

**2024年《机器学习》**

**工程报告**

**（个人版）**



**课 程：** 机器学习

**姓 名：** 杨建华

**学 号：** 2023217487

**完成时间：** 2025-05-21

一． 自己所作工作的简介

**Kaggle平台数据集上的机器学习算法研究**

包括数据分析、数据处理、模型分类预测与效果评估等任务。

数据集：

良恶性肿瘤分类

样本量：569 特征量：30

**八种**机器学习模型：

逻辑回归、决策树、随机森林、KNN、NB、SVM、XGBoost、LightGBM

**四种**降维可视化方法：

PCA、NCA、UMAP、t-SNE

二、研究背景与意义

研究背景

随着人工智能和数据科学的迅猛发展，机器学习与深度学习技术在医疗健康领域得到了广泛应用，尤其是在疾病诊断和预测方面。研究人员借助先进的算法和丰富的数据资源，开发出高效的诊断工具，从而显著提高了疾病早期发现的准确性，进而提升了患者的治愈率和生活质量。

在本研究中，我们选取了具有代表性的数据集进行分析，其中心脏病数据集包含1025个样本和13个特征，用于心脏病的诊断。

主要研究方法

本研究采用了四种降维和可视化方法：

主成分分析（Principal Component Analysis,PCA）

近邻成分分析（Neighborhood Component Analysis,NCA）

UMAP（Uniform Manifold Approximation and Projection）

t-SNE（t-distributed Stochastic Neighbor Embedding）

此外，为了全面评估不同机器学习算法的性能，我们使用了以下机器学习模型：

逻辑回归（Logistic Regression）

决策树（Decision Tree）

随机森林（Random Forest）

K近邻（K-Nearest Neighbors, KNN）

朴素贝叶斯（Naive Bayes, NB）

支持向量机（Support Vector Machine, SVM）

XGBoost

LightGBM

研究意义

通过比较不同算法的性能，本研究旨在找到最适合用于心脏病诊断的模型，从而提高疾病诊断的准确性和效率。

三、模型方法

数据集介绍

威斯康星州乳腺癌（诊断）数据集：

样本量：569

特征量：30

任务：良性和恶性肿瘤分类

模型方法

在本研究中，我们使用了多种机器学习模型，具体如下（本人负责前四个）：

1. 逻辑回归（Logistic Regression）：

逻辑回归是一种经典的线性分类模型，适用于二分类任务。它通过最大化对数似然函数来估计模型参数，从而对样本进行分类。

2. 决策树（Decision Tree）：

决策树是一种基于树结构的模型，通过递归地将数据集分割成更小的子集来构建决策树。每个节点根据某个特征的阈值进行分裂，直至满足停止条件。

3. 随机森林（Random Forest）：

随机森林是一种集成学习方法，通过训练多个决策树并取其预测结果的多数票来提高模型的准确性和稳定性。它能够有效降低单个决策树的过拟合风险。

4. K近邻（K-Nearest Neighbors,KNN）：

KNN是一种基于实例的学习方法，通过计算待分类样本与训练样本之间的距离，选择距离最近的K个邻居进行投票，从而决定分类结果。

降维与可视化方法

1. 主成分分析（Principal Component Analysis,PCA）：

PCA是一种线性降维技术，通过将原始高维数据投影到低维空间，最大化投影后的数据方差，从而实现降维。

2. 近邻成分分析（Neighborhood Component Analysis,NCA）：

NCA是一种监督学习的降维方法，通过学习一个变换矩阵，使得相同类别的样本在变换后的空间中更接近，不同类别的样本更远离。

3. UMAP（Uniform Manifold Approximation and Projection）：

UMAP是一种非线性降维方法，通过构建高维数据的邻域图并在低维空间中优化其投影，保留高维数据的结构。

4. t-SNE（t-distributed Stochastic Neighbor Embedding）：

t-SNE是一种用于数据可视化的非线性降维技术，通过最小化高维空间和低维空间中相邻点之间的概率分布差异，将高维数据映射到二维或三维空间中，便于可视化分析。

四、系统设计

1. **具体流程**

1. 数据加载

使用合适的工具（如 Python 中的 pandas 库）加载乳腺癌数据集。检查数据集的格式，确保数据能够正确读取，例如查看数据的行数和列数，确认特征和目标变量（良性或恶性肿瘤）的对应关系。

2. 数据清洗

检查数据集中是否存在缺失值。如果存在缺失值，根据具体情况选择填充方法，如使用均值、中位数或众数填充，或者删除含有缺失值的行。例如，对于数值型特征，可以使用该特征列的均值进行填充。

检查数据集中是否存在异常值。异常值可能会影响模型的训练效果。可以通过绘制箱线图等方法来识别异常值。对于异常值，可以考虑删除、修正或者使用一些鲁棒的模型来处理。

3. 特征编码

如果数据集中存在分类特征（如某些非数值型的特征），需要对其进行编码。对于二分类特征，可以使用 0-1 编码；对于多分类特征，可以使用独热编码（One-Hot Encoding）。

4. 数据标准化/归一化（可选）

对于一些距离相关的模型（如 KNN、SVM）和一些基于梯度的模型（如 XGBoost、LightGBM），数据的标准化或归一化是很重要的。可以选择合适的方法，如 Z-分数标准化（将数据转换为均值为 0，标准差为 1 的分布）或 Min-Max 归一化（将数据缩放到[0,1]区间）。

5. 划分训练集和测试集

使用合适的方法（如随机划分）将数据集划分为训练集和测试集。一般可以按照 70%-30%或 80%-20%的比例划分。确保在划分过程中保持目标变量（良性或恶性肿瘤）的分布比例在训练集和测试集中大致相同，可以使用分层抽样的方法。

**2、模型训练与评估阶段**

1. 模型训练

逻辑回归

调整模型参数，如正则化参数（L1 或 L2 正则化）。使用训练集数据训练模型，通过优化损失函数（如交叉熵损失）来学习特征与目标变量之间的关系。

决策树

设置决策树的深度、最小样本分割数等参数。训练决策树模型，通过递归划分特征空间来构建树结构，从而对肿瘤的良性或恶性进行分类。

随机森林

确定森林中树的数量、每棵树的最大深度等参数。随机森林通过构建多棵决策树，并在训练过程中引入样本抽样和特征抽样来提高模型的泛化能力。使用训练集数据训练随机森林模型。

KNN（K-最近邻）

选择合适的 K 值（邻居数）。KNN 模型通过计算测试样本与训练样本之间的距离（如欧几里得距离），找出距离最近的 K 个样本，并根据这些邻居的类别进行投票来预测测试样本的类别。

NB（朴素贝叶斯）

选择合适的概率分布假设（如高斯分布、多项式分布等）。朴素贝叶斯模型基于贝叶斯定理和特征条件独立假设来计算目标变量的概率，从而进行分类。

SVM（支持向量机）

选择合适的核函数（如线性核、多项式核、径向基核等）和惩罚参数 。SVM 模型通过寻找最优分割超平面来最大化不同类别样本之间的间隔。

XGBoost

调整模型参数，如树的深度、学习率、树的数量等。XGBoost 是一种基于梯度提升的集成学习方法，通过迭代地构建多棵决策树来提高模型的预测性能。

LightGBM

设置参数，如树的深度、学习率、树的数量等。LightGBM 和 XGBoost 类似，但它的特点是采用了基于直方图的算法，能够更高效地处理大规模数据。

2. 模型评估

使用测试集数据对每个模型进行评估。评估指标可以选择准确率（正确分类的样本数占总样本数的比例）、召回率（真正例被预测为正的比例）、精确率（被预测为正的样本中实际为正的比例）、F1-分数（精确率和召回率的调和平均数）等。对于二分类问题，还可以绘制 ROC 曲线（受试者工作特征曲线）和计算 AUC（曲线下面积）值来评估模型的性能。

**3、降维可视化阶段**

1. 降维方法选择与应用

PCA（主成分分析）

确定降维后的维度数量。PCA 通过将原始数据投影到新的坐标轴（主成分）上，使数据在新坐标轴上的方差最大化，从而实现降维。对训练集和测试集数据分别进行 PCA 降维，以便后续可视化。

NCA（邻域成分分析）

设置参数，如降维后的维度数量。NCA 是一种监督学习的降维方法，它通过学习一个线性变换来最大化样本与其同类邻居之间的相似度。

UMAP（统一流形近似和投影）

调整参数，如降维后的维度数量、邻居数量等。UMAP 是一种基于流形学习的降维方法，能够较好地保留数据的局部结构和全局结构。

t-SNE（t-分布随机邻域嵌入）

设置参数，如降维后的维度数量（一般为 2 或 3）、困惑度（Perplexity）等。t-SNE 通过计算样本之间的相似度，并将其映射到低维空间中，能够有效地将相似的样本聚集在一起。

2. 可视化

使用降维后的数据进行可视化。可以使用散点图来展示不同类别（良性或恶性肿瘤）的样本在低维空间中的分布情况。例如，在二维平面上，用不同的颜色或形状标记良性肿瘤和恶性肿瘤的样本，观察它们之间是否存在明显的分隔区域。通过可视化结果可以直观地了解模型在低维空间中对数据的分类效果，以及不同降维方法对数据结构的保留程度。

**4、结果分析与总结阶段**

1. 模型性能比较

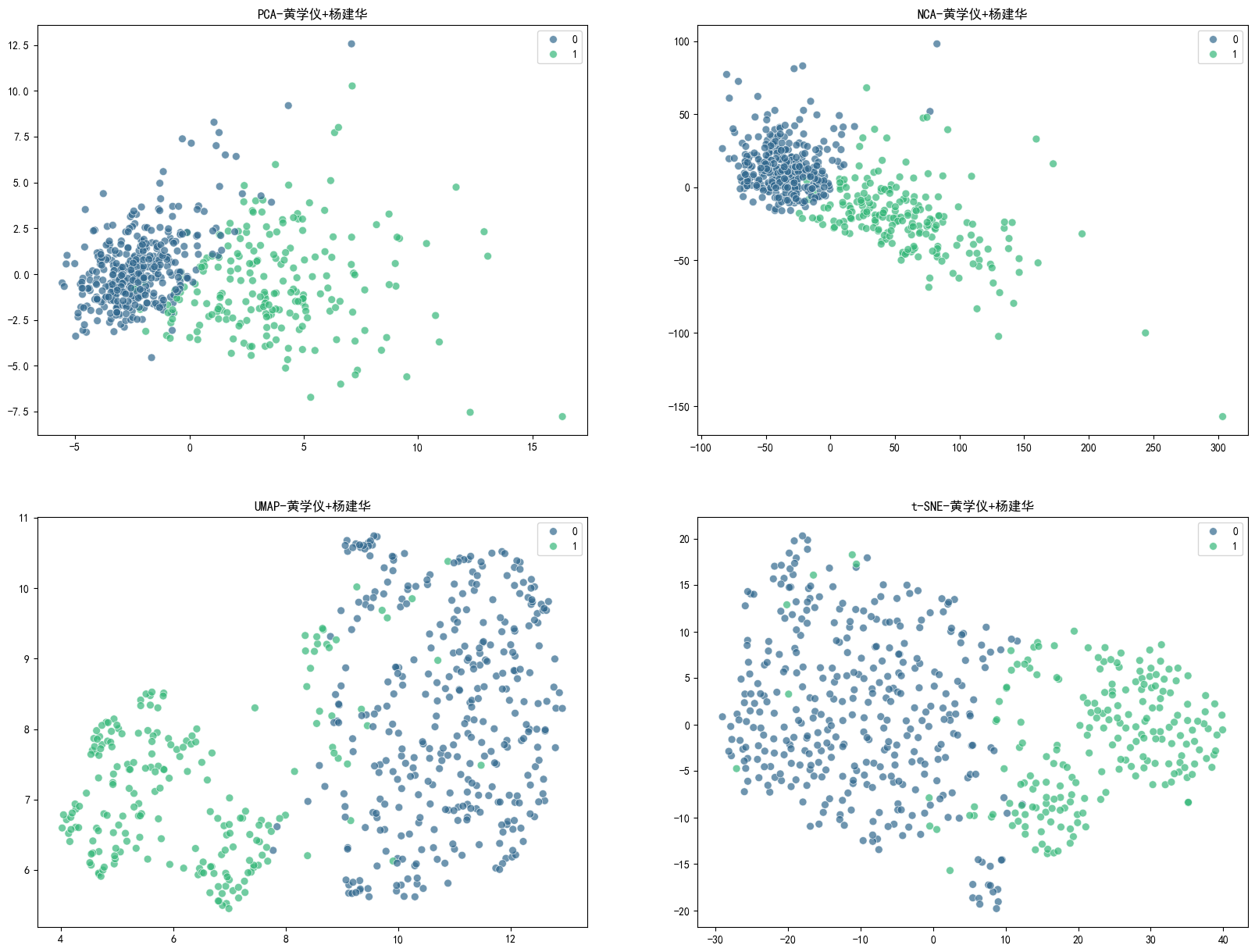
根据模型评估阶段的指标（如准确率、F1-分数、AUC 等），比较不同机器学习模型在乳腺癌数据集上的性能。找出表现最好和最差的模型，分析其原因。例如，某些模型可能对数据中的噪声更敏感，或者某些模型的假设与数据的实际分布不匹配。

2. 降维方法效果分析

观察不同降维方法在可视化结果中的表现。比较它们在保留数据结构（如局部结构和全局结构）方面的差异。例如，PCA 更适合于线性可分的数据，而 UMAP 和 t-SNE 在处理复杂的非线性结构数据时可能表现更好。分析降维后的数据是否能够更好地展示模型的分类效果。

五．实验结果分析、对比和讨论

降维可视化



**机器学习模型效果评估**

Method Model Mean Accuracy Std Dev

0 PCA Logistic Regression 0.942244 0.022565

0 PCA Decision Tree 0.899359 0.043689

0 PCA Random Forest 0.927051 0.032744

0 PCA KNN 0.919423 0.031866

0 PCA Naive Bayes 0.909487 0.020281

0 PCA SVM 0.934679 0.016596

0 PCA XGBoost 0.919423 0.046405

0 PCA LightGBM 0.911923 0.034656

0 NCA Logistic Regression 0.974872 0.015813

0 NCA Decision Tree 0.982372 0.016271

0 NCA Random Forest 0.984936 0.012301

0 NCA KNN 0.977372 0.017520

0 NCA Naive Bayes 0.964808 0.025549

0 NCA SVM 0.979872 0.021974

0 NCA XGBoost 0.982372 0.016271

0 NCA LightGBM 0.979808 0.024930

0 UMAP Logistic Regression 0.959744 0.030090

0 UMAP Decision Tree 0.927051 0.046874

0 UMAP Random Forest 0.957244 0.040440

0 UMAP KNN 0.959679 0.023269

0 UMAP Naive Bayes 0.947179 0.030866

0 UMAP SVM 0.944679 0.031541

0 UMAP XGBoost 0.944551 0.041912

0 UMAP LightGBM 0.952115 0.042755

0 t-SNE Logistic Regression 0.954679 0.027409

0 t-SNE Decision Tree 0.947179 0.028769

0 t-SNE Random Forest 0.954744 0.024684

0 t-SNE KNN 0.964808 0.022972

0 t-SNE Naive Bayes 0.949615 0.032339

0 t-SNE SVM 0.954744 0.018784

0 t-SNE XGBoost 0.959744 0.025727

0 t-SNE LightGBM 0.959744 0.037575

0 Original Logistic Regression 0.978885 0.021953

0 Original Decision Tree 0.919204 0.035189

0 Original Random Forest 0.959555 0.022314

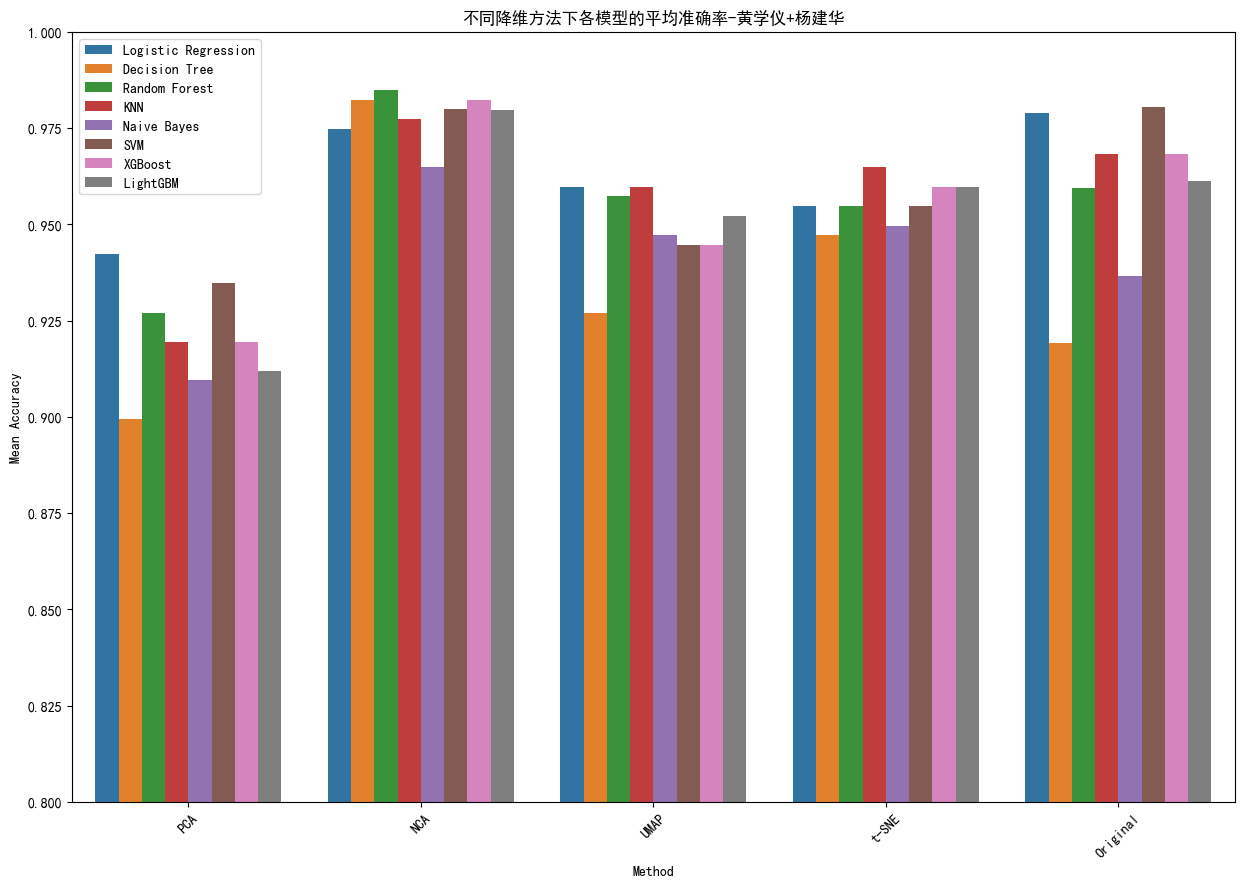
0 Original KNN 0.968296 0.028244

0 Original Naive Bayes 0.936654 0.027697

0 Original SVM 0.980576 0.021714

0 Original XGBoost 0.968390 0.025767

0 Original LightGBM 0.961341 0.024553



**实验结果讨论：**

在该场景中，可能由于数据集规模较小、特征数较少，以及特征重要性和选择等问题，决策树的性能比随机森林更好。对于小规模数据集，单一决策树可能表现得更好，因为其结构简单明了，能够更容易地捕捉数据中的模式。相比之下，随机森林可能会出现过拟合或欠拟合的情况。这是因为随机森林会训练多棵树，而每棵树的表现可能受到数据集规模和复杂性的限制。此外，随机森林在训练过程中会随机选择特征进行分裂，这在某些情况下可能导致忽略了某些重要特征。而单一决策树则会使用所有特征进行分裂，因此在某些特征上可能会表现得更好。

六．其他

无

七．对本门课的感想、意见和建议

本学期的机器学习课程给我留下了深刻的印象。在完成大作业的过程中，我不仅深入理解了机器学习的核心概念和技术，还亲身体验了数据处理和模型构建的实际操作。通过在 Kaggle 平台使用数据集进行实战演练，我将理论知识成功应用于实际问题，感受到了数据分析与建模带来的成就感。尤其是在尝试不同的机器学习模型和降维方法时，每一步的探索和验证都加深了我对机器学习的理解。通过对这些模型的深入对比和代码编写，我对各个模型的特点有了更清晰的认识。同时，复现和应用新技术的过程让我深刻认识到机器学习领域的创新是无止境的，也激励我在未来的学习和研究中不断探索新方法。

总体来看，课程内容设置合理，涵盖了机器学习的基础知识和技术，既有广度又有深度，能够满足不同层次学生的需求。丰富的实践操作机会让我通过实际对比多种模型，极大地提升了动手能力和解决实际问题的能力。课程对机器学习的关键技术进行了深入探讨，并借助多种媒体方式进行展示，开阔了我们的视野，激发了我们对前沿技术的兴趣。课程难度适中，既有基础知识的讲解，也有高难度问题的探讨，让每位学生都能有所收获。

为了进一步提升课程效果，我建议增加更多不同行业的实际案例分析，并持续跟进最新的发展动向，以进一步激发学生的学习兴趣。这门课程不仅让我掌握了扎实的理论基础，还提升了我的实践能力和创新思维。在未来的学习和研究中，我将继续努力，应用所学知识解决实际问题，追求更高目标。

感谢姜烨老师的悉心教导和指导！

|  |  |  |  |
| --- | --- | --- | --- |
| **“机器学习-大作业”评分细则** | | | |
| **成绩等级** | **具体表现** | **教师评分** | |
| 优秀（100-90] | 报告撰写优秀，题目本身难度大，工作量饱满；Notebook代码清晰准确，实验与可视化优秀 | □ |  |
| 良好（90-80] | 报告撰写良好，题目本身难度较大，工作量较多；Notebook代码较好，实验与可视化良好 | □ |  |
| 中等（80-70] | 报告撰写中等，题目本身难度一般，工作量达标；Notebook代码一般，实验与可视化达标 | □ |  |
| 及格（70-60] | 能完成基本要求，工作量较少，非Notebook代码 | □ |  |
| 不及格（<60） | 未达最低要求或抄袭线上线下资源 | □ |  |

教师签名：