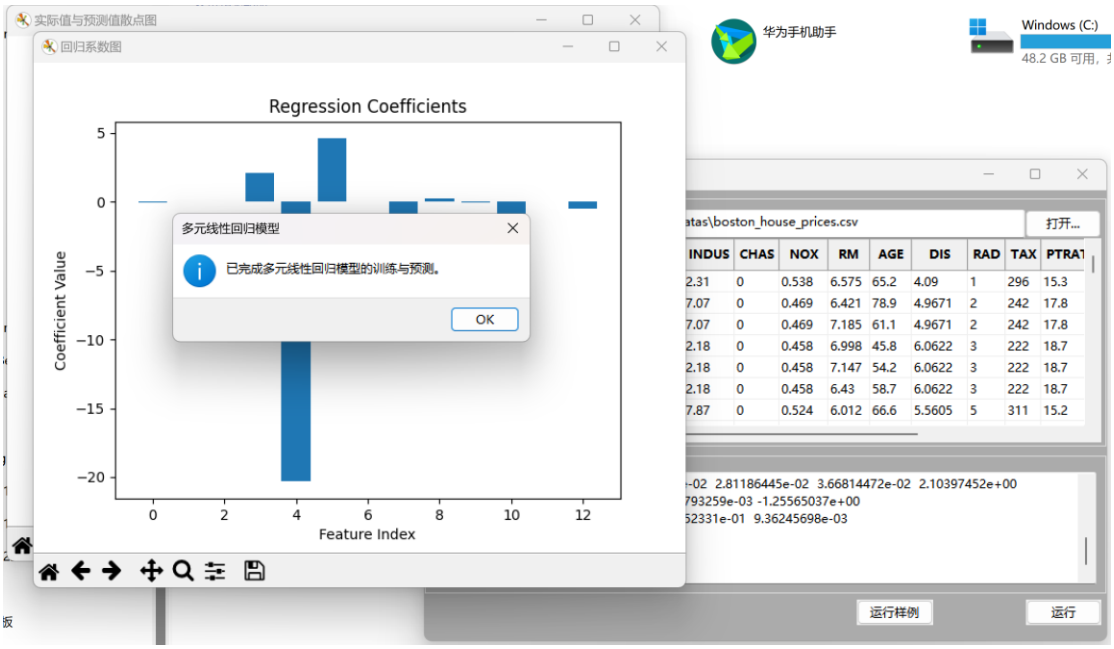
4.1.2运行自定义数据

可以使用用户自定义的 csv、xls、xlsx 数据，进行多元线性回归模型的计算。点击“算法”菜单中的“多元线性回归”子菜单，或者点击主界面中的  按钮，打开多元线性回归模型界面，点击“打开”按钮，选择数据文件，数据文件将被装载到表格中，再点击“运行”按钮，将自动进行自定义数据的模型计算。



	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRAT
1	0.00632	18.0	2.31	0	0.538	6.575	65.2	4.09	1	296	15.3
2	0.02731	0.0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8
3	0.02729	0.0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8
4	0.03237	0.0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7
5	0.06905	0.0	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7
6	0.02985	0.0	2.18	0	0.458	6.43	58.7	6.0622	3	222	18.7
7	0.08829	12.5	7.87	0	0.524	6.012	66.6	5.5605	5	311	15.2

运行结果将在多元线性回归窗体的文本框中输出模型训练等过程和结果信息，以及相关图件，并提示已完成模型的计算。结果及图件输出情况与样例类似。

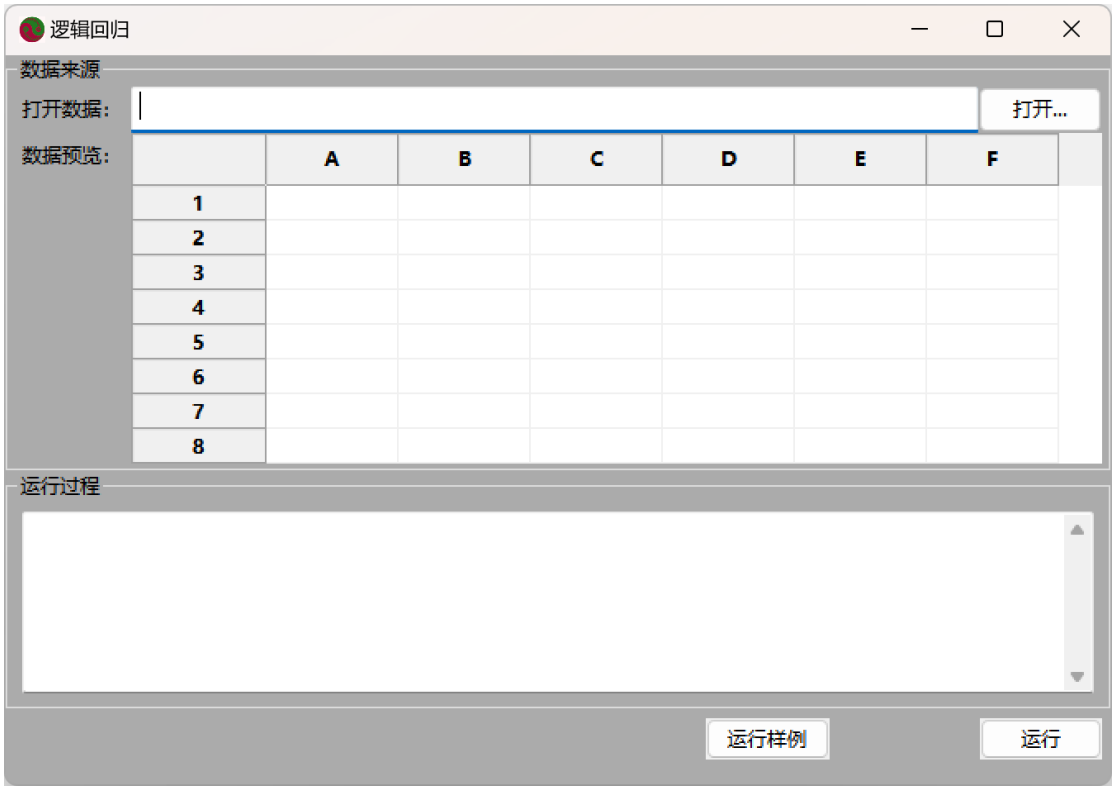


软件也提供了可以进行练习的“boston_house_prices”数据，能够支持对模型进一步的了解，数据介绍可以参考：https://www.kaggle.com/code/prasadperera/the-boston-housing-dataset。

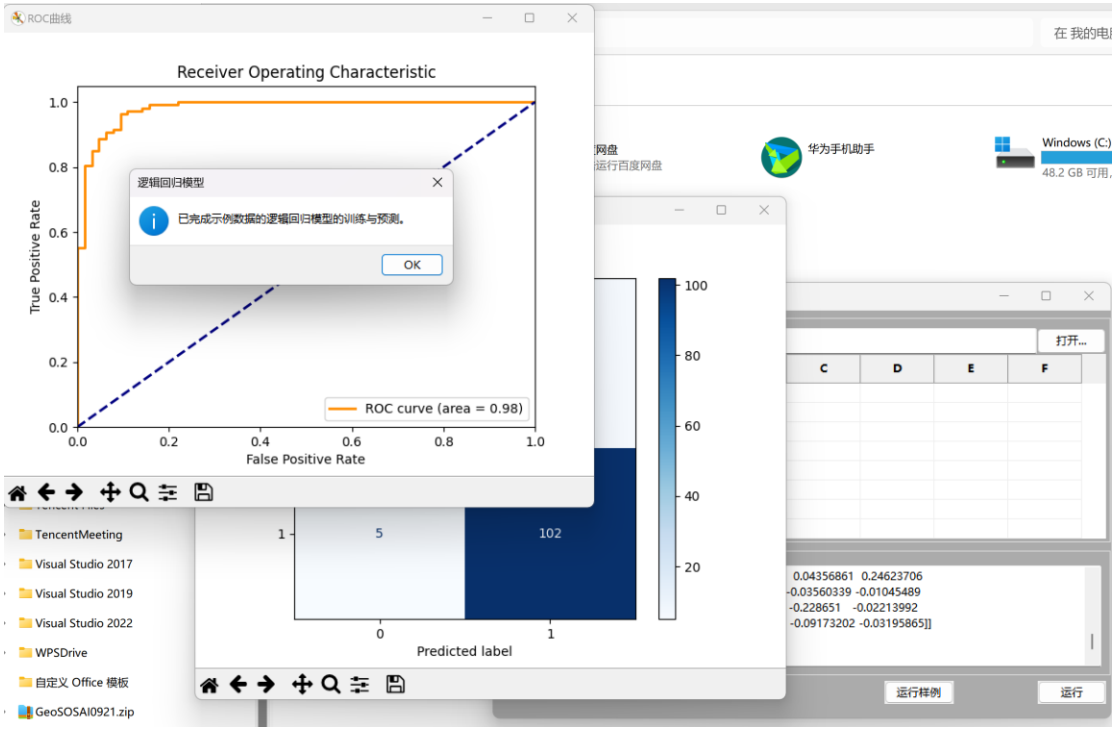
4.2 逻辑回归

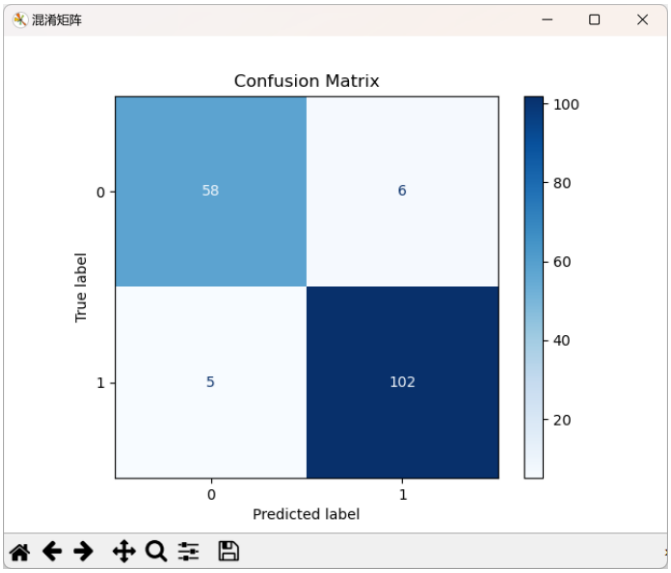
4. 2. 1 运行样例

使用内置的 breast_cancer 样例数据，进行逻辑回归模型计算的示例。点击“算法”菜单中的“逻辑回归”子菜单，或者点击主界面中的 按钮，打开逻辑回归模型界面，点击“运行样例”按钮，将自动进行基于样例数据的模型计算。



运行结果将在逻辑回归窗体的文本框中输出模型训练等过程和结果信息，以及相关图件，并提示已完成模型的计算。相关图件主要包括混淆矩阵图及 ROC 曲线图。





混淆矩阵图的作用主要包括：

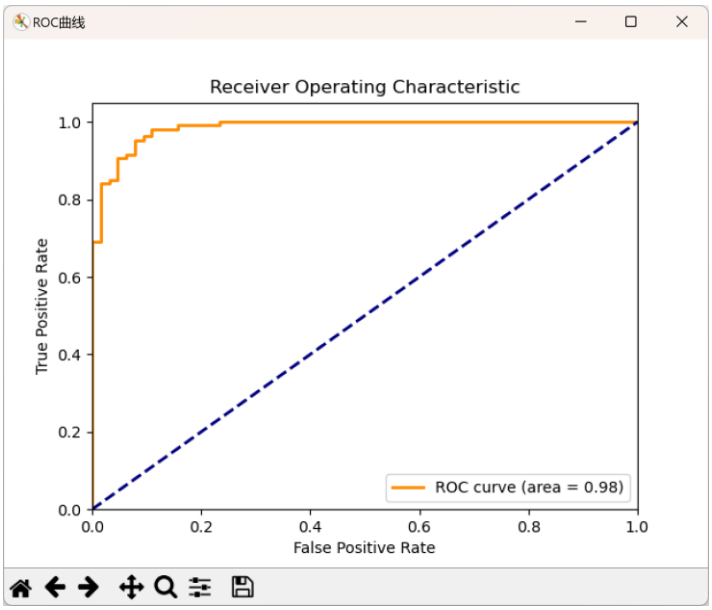
（1）性能评估：混淆矩阵显示了模型在分类任务中的表现，包括真正例（TP）、假正例（FP）、真负例（TN）和假负例（FN）的数量。这有助于计算准确率、精确率、召回率和 F1 分数等评估指标。

（2）错误分析：通过查看混淆矩阵，可以识别模型在哪些类别上表现不佳，帮助分析误分类的原因，进而改进模型或特征选择。

（3）类别不平衡：混淆矩阵可以揭示类别不平衡问题。如果某一类别的 TP 或 TN 数量远低于其他类别，这可能影响模型的性能。

（4）调整阈值：混淆矩阵可以帮助确定最佳的分类阈值，通过观察不同阈值下的混淆矩阵变化，优化模型的分类决策。

（5）可视化：将混淆矩阵以热图形式可视化，可以更直观地了解模型的分类效果，便于交流和报告。



逻辑回归的 ROC（接收器操作特征）曲线图有几个重要作用：

（1）性能评估：ROC 曲线展示了不同分类阈值下的真正率（TPR）与假正率（FPR）的关系，帮助评估模型在各个阈值下的分类性能。

（2）选择最佳阈值：通过观察 ROC 曲线，可以选择一个合适的阈值，使得真正率和假正率达到最佳平衡，优化模型的分类效果。

（3）比较模型：多个模型的 ROC 曲线可以在同一图中展示，便于直观比较不同模型的分类能力，通常曲线越接近左上角，模型性能越好。

（4）计算 AUC：ROC 曲线下的面积（AUC）提供了一个量化的性能指标，AUC 值越接近 1，表示模型的分类能力越强，值为 0.5 则表示模型性能与随机猜测无异。

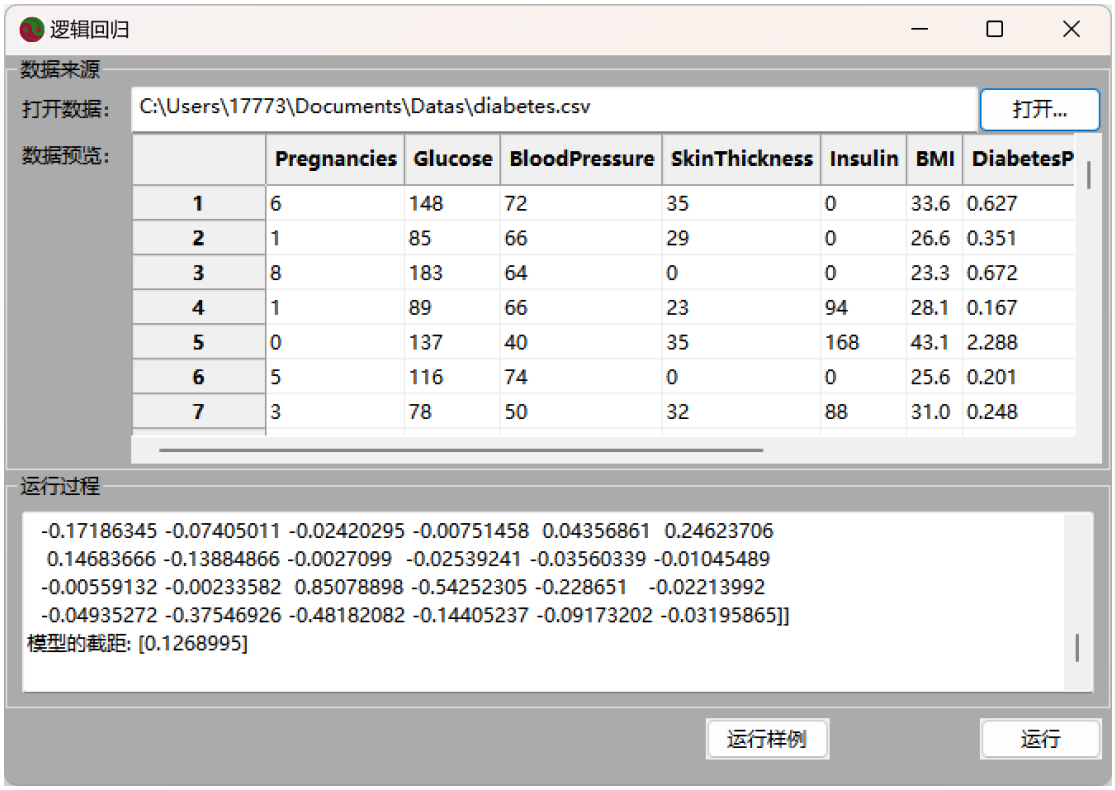
（5）处理不平衡数据：ROC 曲线对类别不平衡不敏感，能够有效反映模型在不同情况下的表现，因此在处理不平衡分类问题时特别有用。

文本框中输出的信息可以进行复制，图表窗口中的图片可以进行浏览、保存等操作。

4.2.2 运行自定义数据

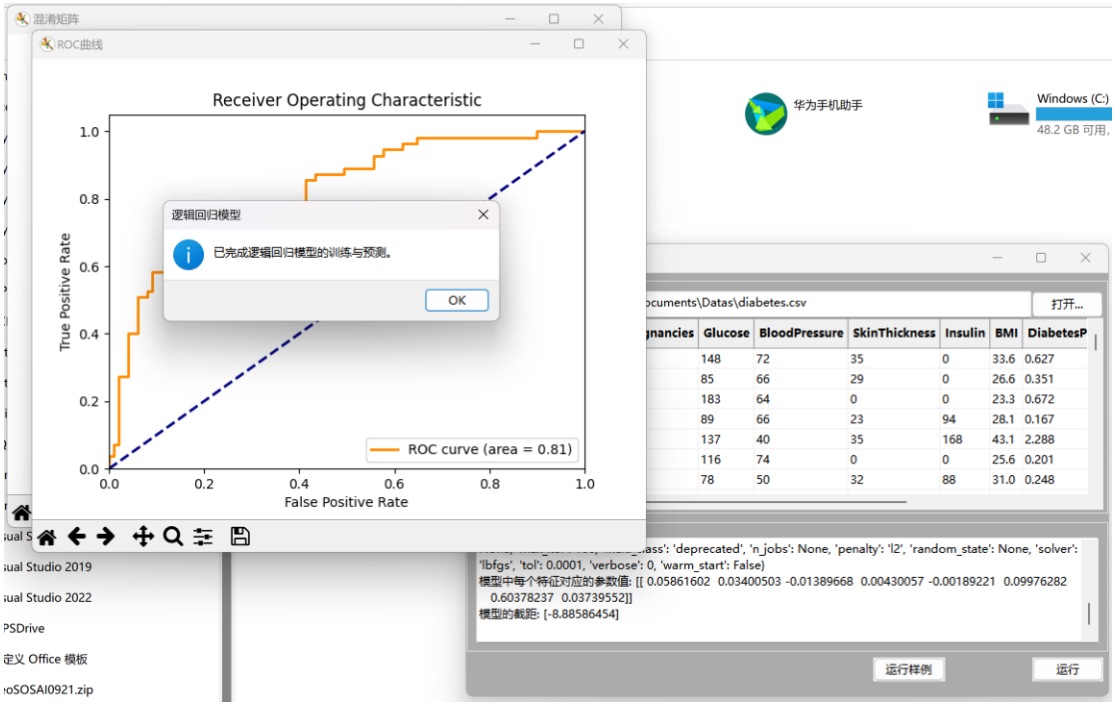
可以使用用户自定义的 csv、xls、xlsx 数据，进行逻辑回归模型的计算。点击“算法”菜单中的“逻辑回归”子菜单，或者点击主界面中的 按钮，打开逻

辑回归模型界面，点击“打开”按钮，选择数据文件，数据文件将被装载到表格中，再点击“运行”按钮，将自动进行自定义数据的模型计算。



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesP
1	6	148	72	35	0	33.6	0.627
2	1	85	66	29	0	26.6	0.351
3	8	183	64	0	0	23.3	0.672
4	1	89	66	23	94	28.1	0.167
5	0	137	40	35	168	43.1	2.288
6	5	116	74	0	0	25.6	0.201
7	3	78	50	32	88	31.0	0.248

运行结果将在逻辑回归窗体的文本框中输出模型训练等过程和结果信息，以及相关图件，并提示已完成模型的计算。结果及图件输出情况与样例类似。

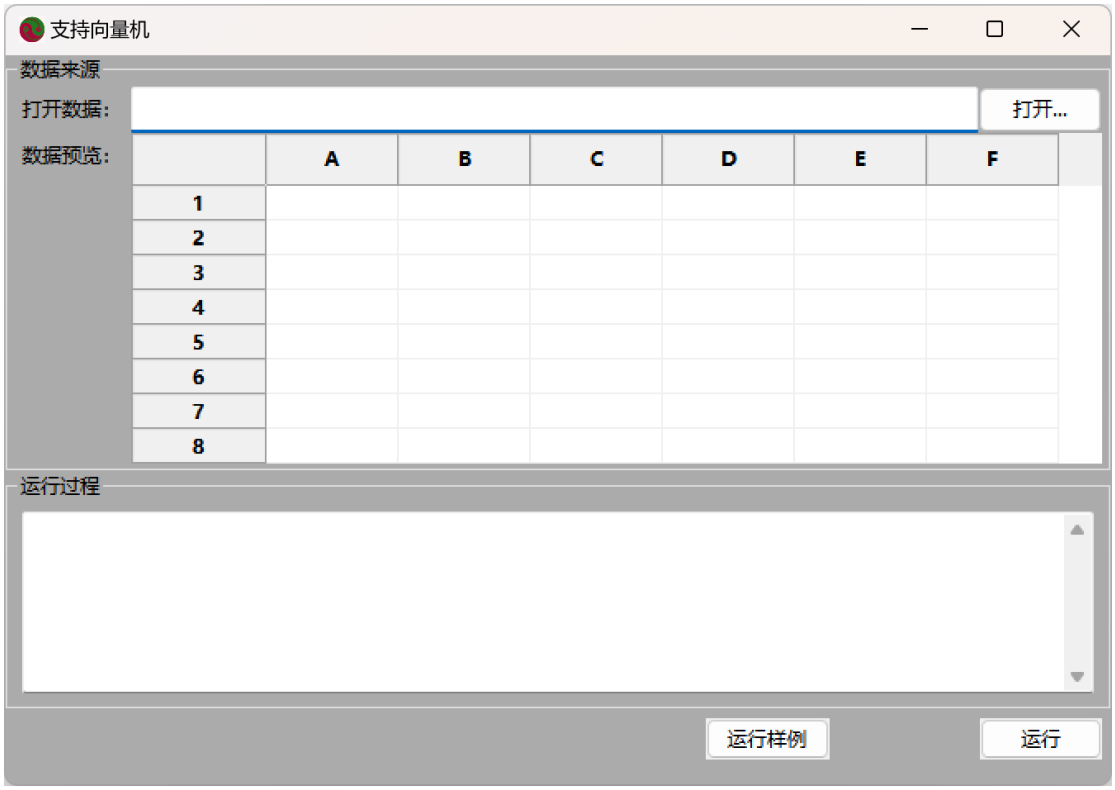


软件也提供了可以进行练习的“diabetes”数据，能够支持对模型进一步的了解，数据介绍可以参考：https://www.kaggle.com/datasets/mathchi/diabetes-data-set。

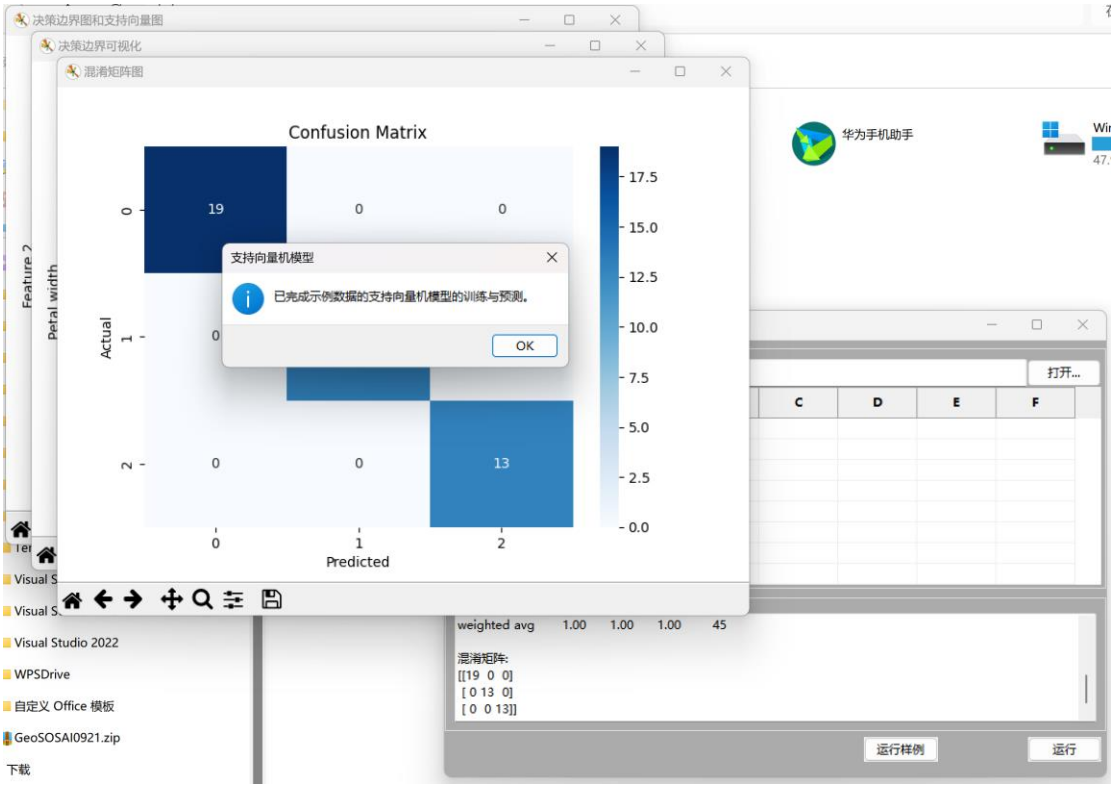
4.3 支持向量机

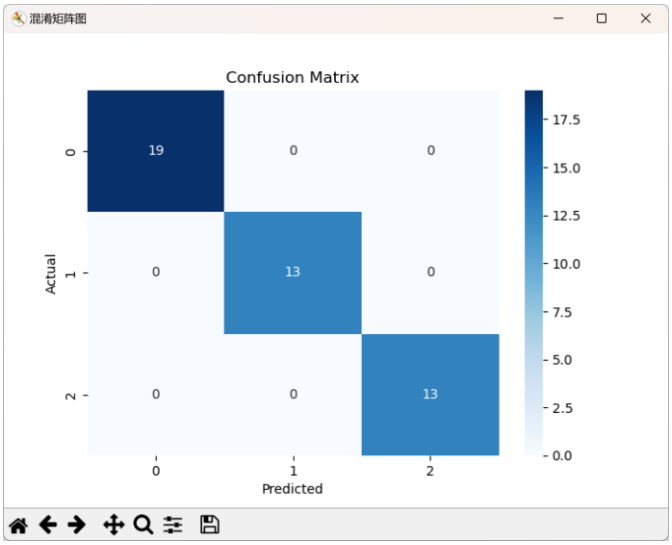
4. 3. 1 运行样例

使用内置的 iris 样例数据，进行支持向量机模型计算的示例。点击“算法”菜单中的“支持向量机”子菜单，或者点击主界面中的按钮，打开支持向量机模型界面，点击“运行样例”按钮，将自动进行基于样例数据的模型计算。



运行结果将在支持向量机窗体的文本框中输出模型训练等过程和结果信息，以及相关图件，并提示已完成模型的计算。图件主要包括混淆矩阵图、决策边界图和支持向量图、决策边界可视化图。





混淆矩阵图的作用主要包括：

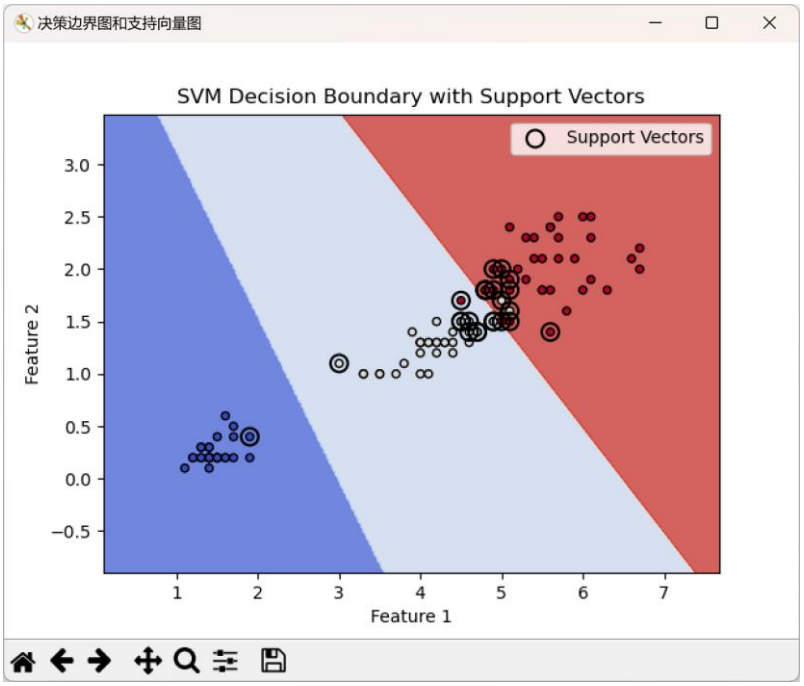
（1）性能评估：混淆矩阵显示了分类模型的性能，包括真正例（TP）、假正例（FP）、真负例（TN）和假负例（FN），从而计算出准确率、精确率、召回率和 F1 分数等指标。

（2）错误分析：通过观察混淆矩阵，可以识别模型在哪些类别上表现不佳，分析误分类的情况，帮助理解模型的局限性。

（3）类别不平衡：混淆矩阵可以揭示类别不平衡问题。如果某个类别的 TP 或 TN 数量明显低于其他类别，可能需要调整模型或数据采样策略。

（4）调整分类阈值：混淆矩阵可以帮助确定最佳的分类阈值，通过观察不同阈值下的混淆矩阵变化，优化模型的分类决策。

（5）可视化效果：将混淆矩阵可视化（如热图）可以使结果更直观，便于向其他人展示模型性能，尤其在报告和演示中非常有用。



决策边界图和支持向量图的用途主要是：

（1）可视化分类决策：决策边界图展示了模型如何将不同类别的数据分开，直观地显示了分类的决策过程，帮助理解模型的工作原理。

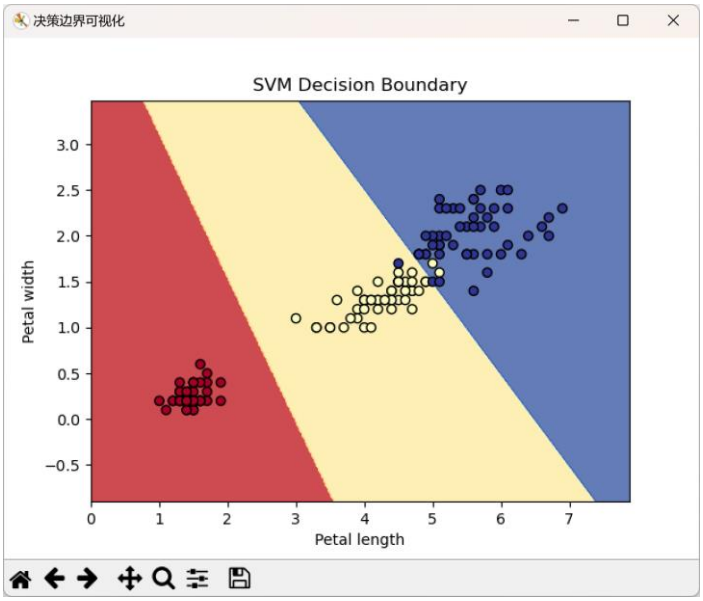
（2）理解支持向量：支持向量图标识了离决策边界最近的样本点，支持向量对模型的建立至关重要，因为它们直接影响决策边界的位置。通过观察支持向量，可以判断哪些数据点对模型的分类效果最为关键。

（3）评估模型的复杂性：通过观察决策边界的形状，可以评估模型的复杂性。如果边界过于复杂，可能会导致过拟合；相反，简单的边界可能表明模型未能捕捉到数据的特征。

（3）选择核函数：在非线性 SVM 中，决策边界图可以帮助验证所选核函数的效果。通过可视化不同核函数下的边界变化，可以判断其对分类效果的影响。

（4）直观展示性能：决策边界图结合样本点的分布，可以直观展示模型在不

同类别上的分类性能，有助于理解模型的优缺点。



决策边界可视化图的主要作用是：

（1）展示分类决策：决策边界图直观地显示了模型如何将不同类别的数据分开，帮助理解分类的逻辑。

（2）理解模型复杂性：通过观察边界的形状和位置，可以判断模型的复杂性，识别是否存在过拟合或欠拟合的问题。

（3）分析支持向量：决策边界图通常会标识出支持向量，这些样本点对决策边界的形成至关重要，有助于分析模型的鲁棒性。

（4）选择合适的核函数：在非线性 SVM 中，不同的核函数会导致不同的决策边界，图形可帮助评估和选择最佳的核函数。

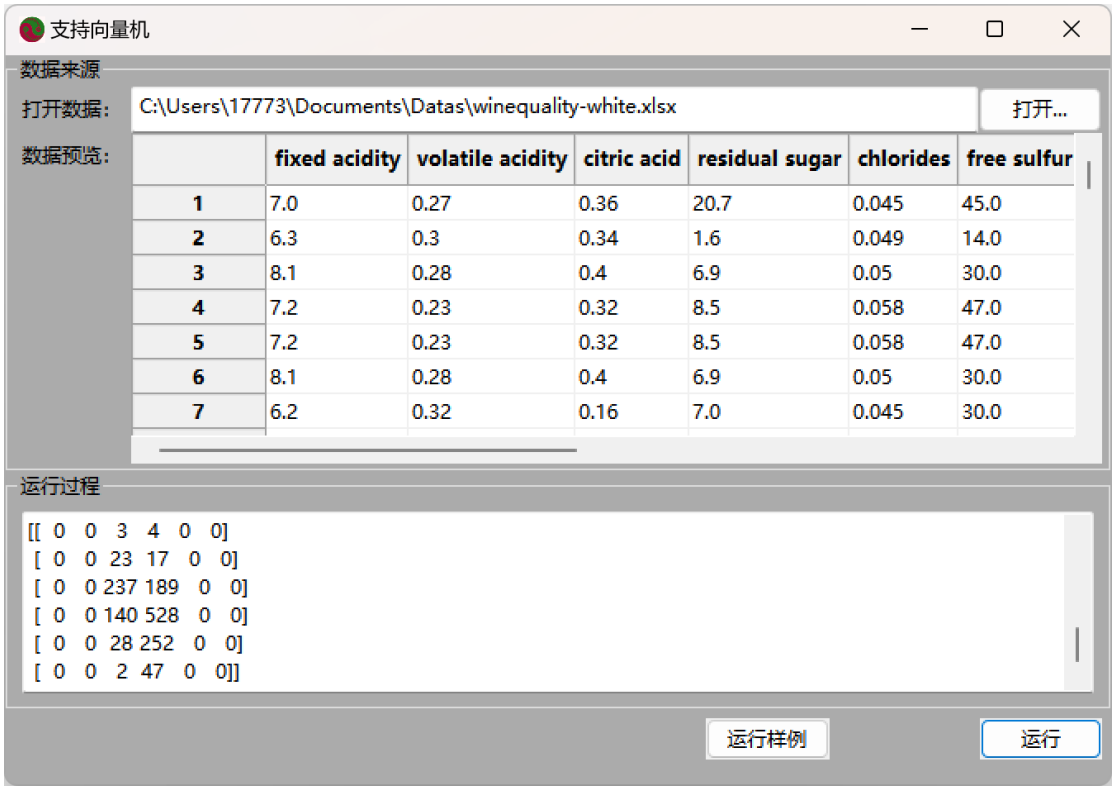
（5）可视化性能评估：结合数据点和决策边界，可以直观展示模型的分类性能，帮助识别潜在的分类错误或不确定区域。

文本框中输出的信息可以进行复制，图表窗口中的图片可以进行浏览、保存等操作。

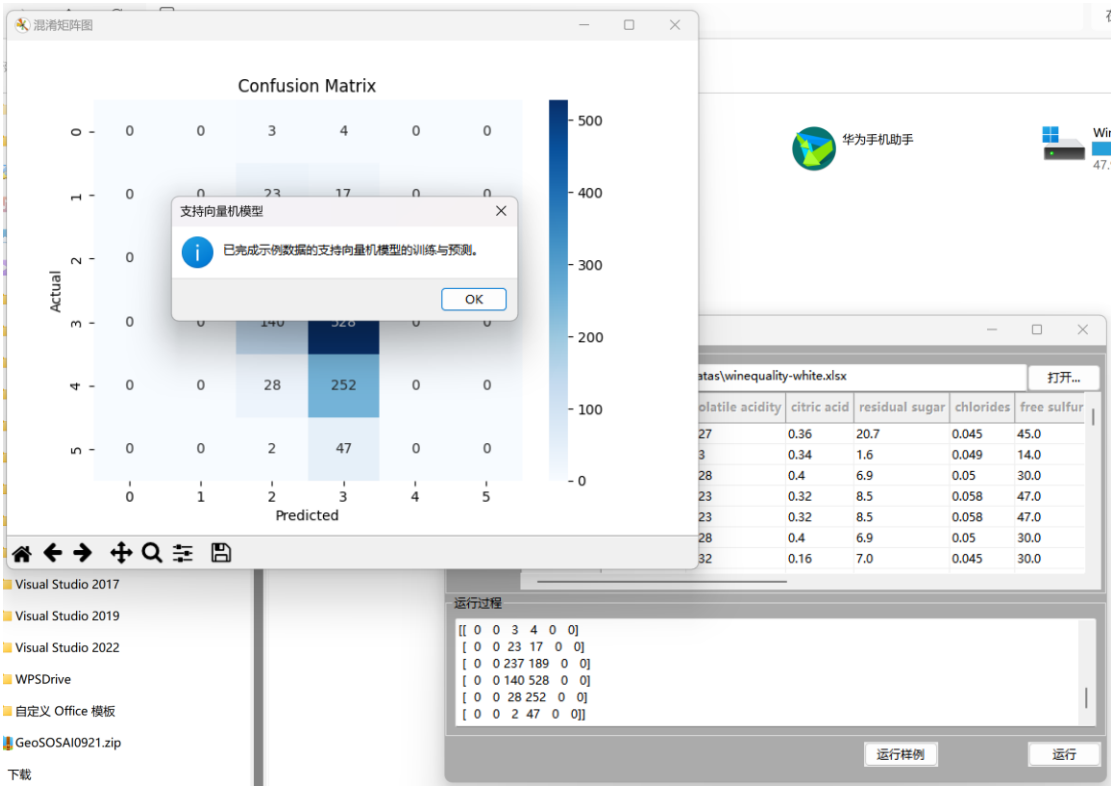
4.3.2 运行自定义数据

可以使用用户自定义的 csv、xls、xlsx 数据，进行支持向量机模型的计算。点击“算法”菜单中的“支持向量机”子菜单，或者点击主界面中的按钮，打

开支持向量机模型界面，点击“打开”按钮，选择数据文件，数据文件将被装载到表格中，再点击“运行”按钮，将自动进行自定义数据的模型计算。



运行结果将在多元线性回归窗体的文本框中输出模型训练等过程和结果信息，以及相关图件，并提示已完成模型的计算。对于非二维平面，无法绘制决策边界图和支持向量图和决策边界可视化图，因此只输出混淆矩阵图。



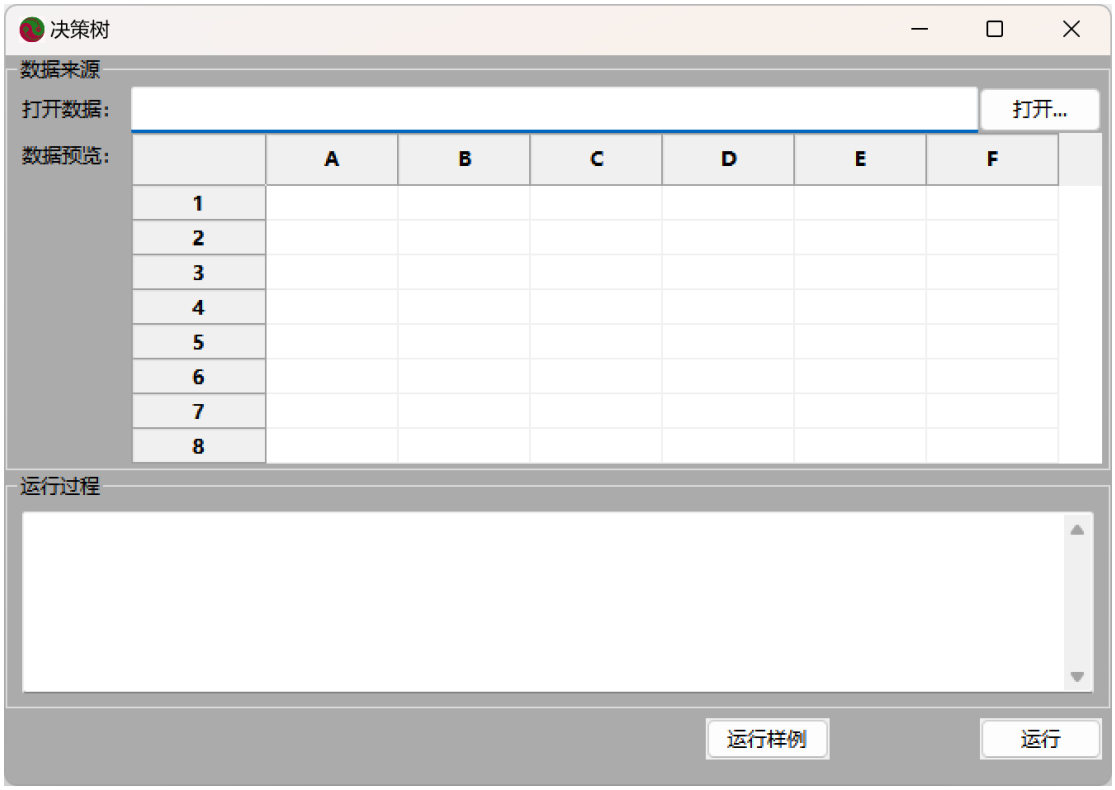
当数据量较大时，SVM 模型的执行时间较长，请耐心等待其执行完毕。

软件也提供了可以进行练习的“winequality-red”数据，能够支持对模型进一步的了解，数据介绍可以参考：https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009。

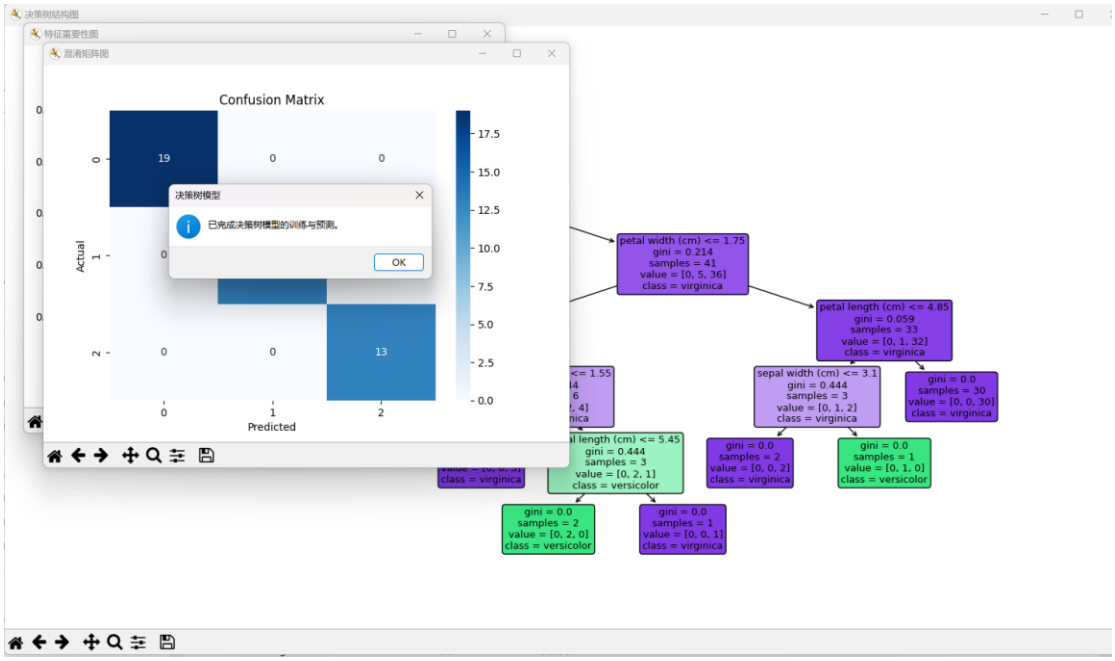
4.4 决策树

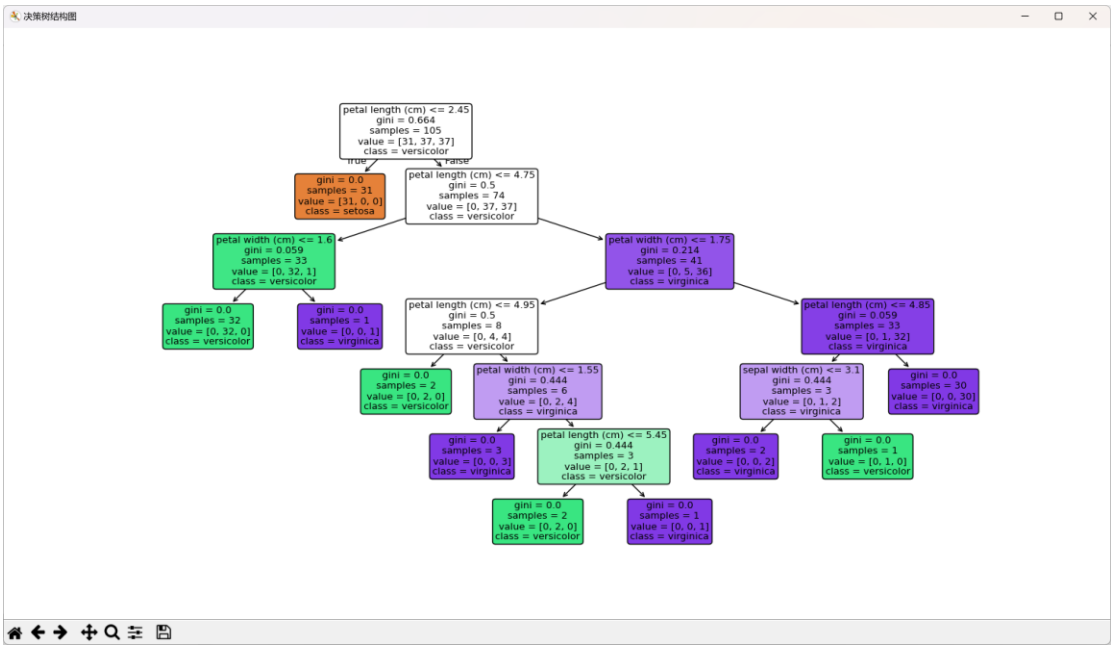
4. 4. 1 运行样例

使用内置的 iris 样例数据，进行决策树模型计算的示例。点击“算法”菜单中的“决策树”子菜单，或者点击主界面中的按钮，打开决策树模型界面，点击“运行样例”按钮，将自动进行基于样例数据的模型计算。



运行结果将在决策树窗体的文本框中输出模型训练等过程和结果信息，以及相关图件，并提示已完成模型的计算。图件包括决策树结构图、混淆矩阵图、特征重要性图。





决策树结构图的作用主要是：

（1）可视化决策过程：决策树图形化地展示了模型的决策过程，显示每个节点的分裂条件，帮助理解模型是如何对数据做出分类或回归的决策。

（2）特征重要性分析：决策树图中节点的分裂依据可以展示每个特征在模型中的重要性程度。靠近树根部的特征通常对预测更为重要，可以快速识别对目标变量影响较大的特征。

（3）解释模型：决策树结构图是少数可以直观解释的机器学习模型之一，图中分裂的路径展示了特征如何组合来做出最终预测，便于与他人沟通模型的工作原理。

（4）诊断过拟合或欠拟合：通过观察树的深度和复杂度，可以识别模型是否过拟合或欠拟合。如果树过深且结构复杂，可能存在过拟合问题；相反，浅层的树可能未能捕捉到数据的特征。

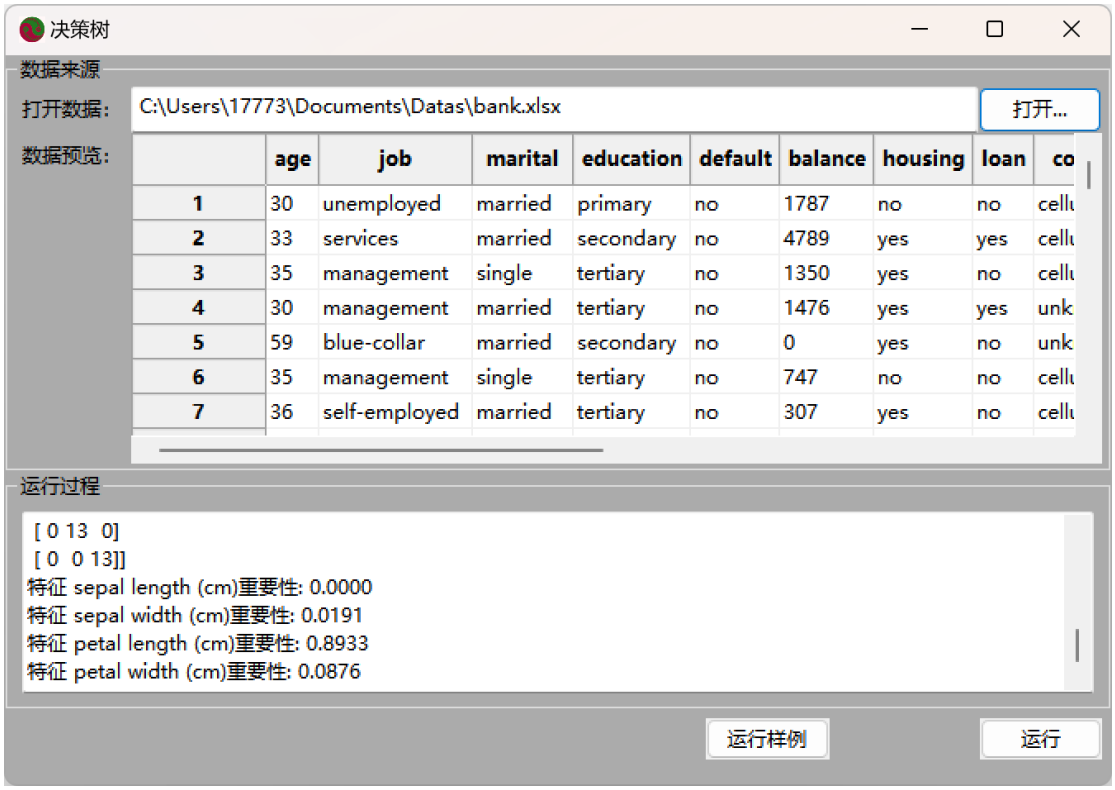
（5）调试与优化模型：决策树图有助于分析和调试模型，识别在树的某些部分中可能存在的分裂问题，可以帮助优化模型，例如修剪树以简化模型，减少过拟合风险。

混淆矩阵图及特征重要性图的作用与前述的其它模型类似。

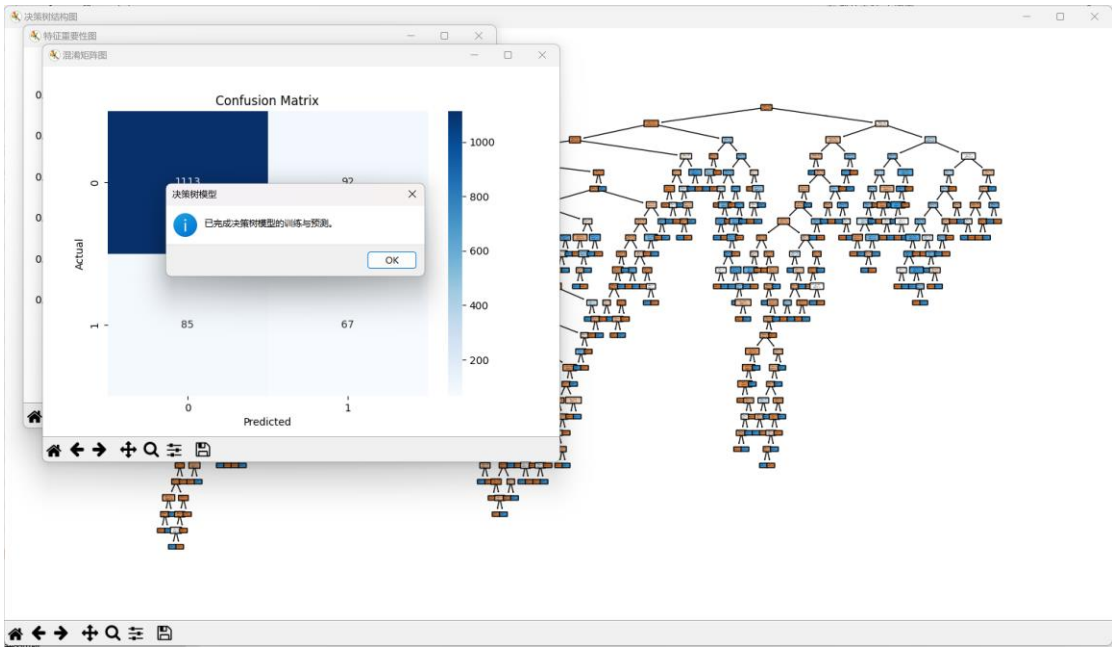
文本框中输出的信息可以进行复制，图表窗口中的图片可以进行浏览、保存等操作。

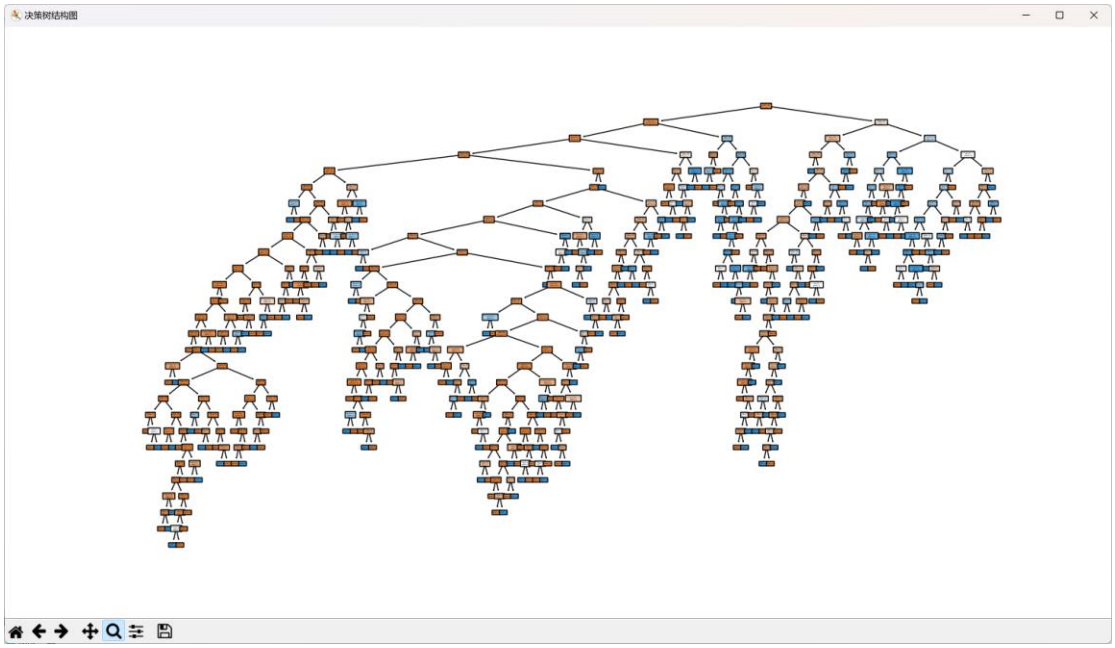
4.4.2运行自定义数据

可以使用用户自定义的 csv、xls、xlsx 数据，进行决策树模型的计算。点击“算法”菜单中的“决策树”子菜单，或者点击主界面中的  按钮，打开决策树模型界面，点击“打开”按钮，选择数据文件，数据文件将被装载到表格中，再点击“运行”按钮，将自动进行自定义数据的模型计算。



运行结果将在多元线性回归窗体的文本框中输出模型训练等过程和结果信息，以及相关图件，并提示已完成模型的计算。结果及图件输出情况与样例类似。





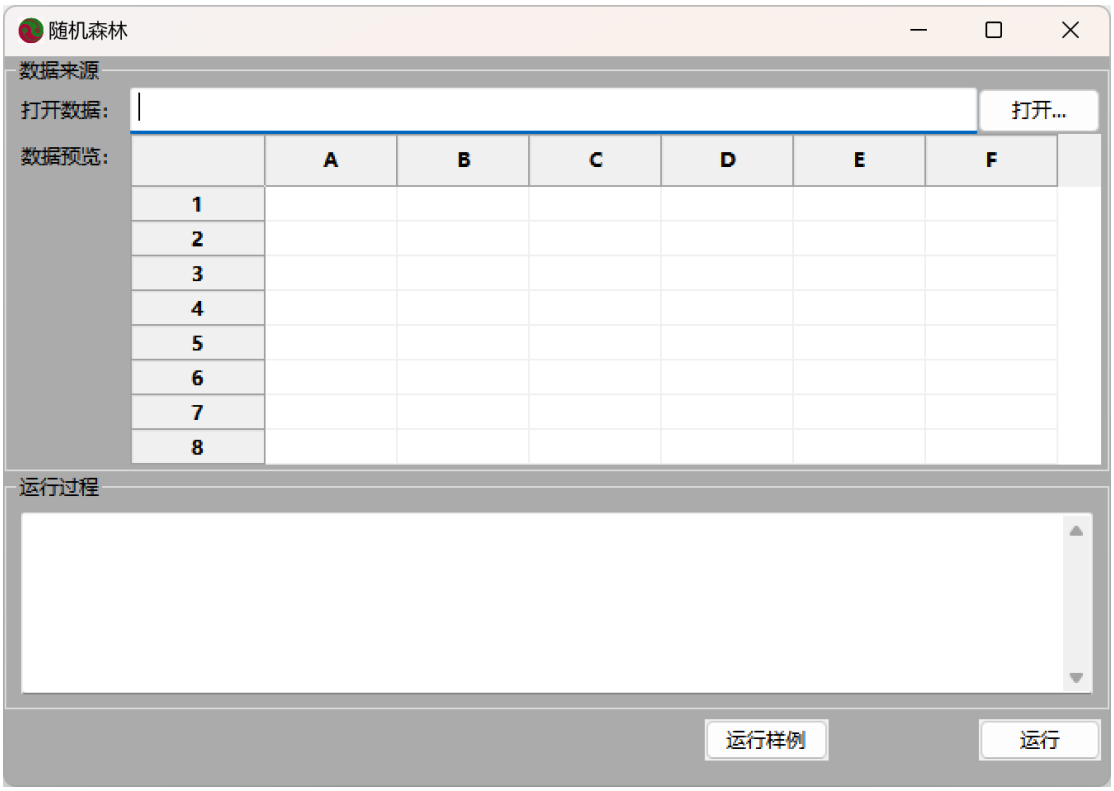
当决策树较为复杂时，绘制决策树结构图的时间较长，请耐心等待其执行完毕。

软件也提供了可以进行练习的“bank”数据，能够支持对模型进一步的了解，数据介绍可以参考：https://archive.ics.uci.edu/dataset/222/bank+marketing。

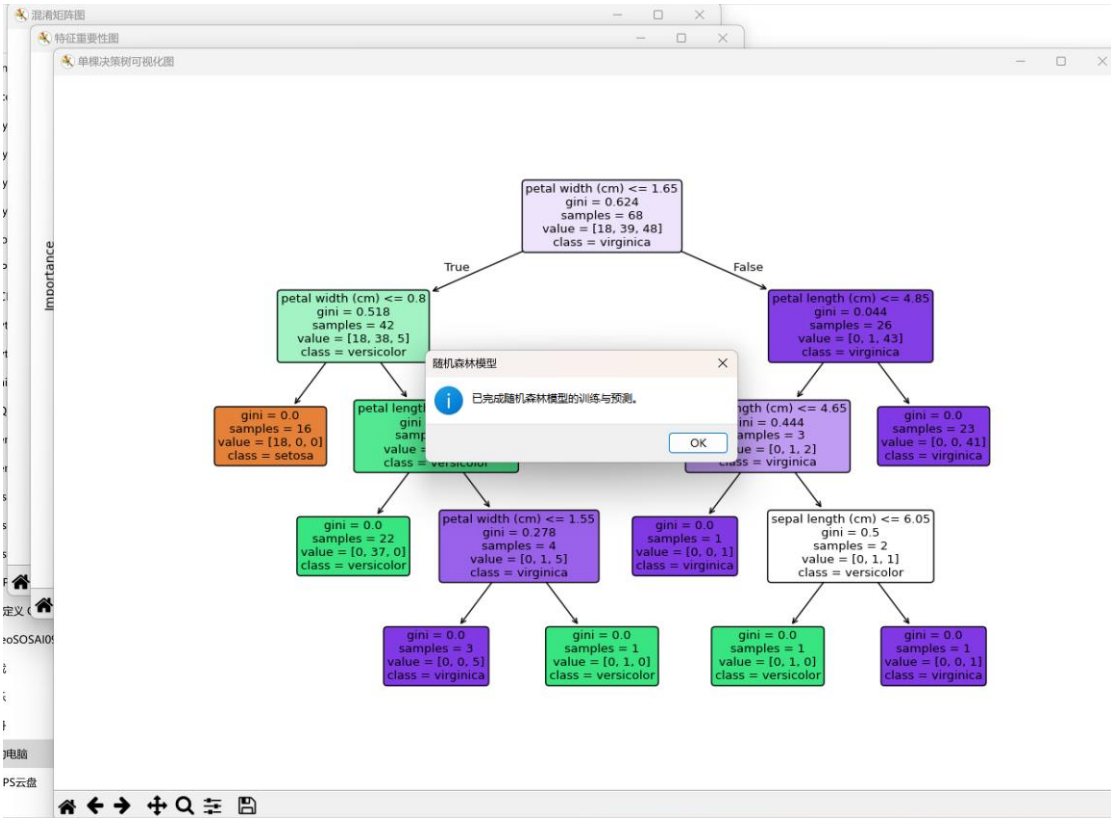
4.5　随机森林

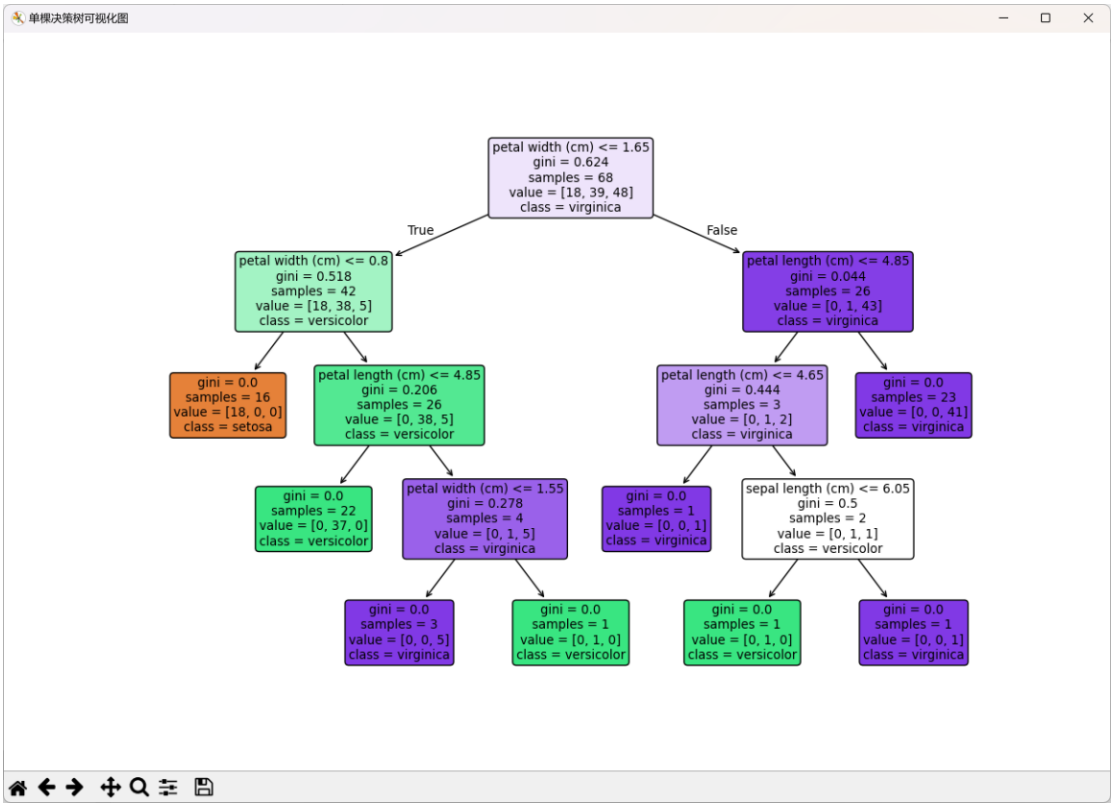
4. 5. 1 运行样例

使用内置的 iris 样例数据，进行随机森林模型计算的示例。点击“算法”菜单中的“随机森林”子菜单，或者点击主界面中的  按钮，打开随机森林模型界面，点击“运行样例”按钮，将自动进行基于样例数据的模型计算。



运行结果将在随机森林窗体的文本框中输出模型训练等过程和结果信息，以及相关图件，并提示已完成模型的计算。图件包括单颗决策树结构图、混淆矩阵图、特征重要性图。





单颗决策树可视化图的作用主要是：

（1）理解模型构建：单颗决策树的可视化可以帮助理解随机森林中每棵树是如何通过特征分裂来进行决策的。这对于理解复杂模型的基本构建块是非常有帮

助的。

（2）特征重要性分析：通过观察单棵决策树的结构，可以识别哪些特征在该树的决策过程中起到了关键作用。这有助于分析特征的重要性并为特征选择提供依据。

（3）可解释性：尽管随机森林是一个集成模型，单颗决策树相对容易解释。通过可视化，用户可以清楚地看到模型是如何基于输入特征做出预测的，这对于与非技术人员沟通模型的工作原理尤其重要。

（4）错误分析：可视化单颗决策树有助于识别模型的潜在错误。例如，可以查看某些分裂是否导致了对特定类别的误分类，从而为改进模型提供线索。

（5）调试和优化：通过观察单颗决策树的结构，可以诊断模型的复杂性，识别可能导致过拟合的特征或分裂，进而进行调整和优化。

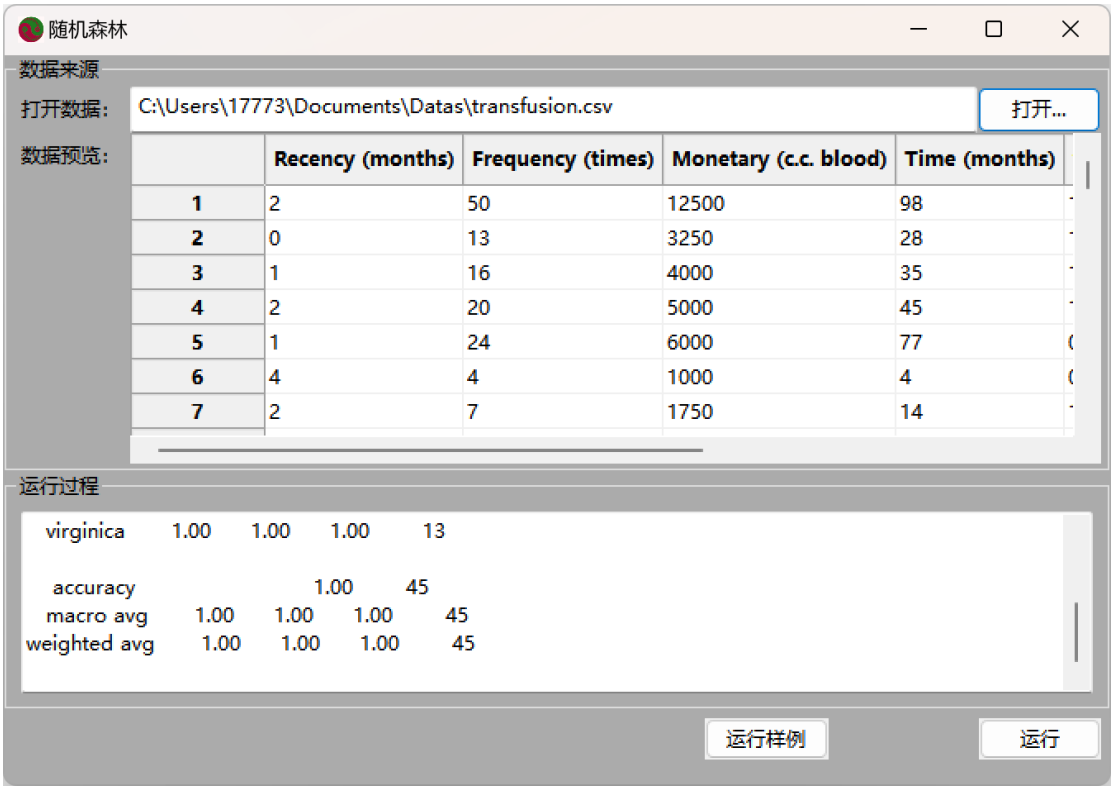
（6）教学工具：在教学和培训中，单颗决策树的可视化是一个很好的示例，可以帮助理解决策树算法的工作原理。

混淆矩阵图及特征重要性图的作用与前述的其它模型类似。

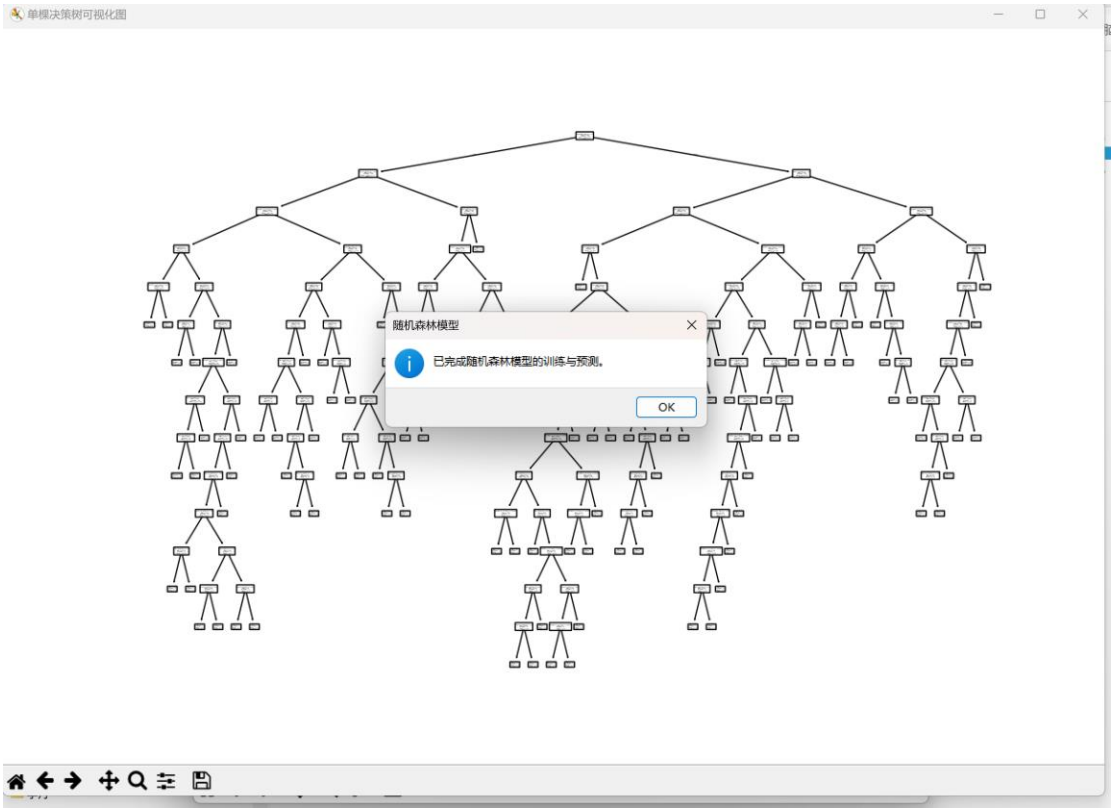

文本框中输出的信息可以进行复制，图表窗口中的图片可以进行浏览、保存等操作。

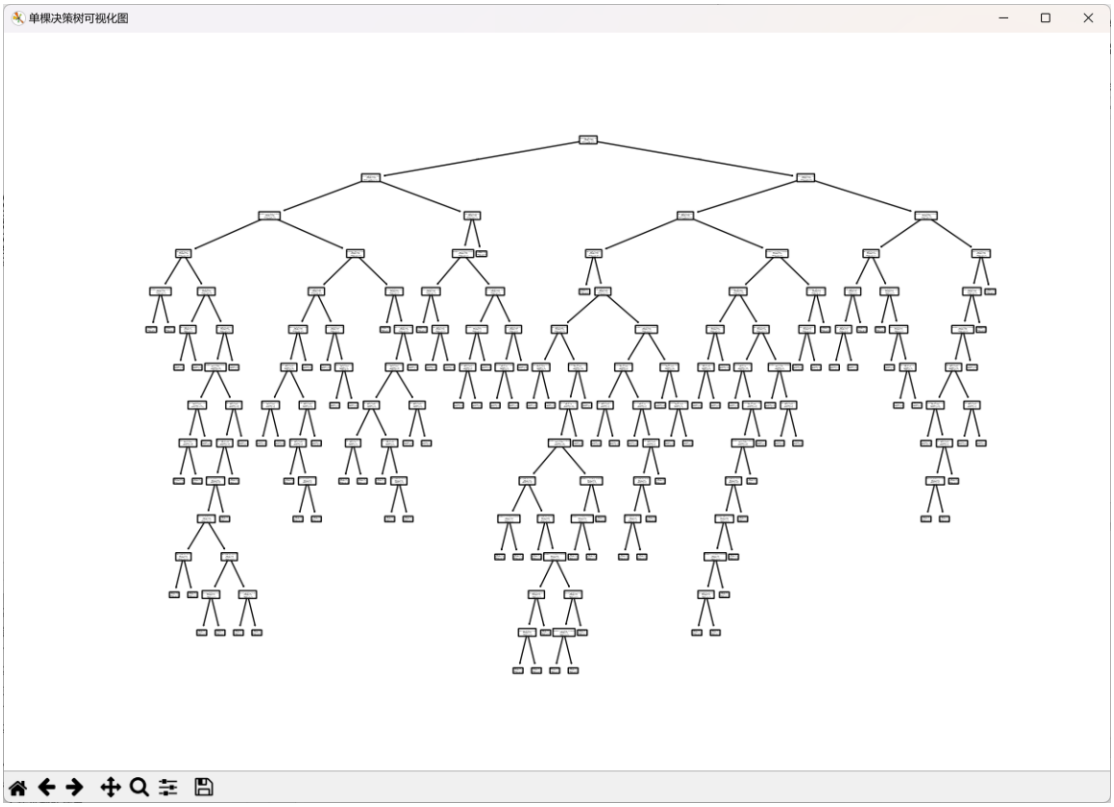
4.5.2 运行自定义数据

可以使用用户自定义的 csv、xls、xlsx 数据，进行随机森林模型的计算。点击“算法”菜单中的“随机森林”子菜单，或者点击主界面中的  按钮，打开随机森林模型界面，点击“打开”按钮，选择数据文件，数据文件将被装载到表格中，再点击“运行”按钮，将自动进行自定义数据的模型计算。



运行结果将在随机森林窗体的文本框中输出模型训练等过程和结果信息，以及相关图件，并提示已完成模型的计算。结果及图件输出情况与样例类似。





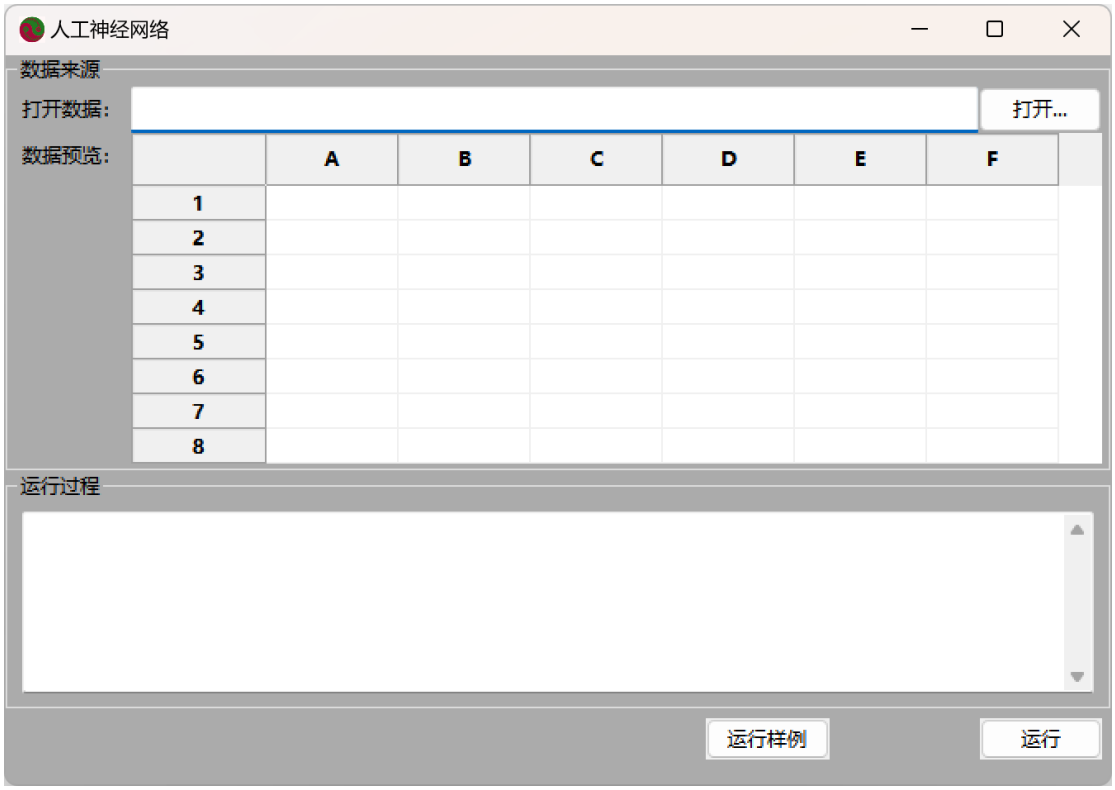
当决策树较为复杂时，绘制决策树结构图的时间较长，请耐心等待其执行完毕。

软件也提供了可以进行练习的“transfusion”数据，能够支持对模型进一步的了解，数据介绍可以参考：https://www.kaggle.com/datasets/whenamancodes/blood-transfusion-dataset?select=transfusion.csv。

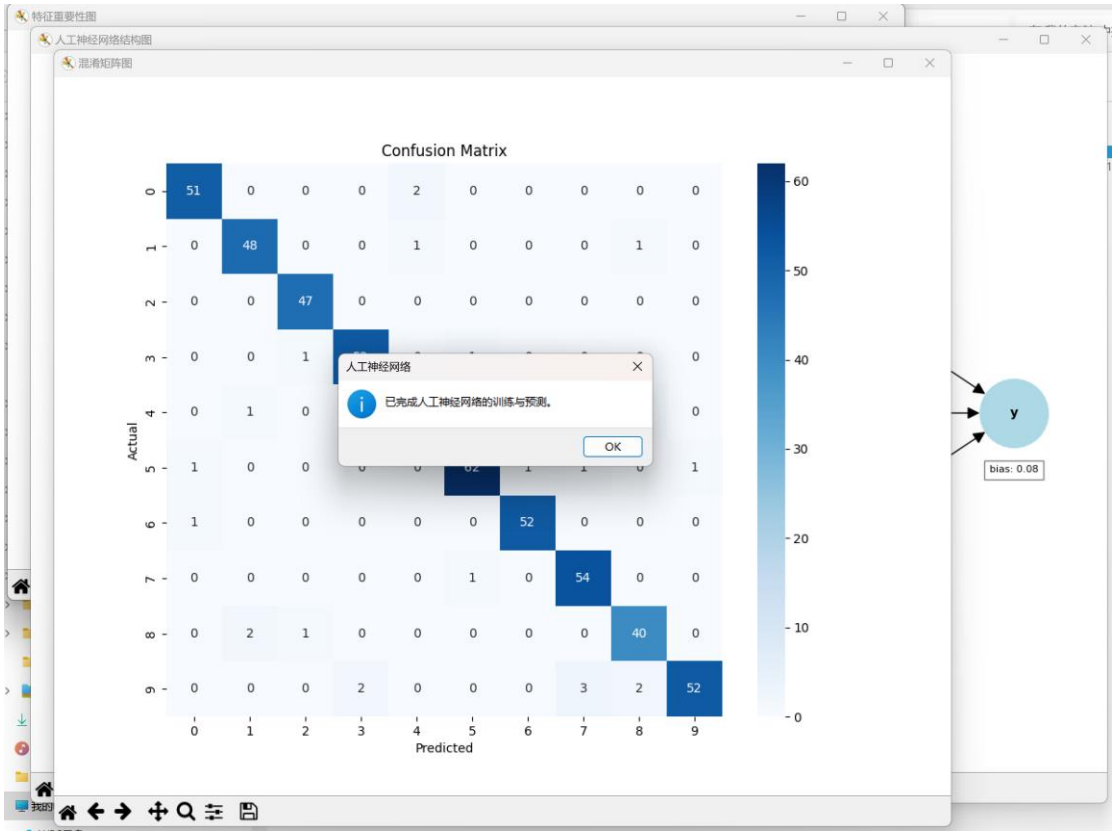
4.6 人工神经网络

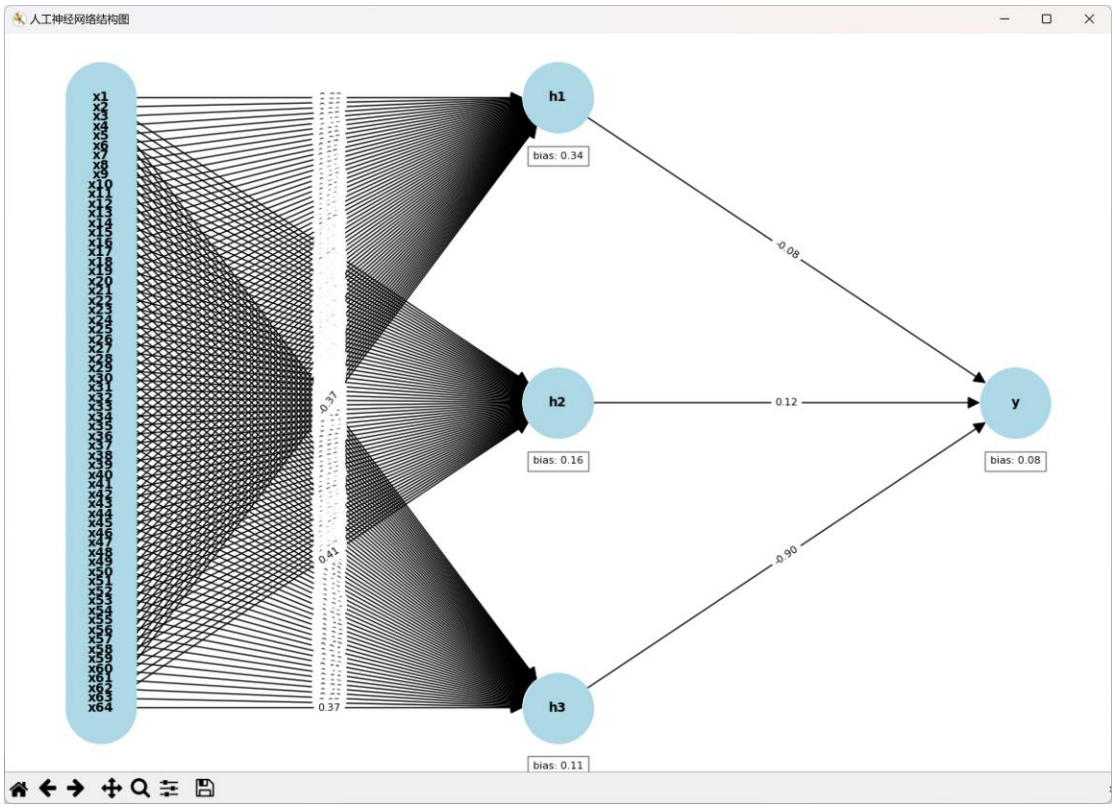
4. 6. 1 运行样例

使用内置的 iris 样例数据，进行人工神经网络模型计算的示例。点击“算法”菜单中的“人工神经网络”子菜单，或者点击主界面中的按钮，打开人工神经网络模型界面，点击“运行样例”按钮，将自动进行基于样例数据的模型计算。



运行结果将在人工神经网络窗体的文本框中输出模型训练等过程和结果信息，以及相关图件，并提示已完成模型的计算。图件包括图件包括人工神经网络结构图、学习曲线图、混淆矩阵图、特征重要性图。





人工神经网络结构图的作用主要是：

（1）可视化网络架构：结构图直观展示了神经网络的层次结构，包括输入层、隐藏层和输出层的节点（神经元）及其连接。这有助于理解网络的整体设计和复杂性。

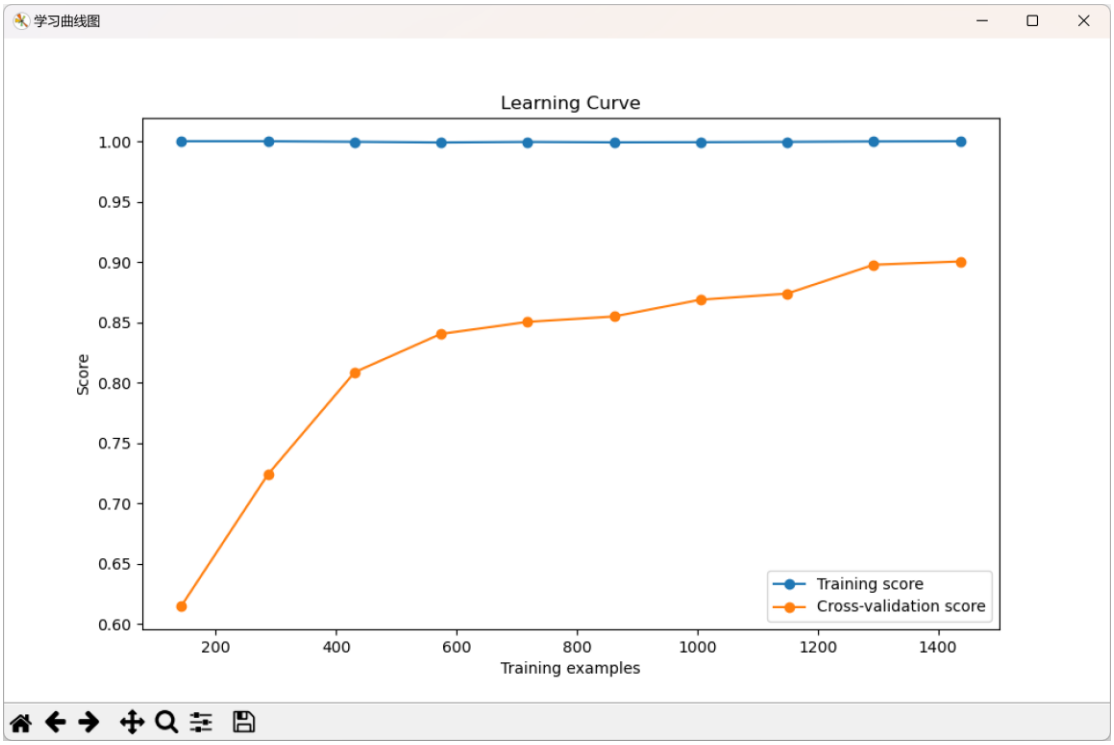
（2）理解特征处理：通过观察网络结构，可以了解数据是如何通过不同层进行处理和变换的，从输入到输出的过程，帮助分析特征提取和信息传递。

（3）识别层和节点的作用：结构图可以帮助识别每一层和每个节点在模型中的具体作用，分析不同层次的神经元如何对输入数据产生影响。

（4）调试与优化：可视化网络结构可以帮助诊断模型的潜在问题，例如层数过多导致的过拟合、节点数量不足导致的欠拟合等，从而为模型的优化提供依据。

（5）教学与交流：在教学和交流中，结构图是一种有效的工具，能够更容易地理解复杂的神经网络概念和算法。

（6）展示模型复杂性：结构图可以展示模型的复杂性，包括层数、每层的节点数、激活函数等，这有助于评估模型的能力和适用性。



学习曲线图的作用主要是：

（1）监控模型训练过程：学习曲线图展示了训练损失和验证损失随训练轮数的变化。这有助于监控模型在训练过程中的表现，判断模型是否在正确学习。

（2）识别过拟合和欠拟合：通过比较训练和验证损失的趋势，可以判断模型是否过拟合或欠拟合：

- 过拟合：当训练损失持续下降，而验证损失在某个点后上升时，表明模型在训练集上表现良好，但在验证集上表现不佳。
- 欠拟合：当训练和验证损失都高且相对接近时，表明模型无法捕捉数据的基本模式。

（3）优化训练参数：学习曲线图可以帮助确定训练过程中的超参数，如学习率、批量大小和训练轮数。通过分析损失的变化，可以调整这些参数以提高模型性能。

（4）评估数据集规模的影响：随着训练样本数量的增加，学习曲线可以展示模型性能的变化，帮助判断当前数据集是否足够大，是否需要更多数据以提高泛化能力。

（5）制定训练策略：通过观察学习曲线，可以制定适当的训练策略，例如提前停止（early stopping），以避免过拟合，并确保模型在验证集上的最佳表现。

（6）沟通模型表现：学习曲线图是与团队成员或利益相关者沟通模型表现和训练过程的有效工具，帮助他们理解模型的学习过程和性能趋势。

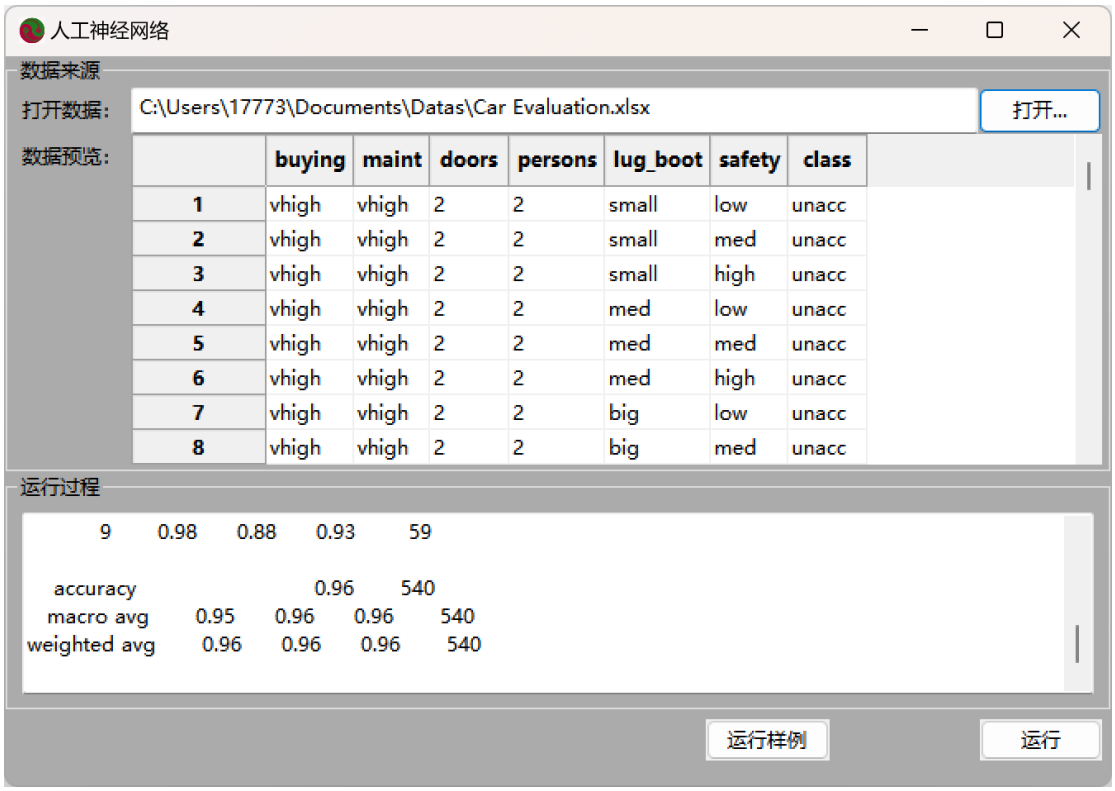
混淆矩阵图及特征重要性图的作用与前述的其它模型类似。

当网络结构较为复杂时，绘制结构图的时间较长，请耐心等待其执行完毕。

文本框中输出的信息可以进行复制，图表窗口中的图片可以进行浏览、保存等操作。

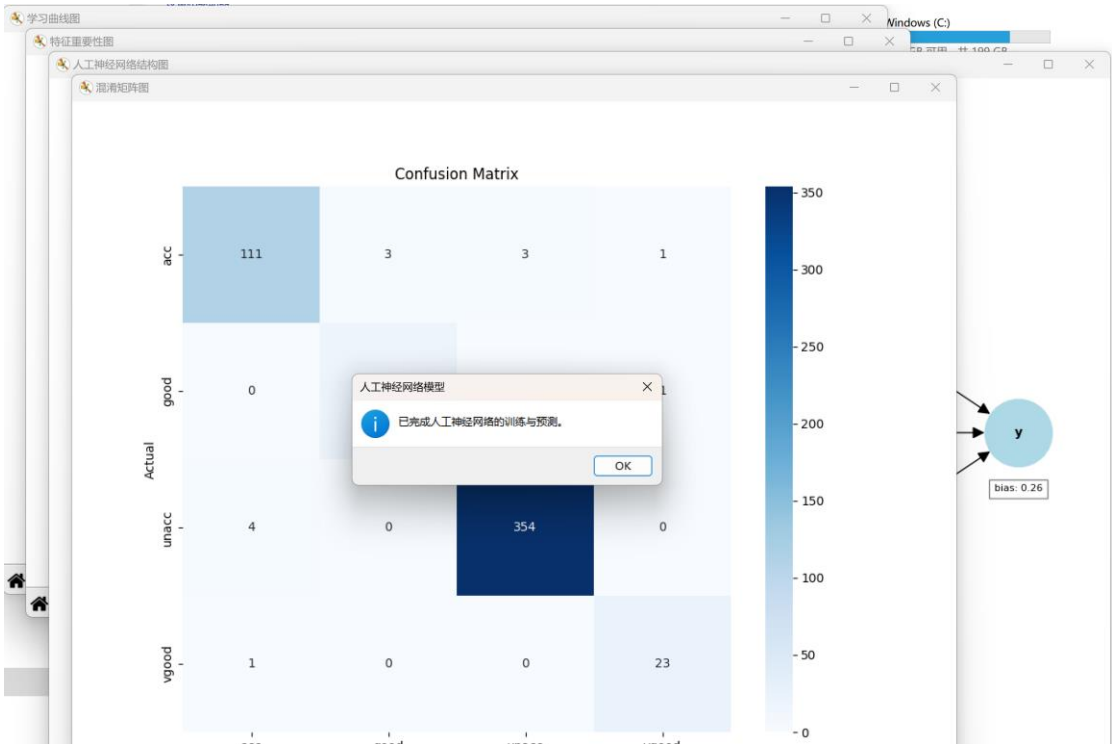
4.6.2 运行自定义数据

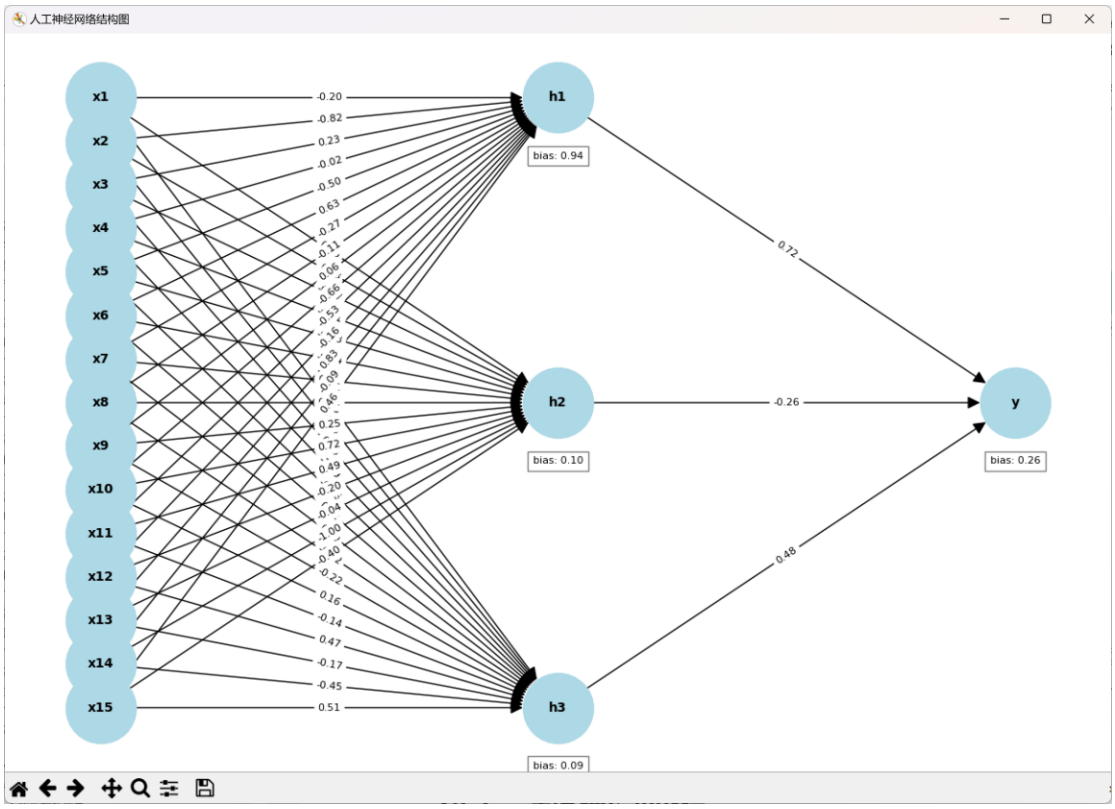
可以使用用户自定义的 csv、xls、xlsx 数据，进行人工神经网络模型的计算。点击“算法”菜单中的“人工神经网络”子菜单，或者点击主界面中的按钮，打开人工神经网络模型界面，点击“打开”按钮，选择数据文件，数据文件将被装载到表格中，再点击“运行”按钮，将自动进行自定义数据的模型计算。



运行结果将在人工神经网络窗体的文本框中输出模型训练等过程和结果信

息，以及相关图件，并提示已完成模型的计算。结果及图件输出情况与样例类似。



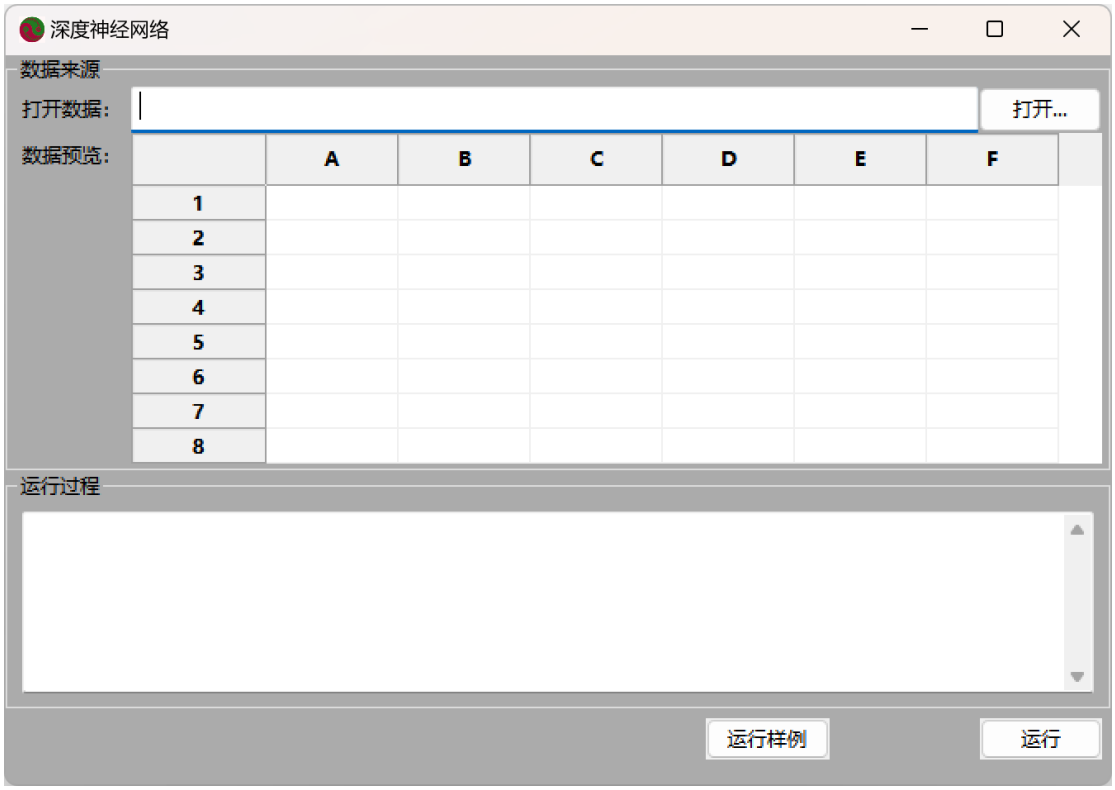


软件也提供了可以进行练习的“Car Evaluation”数据，能够支持对模型进一步的了解，数据介绍可以参考：https://archive.ics.uci.edu/dataset/19/car+evaluation。

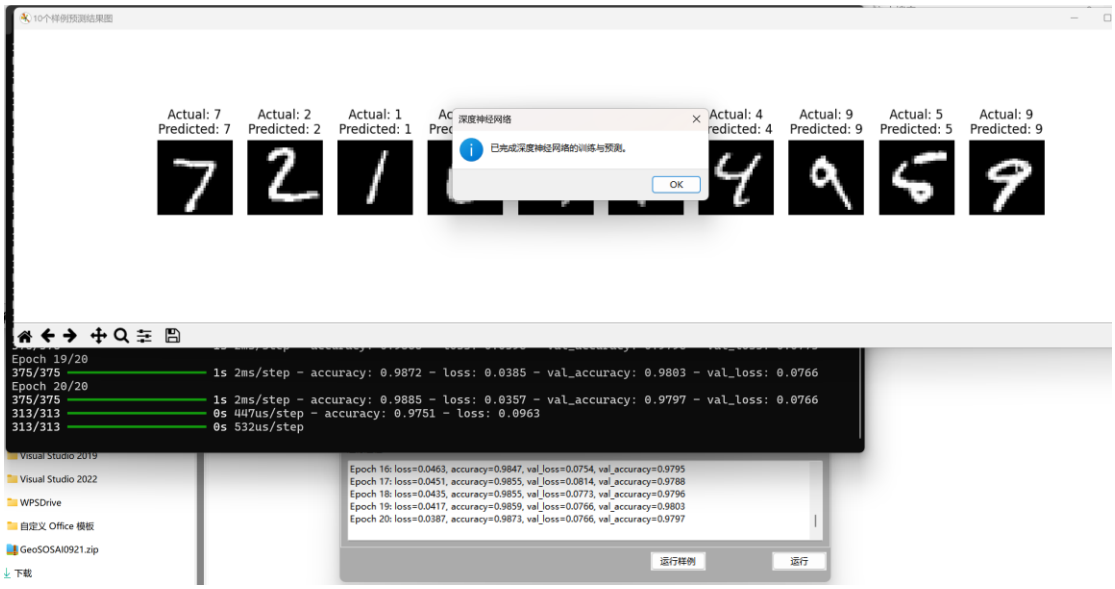
4.7　深度神经网络

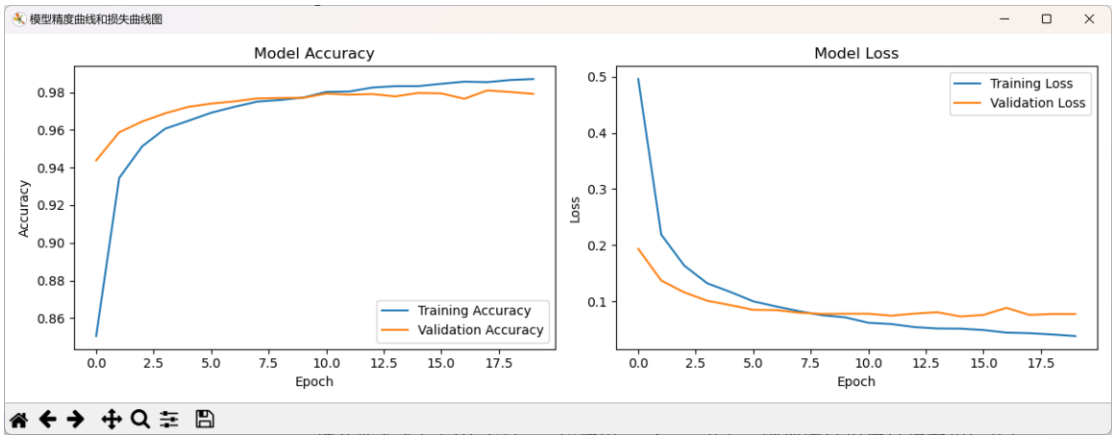
4. 7. 1 运行样例

使用内置的 mnist 样例数据，进行深度神经网络模型计算的示例。点击“算法”菜单中的“深度神经网络”子菜单，或者点击主界面中的  按钮，打开深度神经网络模型界面，点击“运行样例”按钮，将自动进行基于样例数据的模型计算。



运行结果将在深度神经网络窗体的文本框中输出模型训练等过程和结果信息，以及相关图件，并提示已完成模型的计算。包括模型精度曲线和损失曲线图、随机 10 个样例预测结果图。





模型精度曲线图的作用主要是：

（1）性能评估：精度曲线显示了训练和验证集的分类精度随训练轮数的变化，直观展示模型在不同阶段的表现，帮助评估模型的性能。

（2）识别过拟合和欠拟合：

- 过拟合：如果训练精度持续上升而验证精度在某一轮后不再上升或下降，表明模型对训练数据的拟合过强，但在验证集上表现不佳。
- 欠拟合：当训练和验证精度都较低且相近时，表明模型未能充分学习数据的特征。

（3）优化模型参数：通过观察精度曲线，可以判断训练过程中的超参数（如学习率、批量大小等）是否需要调整，以提高模型的表现。

（4）跟踪进展：精度曲线可以帮助监控模型在训练过程中的进展，了解模型何时达到了最佳表现。

（5）验证集选择：比较不同模型的精度曲线，可以帮助选择最佳的模型架构或超参数设置。

损失曲线图的作用主要是：

（1）监控学习过程：损失曲线显示了训练和验证集的损失值随训练轮数的变化，能够直观展示模型学习的进展。

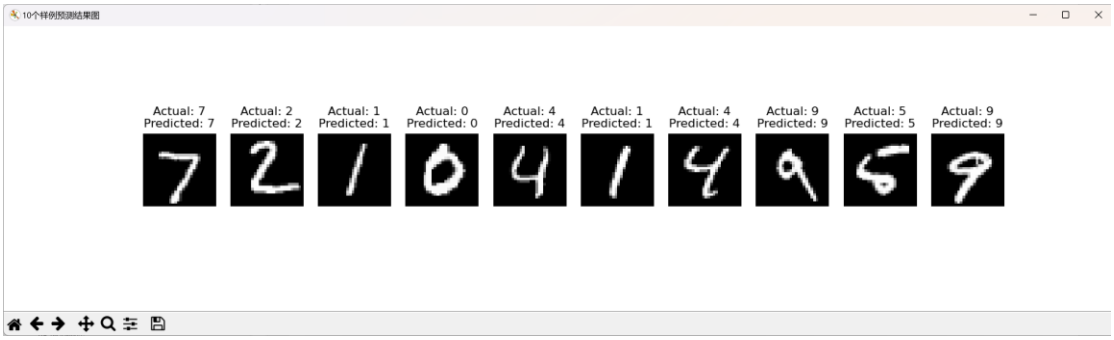
（2）识别模型问题：

- 过拟合：如果训练损失持续下降而验证损失在某一轮后上升，表明模型在训练集上表现良好，但在验证集上表现不佳。
- 欠拟合：如果训练损失和验证损失都较高且相对接近，表明模型未能捕捉数据的基本模式。

（3）确定训练停止：观察损失曲线有助于实施提前停止策略，以避免过拟合，并确保模型在验证集上的最佳性能。

（4）调整学习率：损失曲线的变化可以用来判断学习率的合适性。损失如果下降过慢，可能需要增加学习率；如果损失波动较大，可能需要减小学习率。

（5）数据集规模影响：随着训练样本数量的增加，损失曲线可以帮助评估模型的性能变化，判断是否需要更多数据来提高泛化能力。



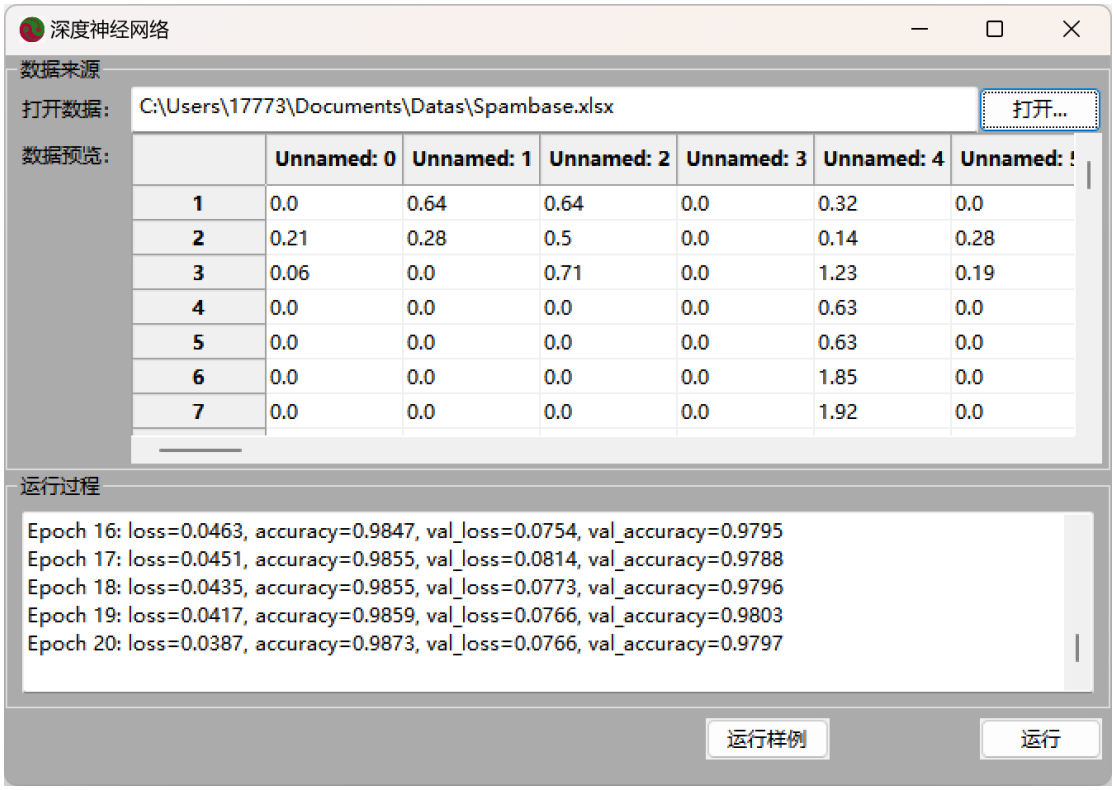
随机 10 个样例预测结果图主要是查看模型对于 10 个随机样例的预测程度如何，是否基本达到正确的预测结果。

如果数据量较大，深度神经网络的训练过程将较长，程序主窗体暂时无信息更新，请耐心等待其执行完毕。如使用运行包执行方式，则训练过程信息将在打开的命令行窗体中显示；如使用 PyCharm 执行，则训练过程信息将在程序执行面板中显示；使用 Python 命令行执行的方式，则训练过程信息也将在打开的命令行窗体中显示。

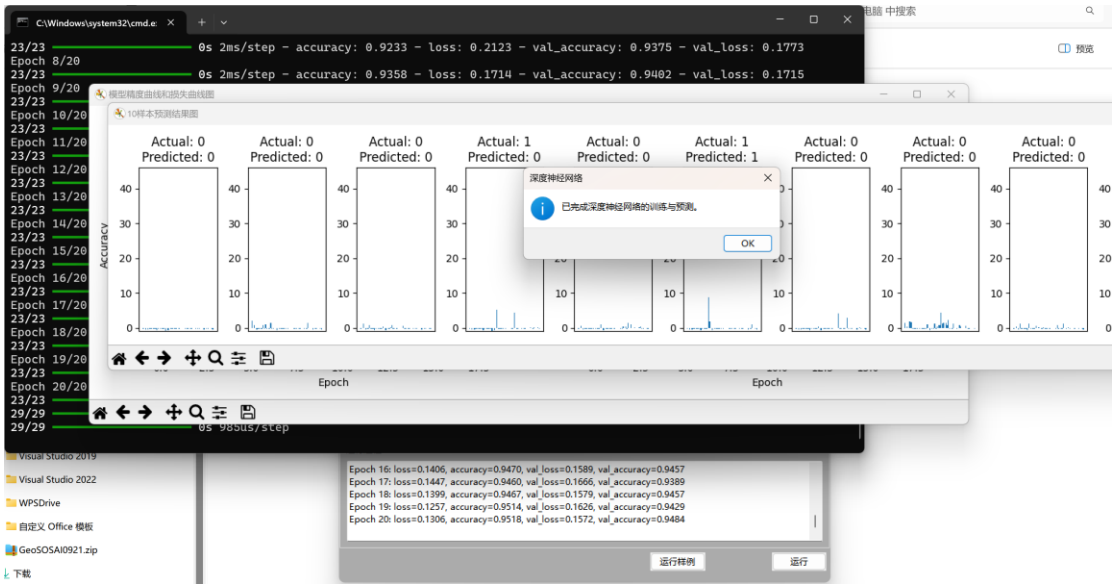
1. 在软件执行的命令行窗口中或 PyCharm 输出窗口中，可以看到详细的深度神经网络训练过程信息。2.文本框中输出的信息可以进行复制，图表窗口中的图片可以进行浏览、保存等操作。

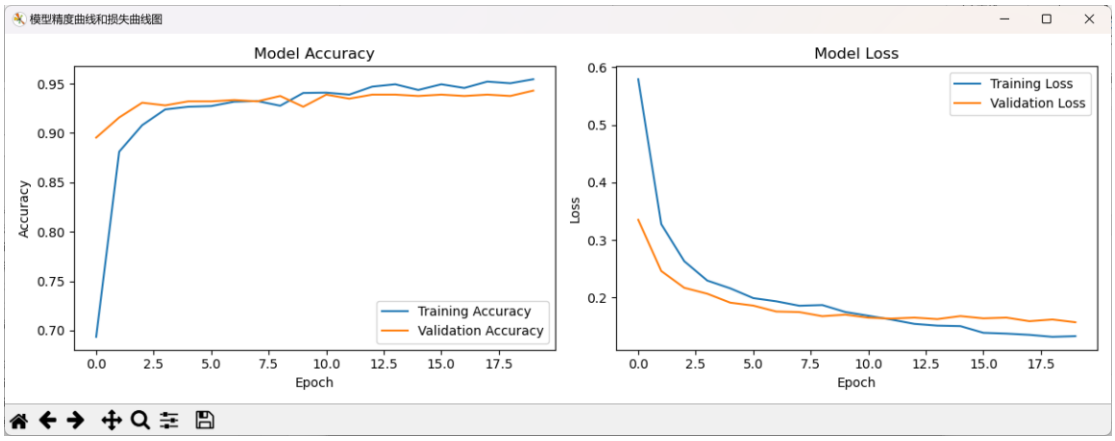
4.7.2 运行自定义数据

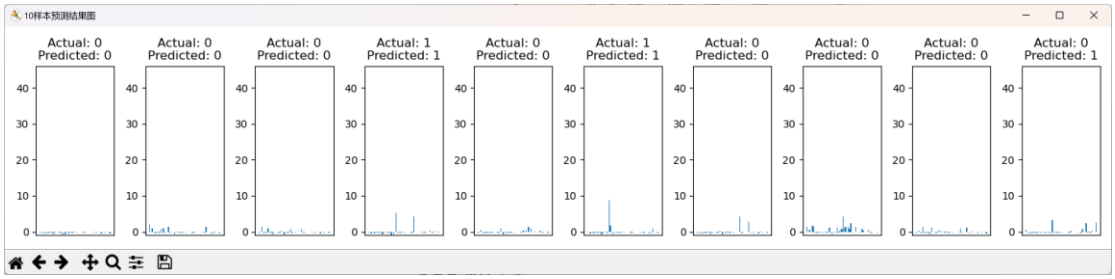
可以使用用户自定义的 csv、xls、xlsx 数据，进行深度神经网络模型的计算。点击“算法”菜单中的“深度神经网络”子菜单，或者点击主界面中的  按钮，打开深度神经网络模型界面，点击“打开”按钮，选择数据文件，数据文件将被装载到表格中，再点击“运行”按钮，将自动进行自定义数据的模型计算。



运行结果将在深度神经网络窗体的文本框中输出模型训练等过程和结果信息，以及相关图件，并提示已完成模型的计算。结果及图件输出情况与样例类似。





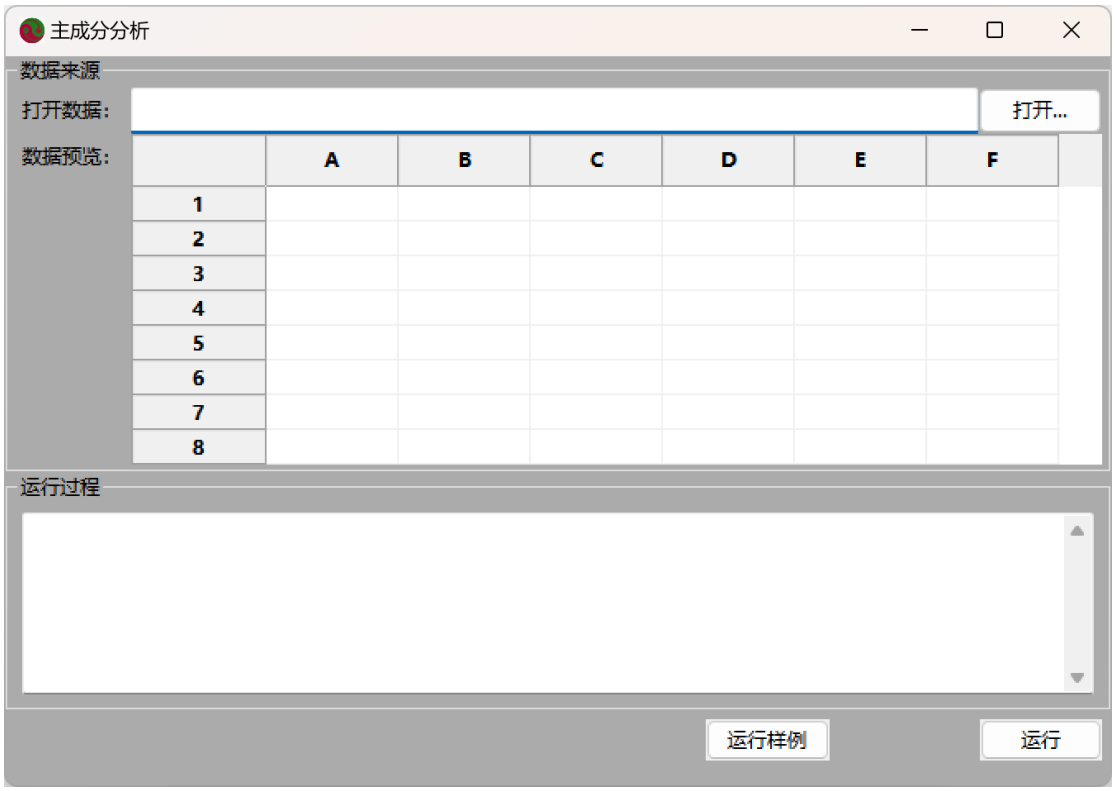


软件也提供了可以进行练习的“Spambase”数据，能够支持对模型进一步的了解，数据介绍可以参考：https://archive.ics.uci.edu/dataset/94/spambase。

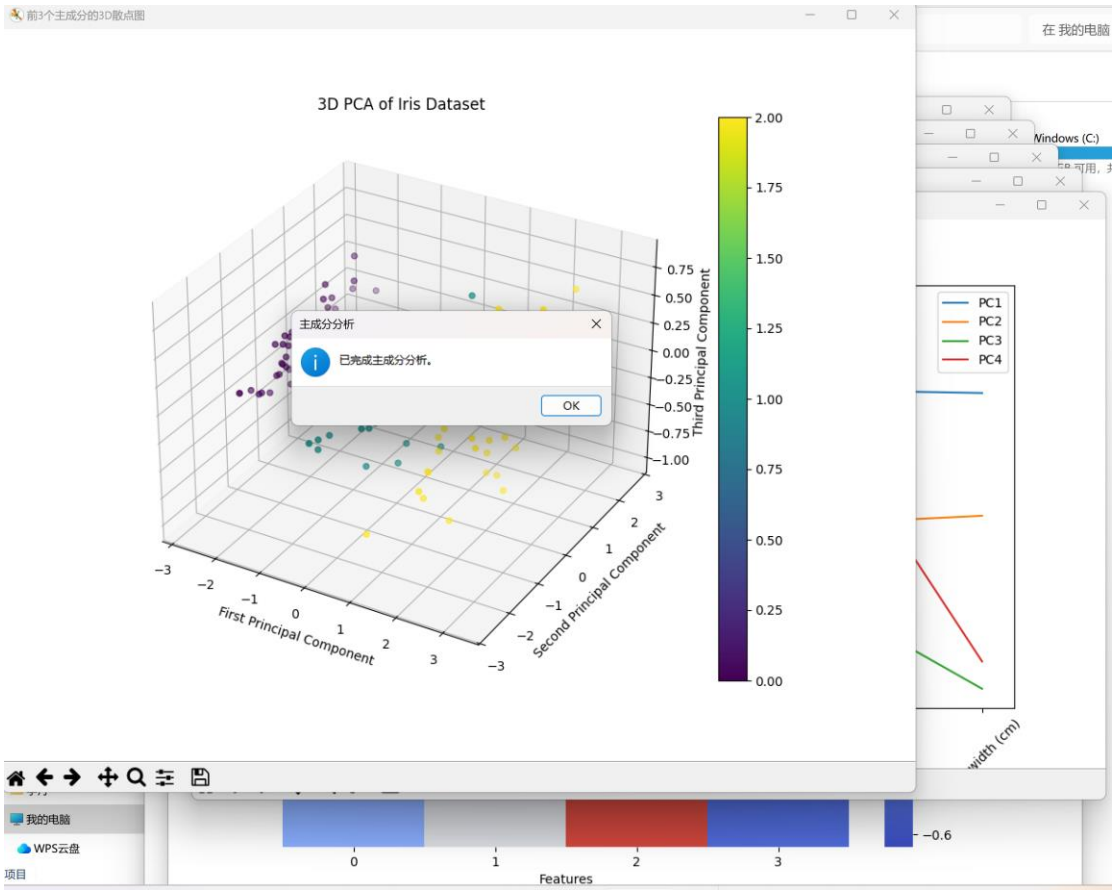
4.8　主成分分析

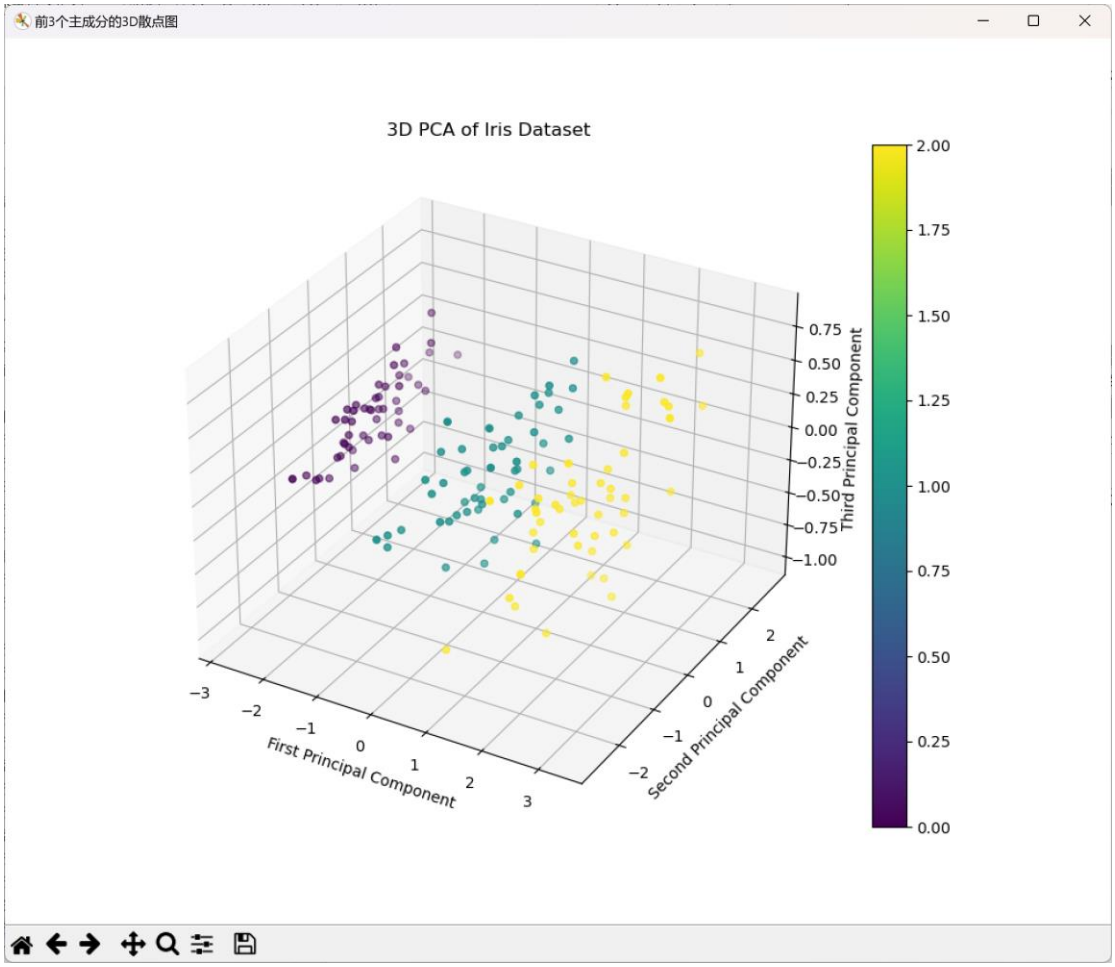
4. 8. 1 运行样例

使用内置的 iris 样例数据，进行主成分分析模型计算的示例。点击“算法”菜单中的“主成分分析”子菜单，或者点击主界面中的  按钮，打开主成分分析模型界面，点击“运行样例”按钮，将自动进行基于样例数据的模型计算。



运行结果将在主成分分析窗体的文本框中输出模型训练等过程和结果信息，以及相关图件，并提示已完成模型的计算。相关图件包括前 3 个主成分的 3D 散点图、负载图、热力图、双标图、累计解释方差比图、解释方差比图。


前3个主成分的3D散点图
3D PCA of Iris Dataset
First Principal Component
Second Principal Component
Third Principal Component
主成分分析
已完成主成分分析。
OK
PC1
PC2
PC3
PC4
Features
我的电脑
WPS云盘
项目



主成分的 3D 散点图的作用主要是：

（1）数据可视化：3D 散点图能够将高维数据投影到三维空间中，使得数据的分布、结构和模式更加直观，便于理解和分析。

（2）识别数据结构：通过观察不同主成分的散点图，可以识别数据中的聚类、分组和趋势，帮助发现潜在的结构或关系。例如，可能会发现某些数据点在特定方向上聚集，表明它们具有相似的特征。

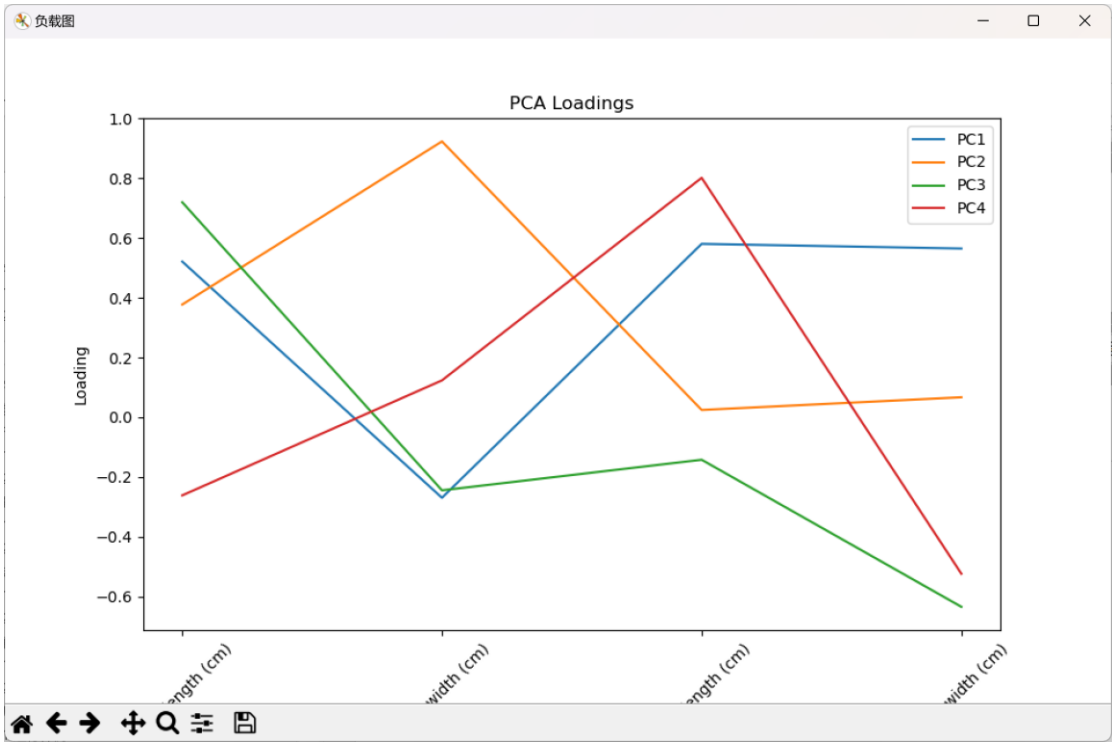
（3）降维效果评估：3D 散点图可以帮助评估 PCA 降维的有效性。通过观察主成分的分布，可以判断是否能够保留数据的主要变异性，并是否能够区分不同类别的数据。

（4）异常值检测：在 3D 散点图中，异常值或离群点通常会显得显眼。通过观察散点图，可以轻松识别出不符合其他数据模式的异常数据点。

（5）比较不同类别：如果数据集中包含多个类别（如不同的标签或组），可以在散点图中使用不同的颜色或标记来表示不同类别。这有助于直观理解各类别

在主成分空间中的分布和差异。

（6）辅助决策：3D 散点图提供了一种直观的方式来辅助决策。例如，在机器学习模型构建过程中，分析主成分的分布可以帮助选择特征和优化模型。



主成分分析（PCA）中的主成分负载图具有以下几个重要作用：

（1）特征贡献分析：负载图显示了每个原始特征对主成分的贡献程度。通过观察负载值，可以识别哪些特征对主成分的形成起到了关键作用，有助于理解数据的结构。

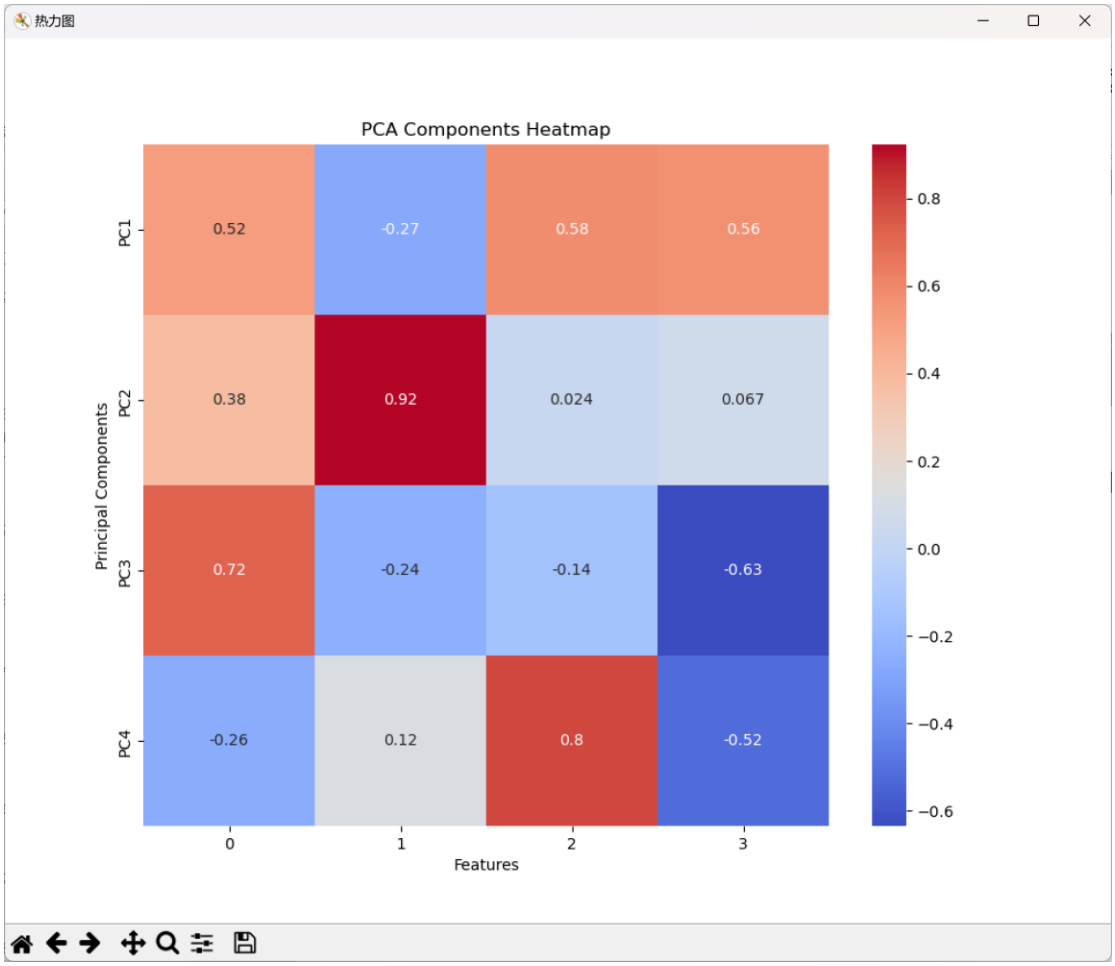
（2）特征选择：负载图可以帮助在特征选择过程中识别重要特征。特征负载值较高的特征通常对模型的解释能力和预测能力有重要影响，便于在后续分析中进行选择。

（3）数据解释：负载图提供了一种直观的方法来解释主成分的含义。通过分析负载值的大小和方向，可以理解每个主成分所代表的特征组合，进而为主成分赋予具体的解释。

（4）关联性分析：负载图可以揭示特征之间的关系。负载值的正负符号及其大小可以表明特征之间的相关性，如果两个特征的负载值在某个主成分上方向相似，它们可能在数据中有相似的表现。

（5）降维效果评估：通过分析主成分负载图，可以评估 PCA 降维的有效性，判断降维后是否保留了原始数据中的重要特征和信息。

（6）可视化数据结构：负载图能够直观展示高维数据在低维空间中的投影，有助于理解数据的整体结构和分布。



主成分分析（PCA）中的主成分热力图具有以下几个重要作用：

（1）可视化特征负载：热力图直观展示了各个原始特征在主成分上的负载值（权重）。通过颜色的深浅，可以快速识别哪些特征对主成分的贡献较大，帮助理解数据的结构。

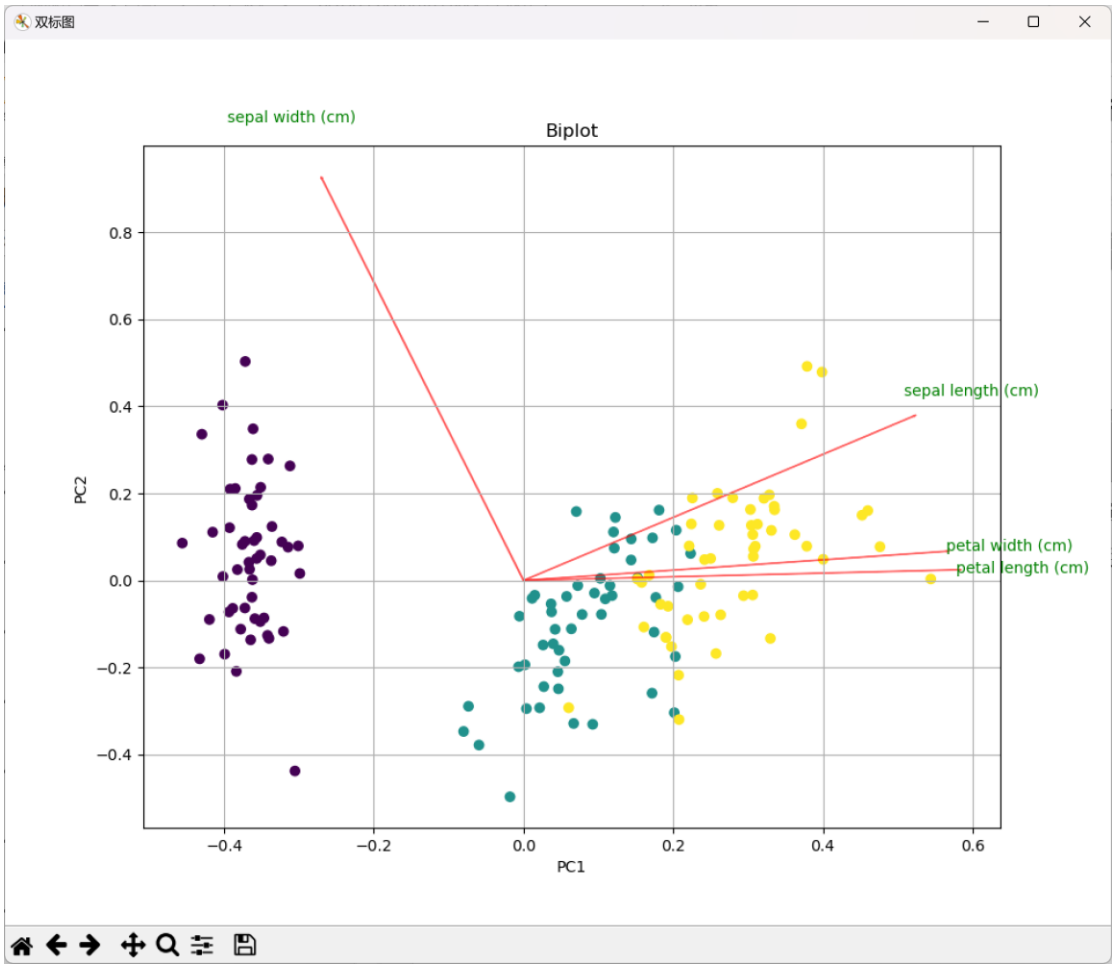
（2）识别重要特征：热力图使得重要特征的识别变得更加直观。负载值较高的特征通常对主成分的形成起到了关键作用，有助于后续的特征选择和分析。

（3）特征间的相关性分析：热力图能够揭示不同特征之间的关系。如果多个特征在某个主成分上具有相似的负载模式，可能表明这些特征之间存在一定的相关性。

（4）降维效果评估：通过观察热力图，可以评估 PCA 降维后对特征信息的保留情况，判断降维是否有效保留了原始数据中的重要信息。

（5）数据解释：热力图可以帮助分析和解释主成分的含义。通过观察哪些特征在主成分上的负载值较高，可以为主成分赋予具体的解释，从而更好地理解数据。

（6）直观展示数据模式：热力图以视觉化的方式展示了高维数据的模式和结构，使得数据分析过程更具可操作性，便于与他人沟通分析结果。



主成分分析（PCA）中的主成分双标图（biplot）具有以下几个重要作用：

（1）数据可视化：双标图将主成分的样本投影与特征负载可视化在同一图中，使得高维数据在二维空间中的分布一目了然。这有助于理解数据的整体结构和模式。

（2）样本与特征的关系：双标图同时展示样本点和特征向量，能够清晰地说明样本如何在主成分空间中分布，以及特征对样本分布的影响。特征向量的方向

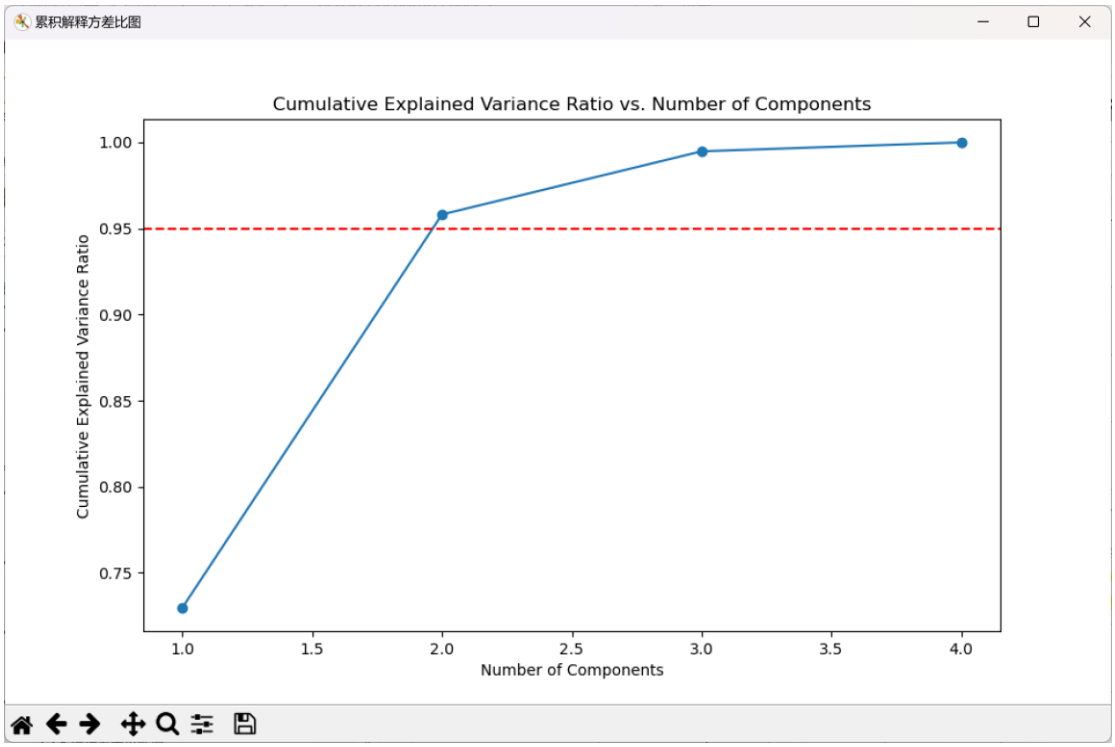
和长度指示了特征的重要性和影响力。

（3）聚类与分组识别：通过观察样本点的分布，可以识别出潜在的聚类或分组。如果样本点在双标图上靠近，可能表明它们在特征空间中有相似的特征。

（4）特征相关性分析：在双标图中，特征向量之间的夹角可以用来评估特征间的相关性。向量之间的角度越小，表示特征之间的相关性越高；如果向量呈现直角，则表示特征间无相关性。

（5）降维效果评估：通过观察样本在主成分上的分布，双标图可以帮助评估 PCA 降维的有效性。良好的降维应该能保留样本之间的结构和关系。

（6）数据解释：双标图能够帮助分析和解释主成分的意义。通过观察特征在图中的位置，可以更好地理解哪些特征对样本的主成分表达产生了重要影响。

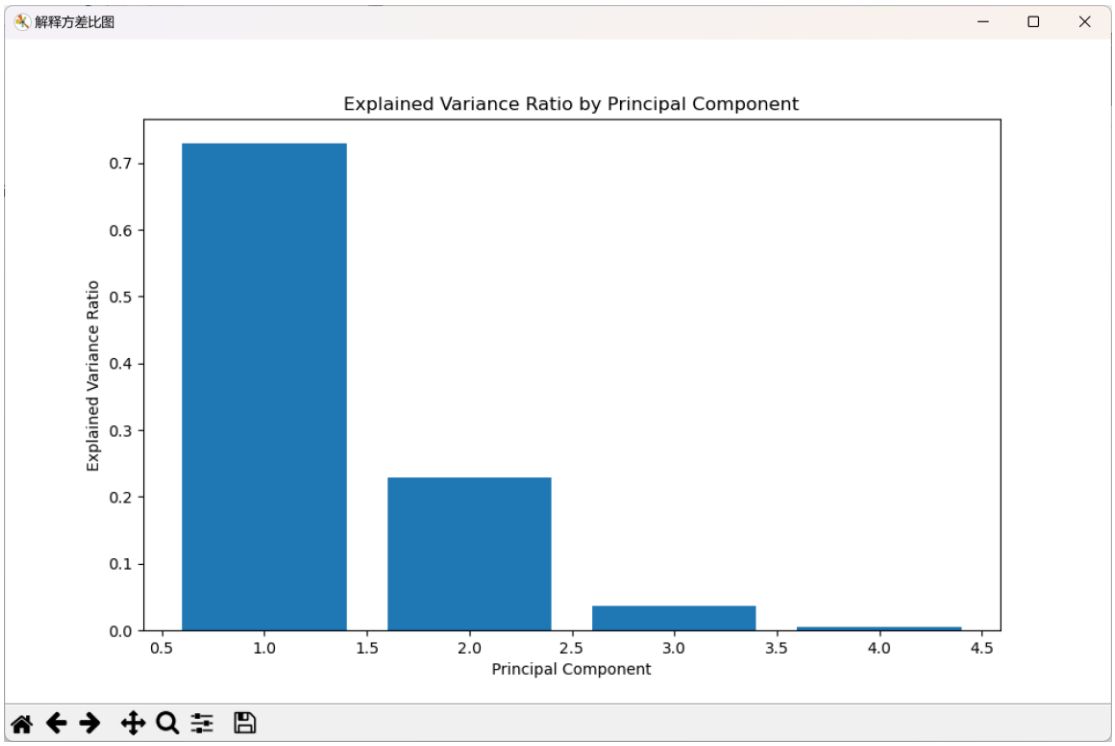


累计解释方差比图的作用主要是：

（1）整体变异性解释：累计解释方差比图展示了前 n 个主成分所累计解释的总方差比例。这使得分析者可以直观理解随着主成分数量的增加，模型解释能力的提升情况。

（2）选择主成分的阈值：通过观察图形，可以容易识别“肘部”位置，即累计解释方差开始趋于平稳的地方。这个点通常是选择主成分的理想数量，表示在此数量后增加更多主成分对解释能力的提升有限。

（3）直观展示数据结构：该图提供了一种直观的方式，展示如何通过少量主成分来捕捉数据的主要变异性，有助于说明降维的有效性和必要性。



解释方差比图的作用主要是：

（1）主成分的重要性：解释方差比图显示每个主成分所解释的方差比例，直观展示了各主成分对数据变异性的贡献。通过观察每个主成分的解释方差比，可以识别哪些主成分在数据中起到了关键作用。

（2）降维决策：该图有助于决定保留多少主成分。通常希望选择前几个主成分以最大化数据的解释能力，同时减少维度。例如，可以选择累积解释方差达到某个阈值（如 90%或 95%）的主成分数量。

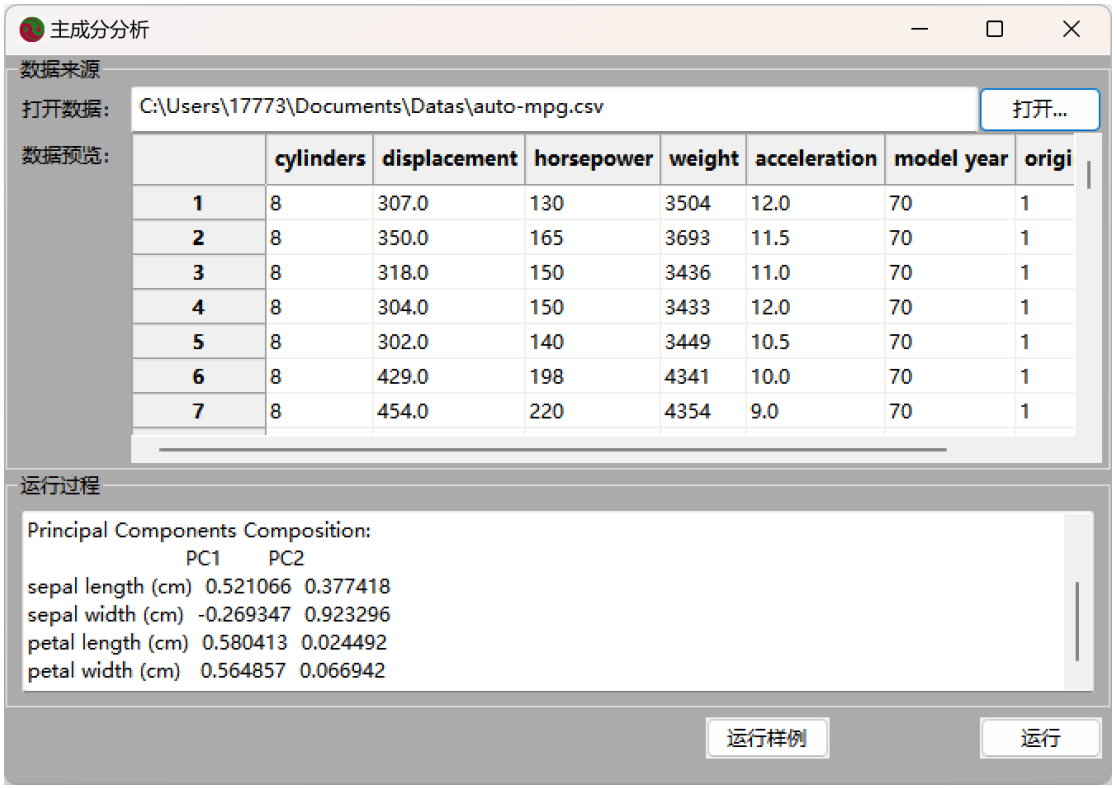
（3）评估特征的冗余性：如果前几个主成分的解释方差比很高，而后面的主成分解释的方差比很低，这可能表明数据中存在冗余特征，意味着可以通过降维简化数据集。

文本框中输出的信息可以进行复制，图表窗口中的图片可以进行浏览、保存等操作。

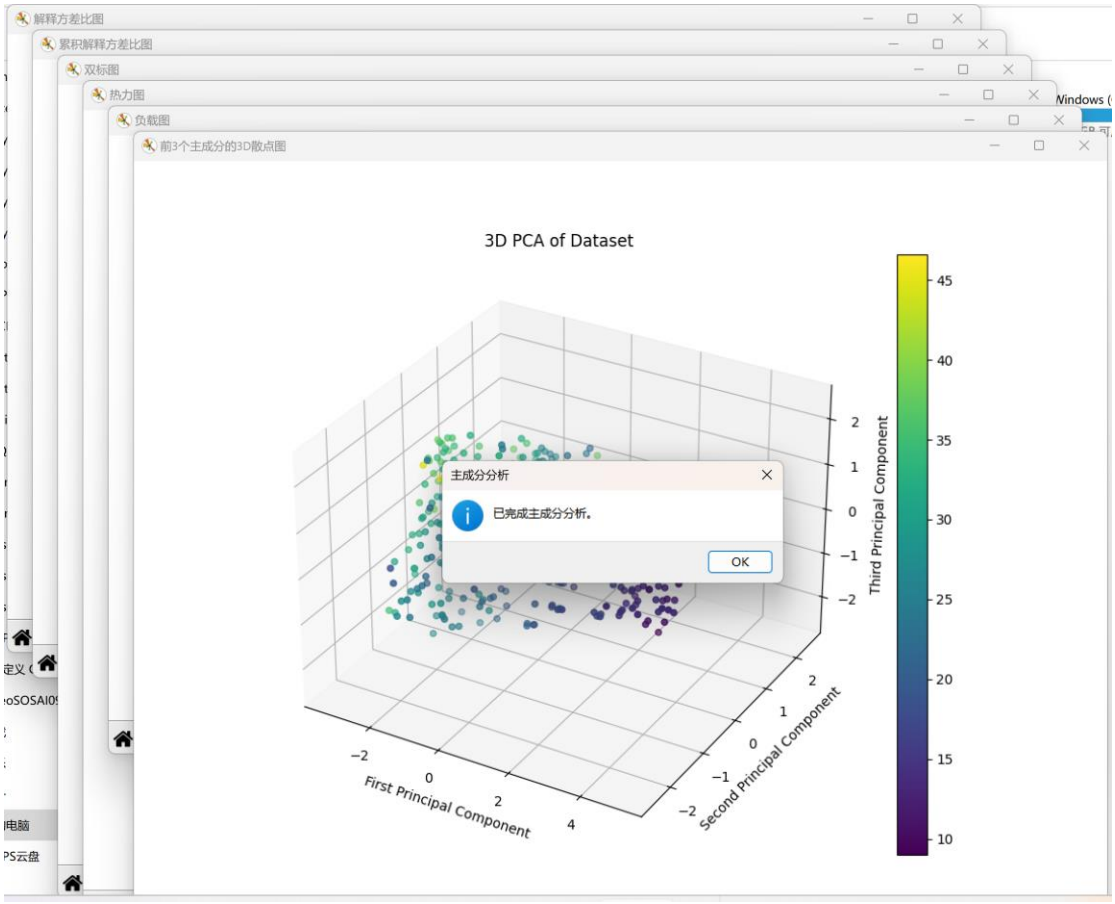
4.8.2 运行自定义数据

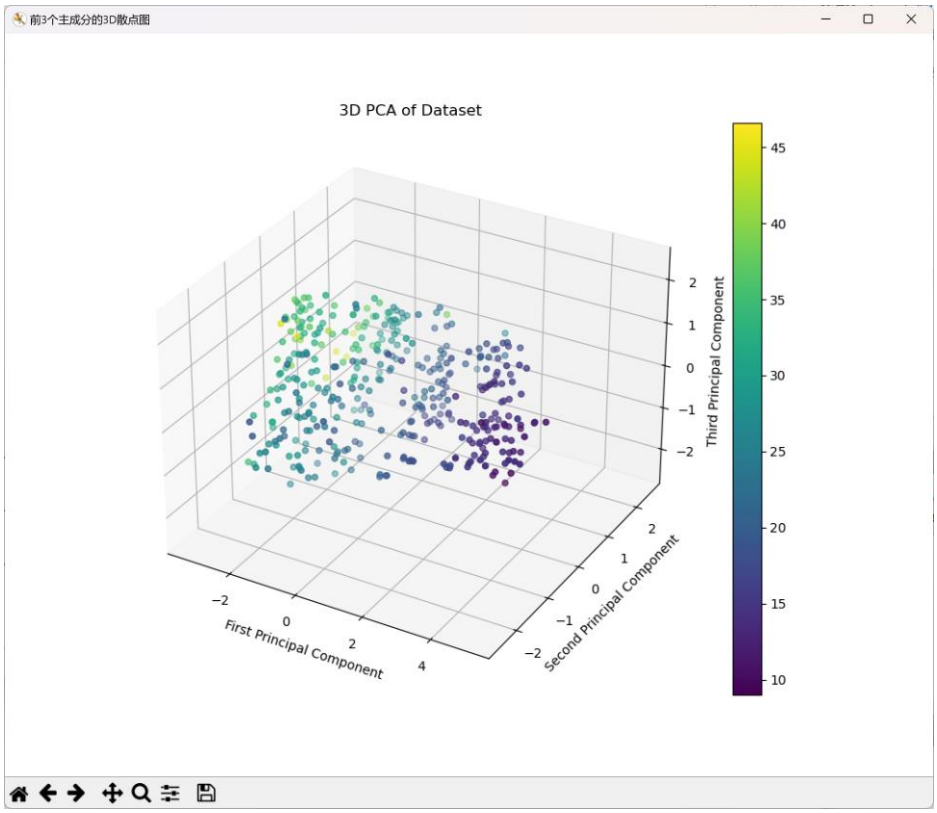
可以使用用户自定义的 csv、xls、xlsx 数据，进行主成分分析模型的计算。点

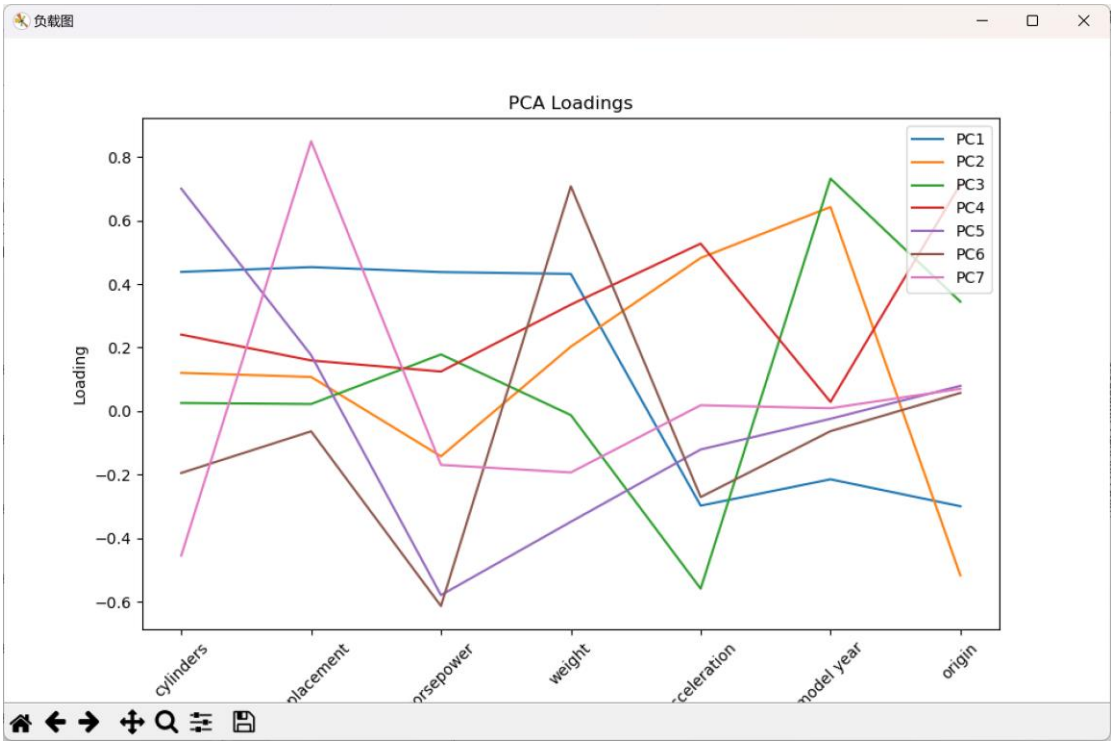
击“算法”菜单中的“主成分分析”子菜单，或者点击主界面中的  按钮，打开主成分分析模型界面，点击“打开”按钮，选择数据文件，数据文件将被装载到表格中，再点击“运行”按钮，将自动进行自定义数据的模型计算。

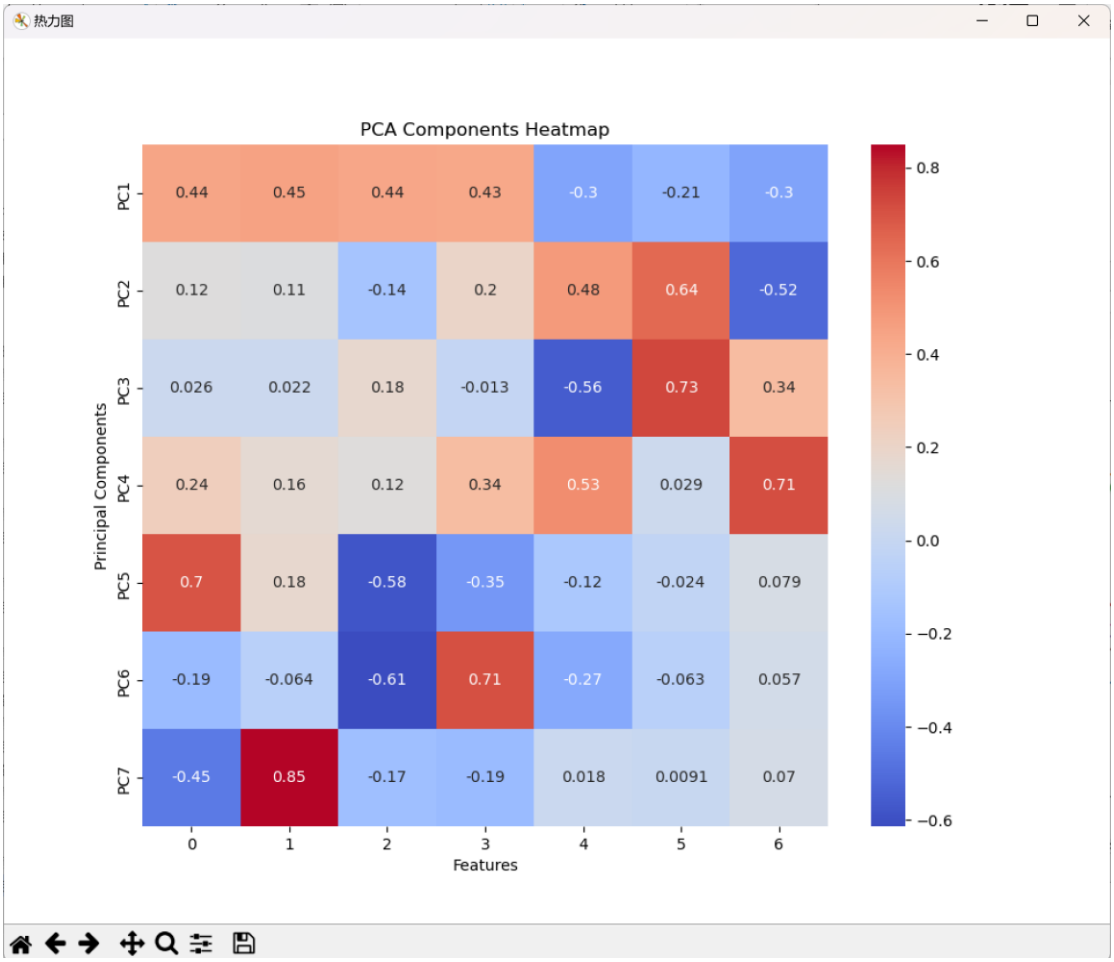


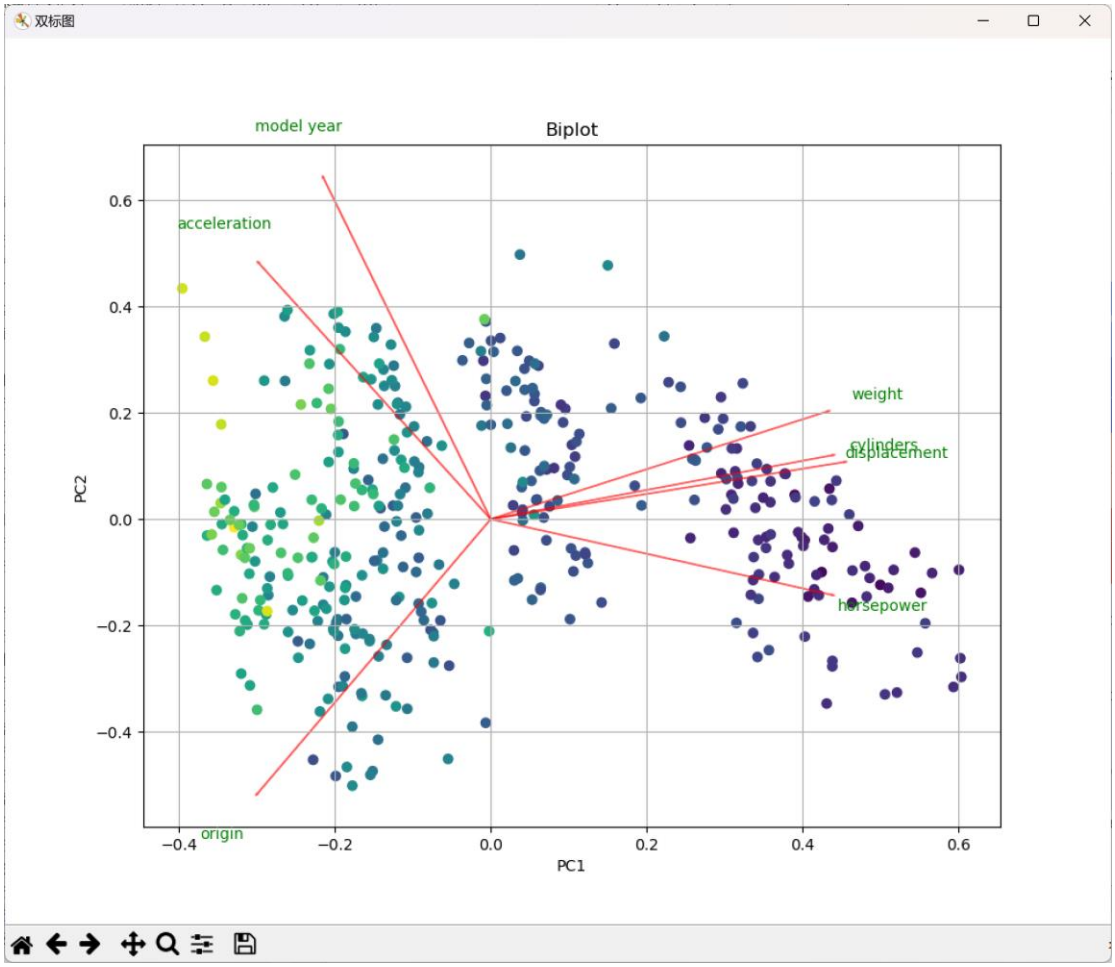
运行结果将在主成分分析窗体的文本框中输出模型训练等过程和结果信息，以及相关图件，并提示已完成模型的计算。结果及图件输出情况与样例类似。

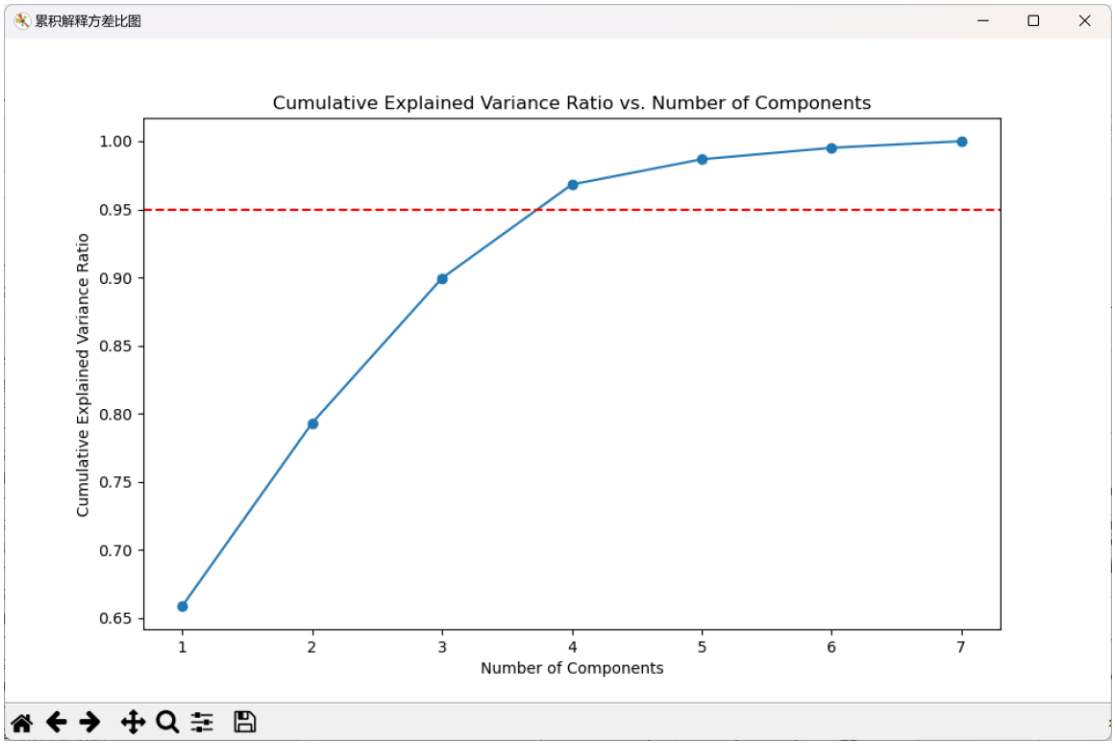

解释方差比图
累积解释方差比图
双标图
热力图
负载图
前3个主成分的3D散点图
3D PCA of Dataset
主成分分析
已完成主成分分析。
OK
First Principal Component
Second Principal Component
Third Principal Component

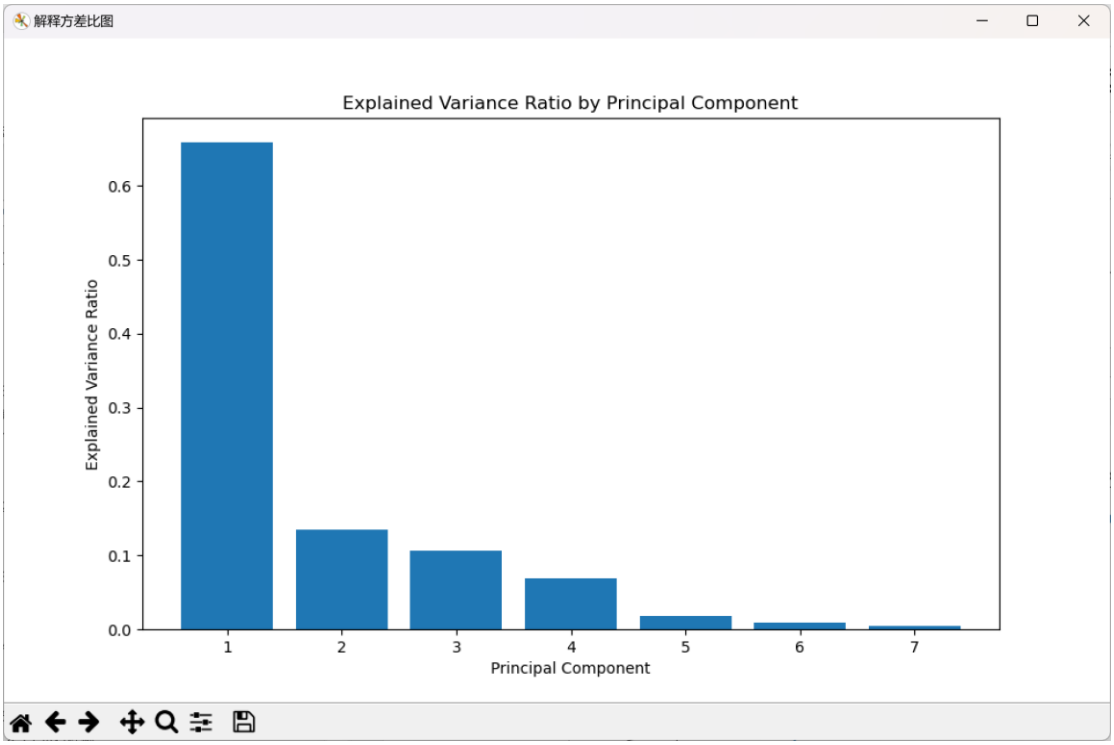

前3个主成分的3D散点图
3D PCA of Dataset
First Principal Component
Second Principal Component
Third Principal Component


负载图
PCA Loadings
PC1
PC2
PC3
PC4
PC5
PC6
PC7
Loading
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
cylinders
weight
origin


热力图
PCA Components Heatmap
Principal Components
PC1
PC2
PC3
PC4
PC5
PC6
PC7
0.44 0.45 0.44 0.43 -0.3 -0.21 -0.3
0.12 0.11 -0.14 0.2 0.48 0.64 -0.52
0.026 0.022 0.18 -0.013 -0.56 0.73 0.34
0.24 0.16 0.12 0.34 0.53 0.029 0.71
0.7 0.18 -0.58 -0.35 -0.12 -0.024 0.079
-0.19 -0.064 -0.61 0.71 -0.27 -0.063 0.057
-0.45 0.85 -0.17 -0.19 0.018 0.0091 0.07
0
1
2
3
4
5
6
Features
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6


双标图
Biplot
model year
acceleration
weight
cylinders
displacement
horsepower
origin
PC1
PC2
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.4
−0.2
0.0
0.2
0.4
0.6

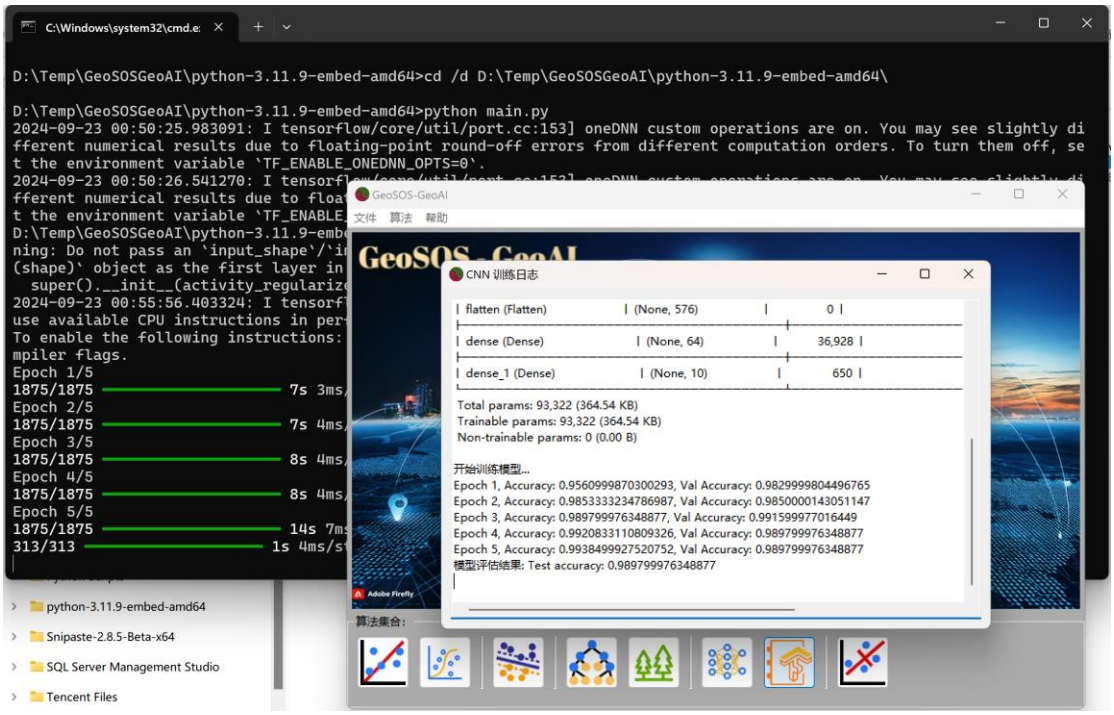

累积解释方差比图
Cumulative Explained Variance Ratio vs. Number of Components
Cumulative Explained Variance Ratio
Number of Components
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
1
2
3
4
5
6
7



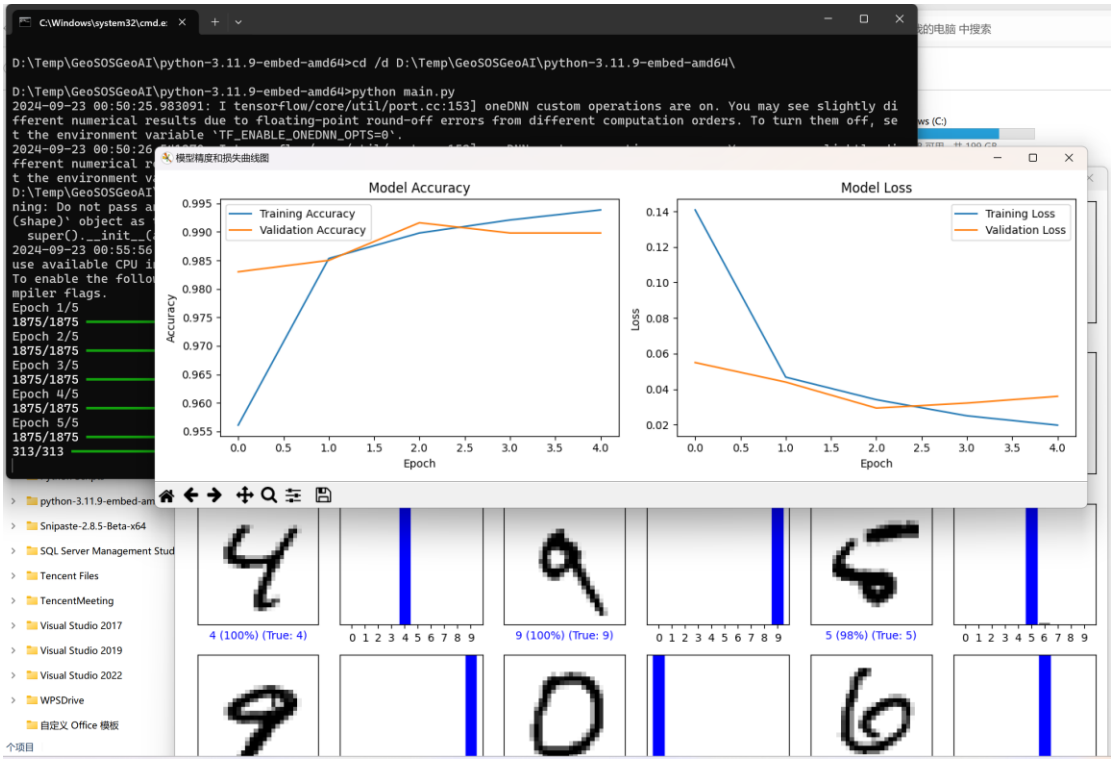
软件也提供了可以进行练习的“auto-mpg”数据，能够支持对模型进一步的了解，数据介绍可以参考：https://www.kaggle.com/datasets/uciml/autompg-dataset?resource=download。

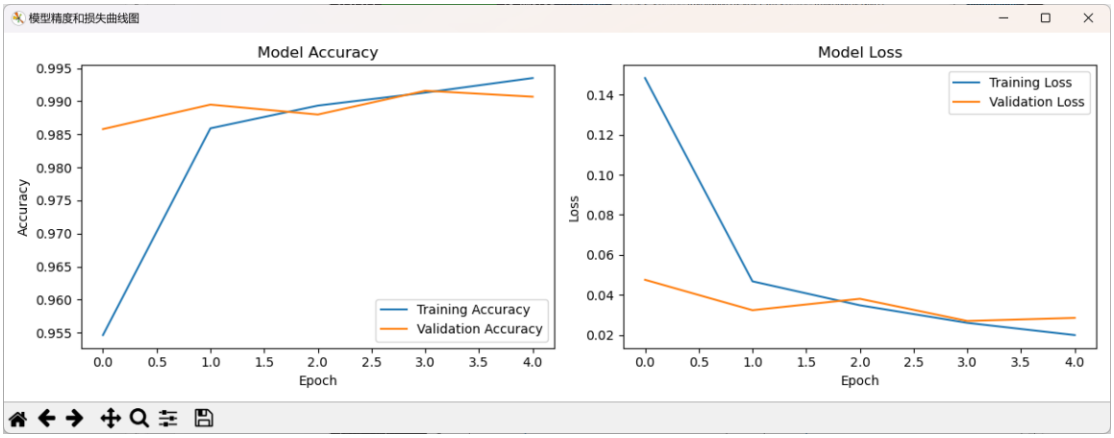
4.9 CNN 示例

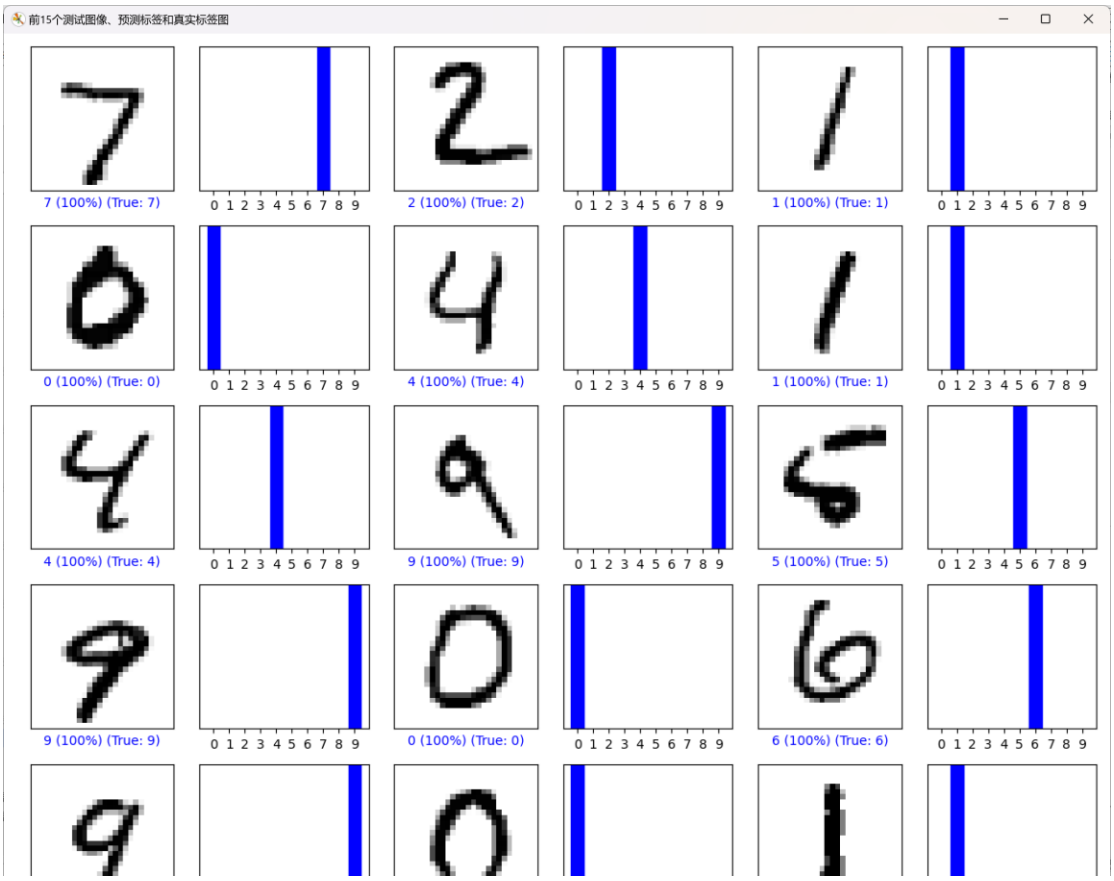
使用内置的 MNIST 样例数据，进行卷积神经网络（CNN）模型计算的示例。点击“算法”菜单中的“卷积神经网络（CNN）”子菜单，将自动进行基于样例数据的模型计算，命令行将显示 CNN 模型训练信息。



运行结果将在卷积神经网络（CNN）窗体的文本框中输出模型训练等过程和结果信息，以及相关图件，并提示已完成模型的计算。输出的信息包括 CNN 训练日志，相关图件包括模型精度和损失曲线图和前 15 个测试图像、预测标签和真实标签图。



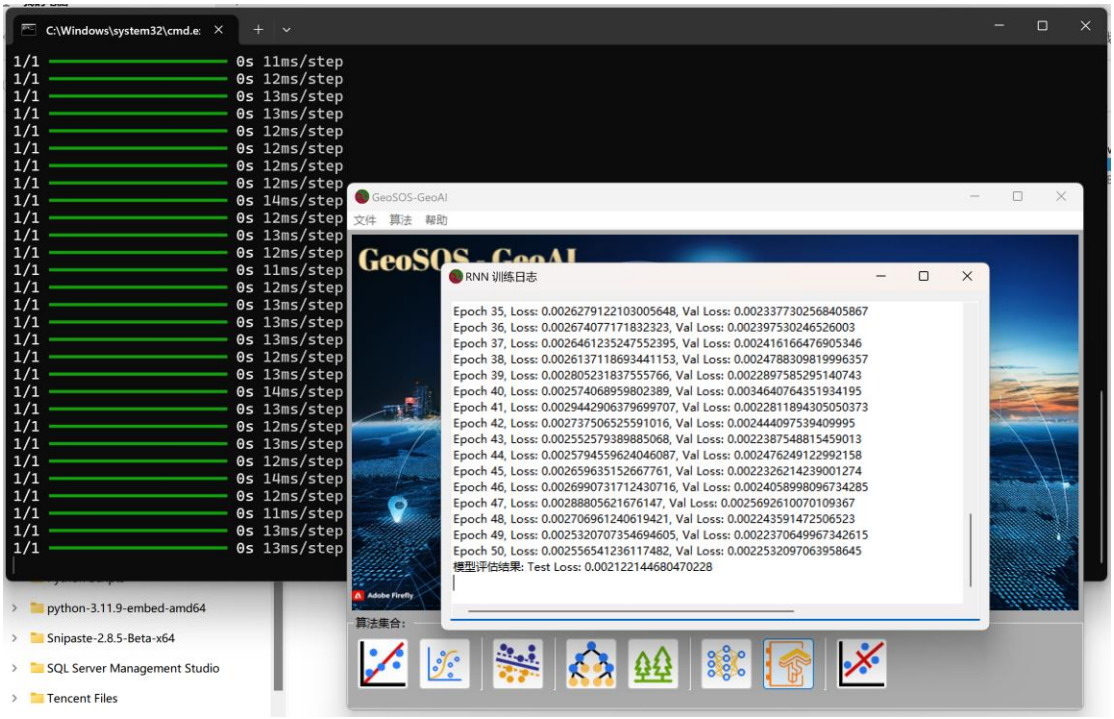




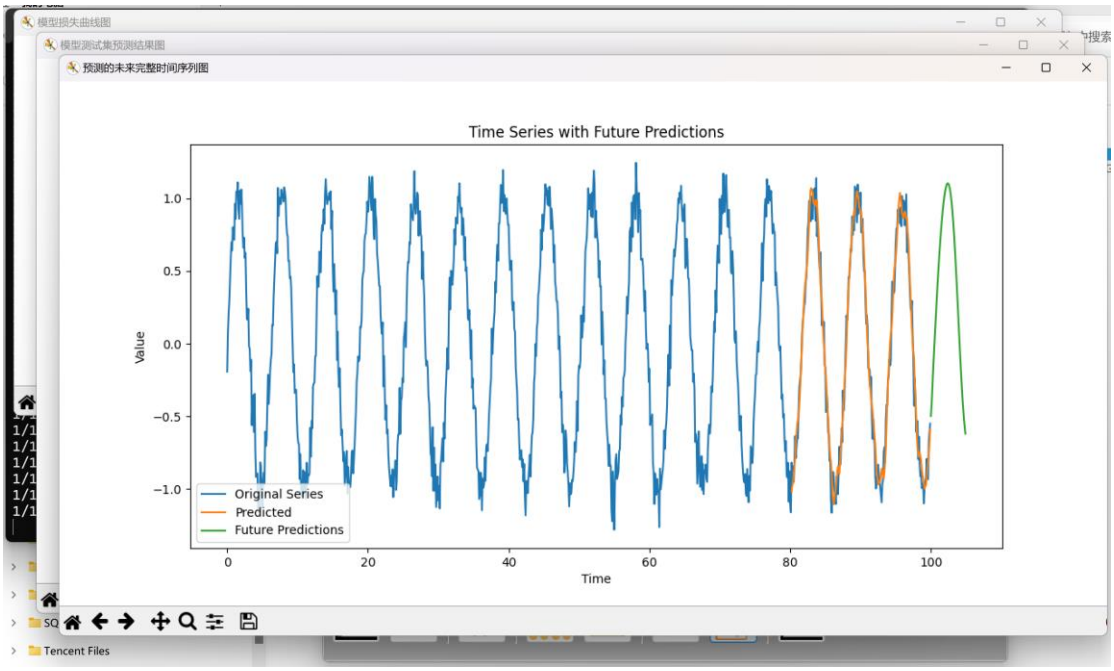
CNN 模型训练是计算密集型任务，将占用大量的计算机资源，并执行较长的时间，因此建议训练时不继续其它操作，等待训练完成。

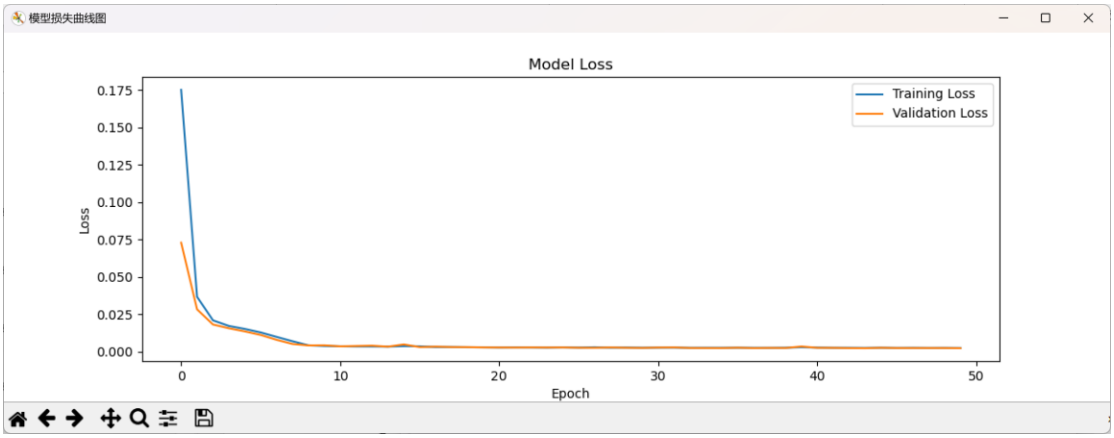
4.10 RNN 示例

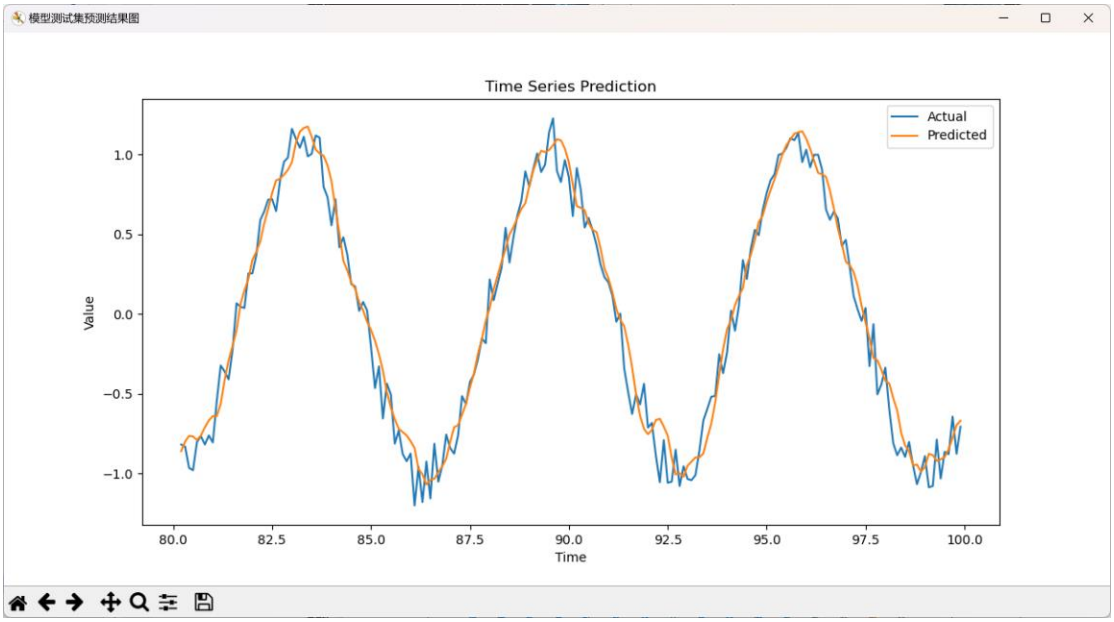
将自动生成样例数据，进行循环神经网络（RNN）模型计算的示例。点击“算法”菜单中的“循环神经网络（RNN）”子菜单，将自动进行基于样例数据的模型计算，命令行将显示 RNN 模型训练信息。

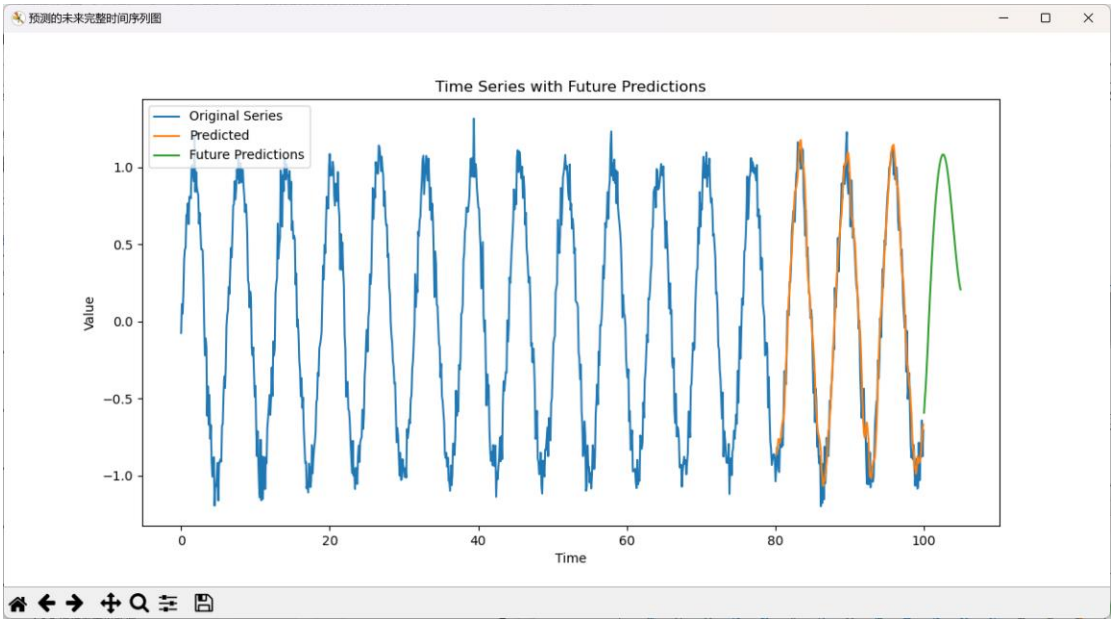


运行结果将在循环神经网络（RNN）窗体的文本框中输出模型训练等过程和结果信息，以及相关图件，并提示已完成模型的计算。输出的信息包括 RNN 训练日志，相关图件包括模型损失曲线图、模型测试集预测结果图和预测的未来完整时间序列图。







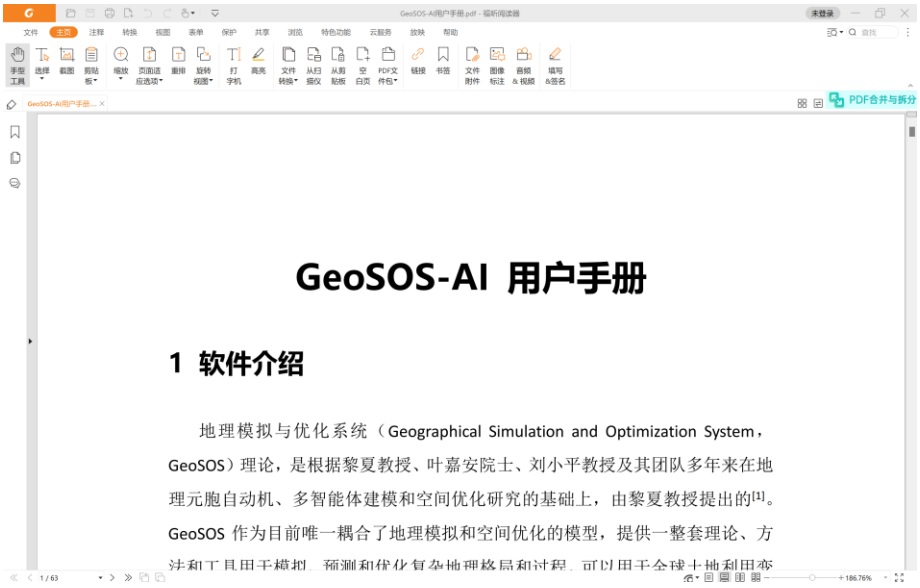


RNN 模型训练是计算密集型任务，将占用大量的计算机资源，并执行较长的时间，因此建议训练时不继续其它操作，等待训练完成。

4.11 帮助及关于

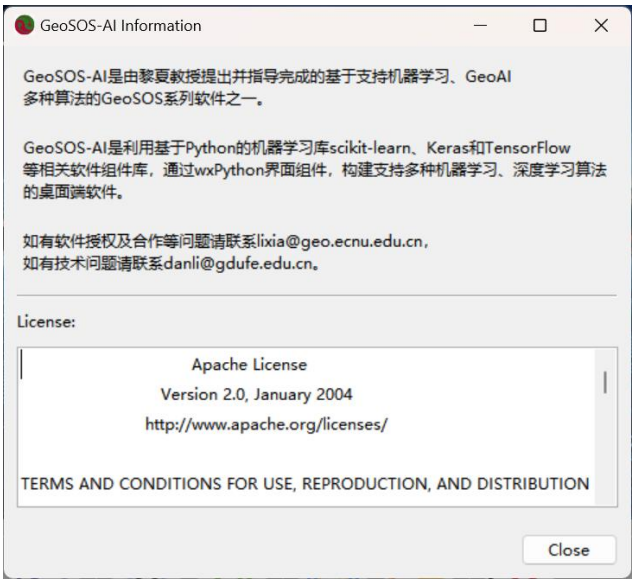
4.11.1 软件帮助信息

点击“帮助”菜单中的“帮助”子菜单，可以打开软件附带的 PDF 格式的帮助文件，可以查看本用户手册。



4.11.2 软件关于信息

点击“帮助”菜单中的“关于...”子菜单，，将显示本软件的相关信息，如名称、简要说明、联系方式、权利信息等。



5 软件权利声明

软件为免费软件，以帮助学术界、规划业界和政府部门更好地从事研究和决策，**但使用其进行学术论文发表及其他工作时应公开注明使用了本软件**。

本软件由黎夏教授进行理论指导，李丹总负责其软件的设计和开发。如有相关学术问题、软件授权及合作等问题，请联系黎夏教授（lixia@geo.ecnu.edu.cn）。如在软件安装、使用过程中存在问题，或有软件改进方面的意见和建议，请联系李丹（danli@gdufe.edu.cn）。感谢您对 GeoSOS 的关注和支持！

本软件采用 Apache License Version 2.0 软件协议，将源代码放置在 GitHub 库中，非常欢迎软件用户等感兴趣的个人、机构共同开发和维护该软件，促进软件的升级和更新，以促进软件更好的应用。