**分类算法在3D-Moon数据集上的性能比较实验报告**

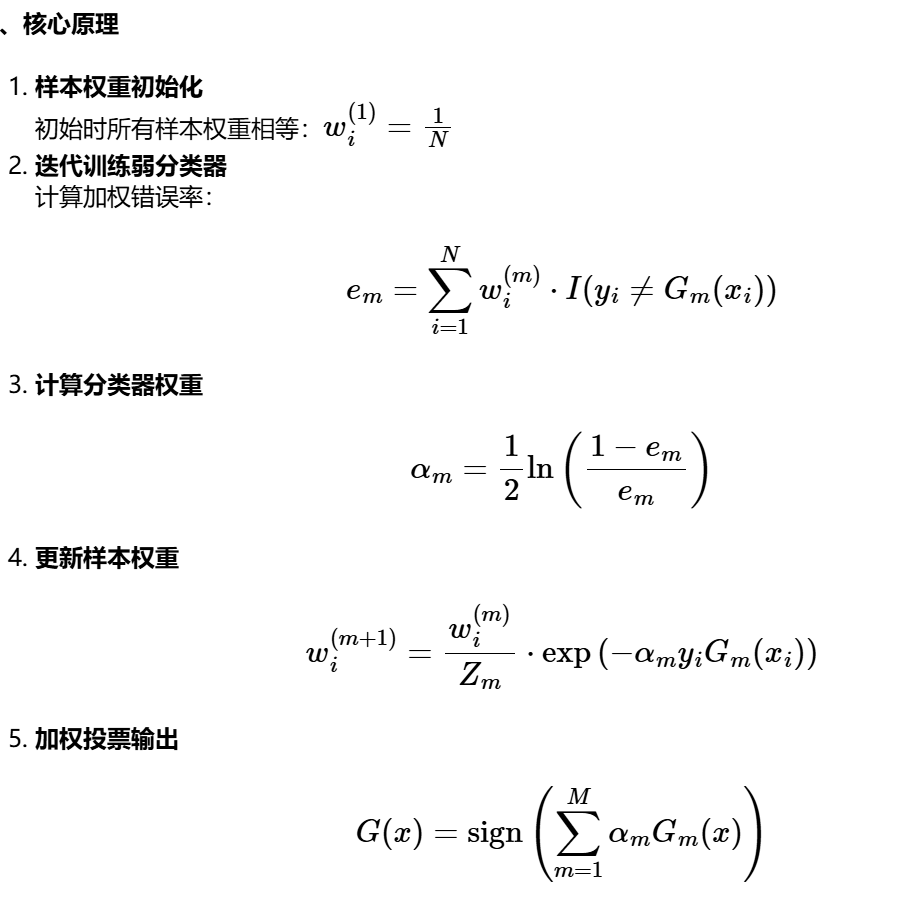
1. **摘要**

本次作业报告应用Decision Trees、AdaBoost + DecisionTrees和SVM对3D数据集进行分类，对比不同算法的性能差异。通过标准化数据集、可视化分类边界和量化评估指标，验证了RBF核SVM在非线性可分数据中的最优表现。

1. **方法原理**
2. **决策树（Decision Trees）**：是一种树形结构的监督学习算法，通过递归划分特征空间实现分类或回归。其核心组成包括：
3. 根节点：代表初始数据集
4. 内部节点：对应特征测试，产生分支决策
5. 叶节点：存储最终预测结果（类别或数值）

基于信息增益递归划分特征空间，作业中设置max\_depth=5防止过拟合。优势在于直观解释性强，无需数据标准化处理，但是对噪声敏感容易过拟合。

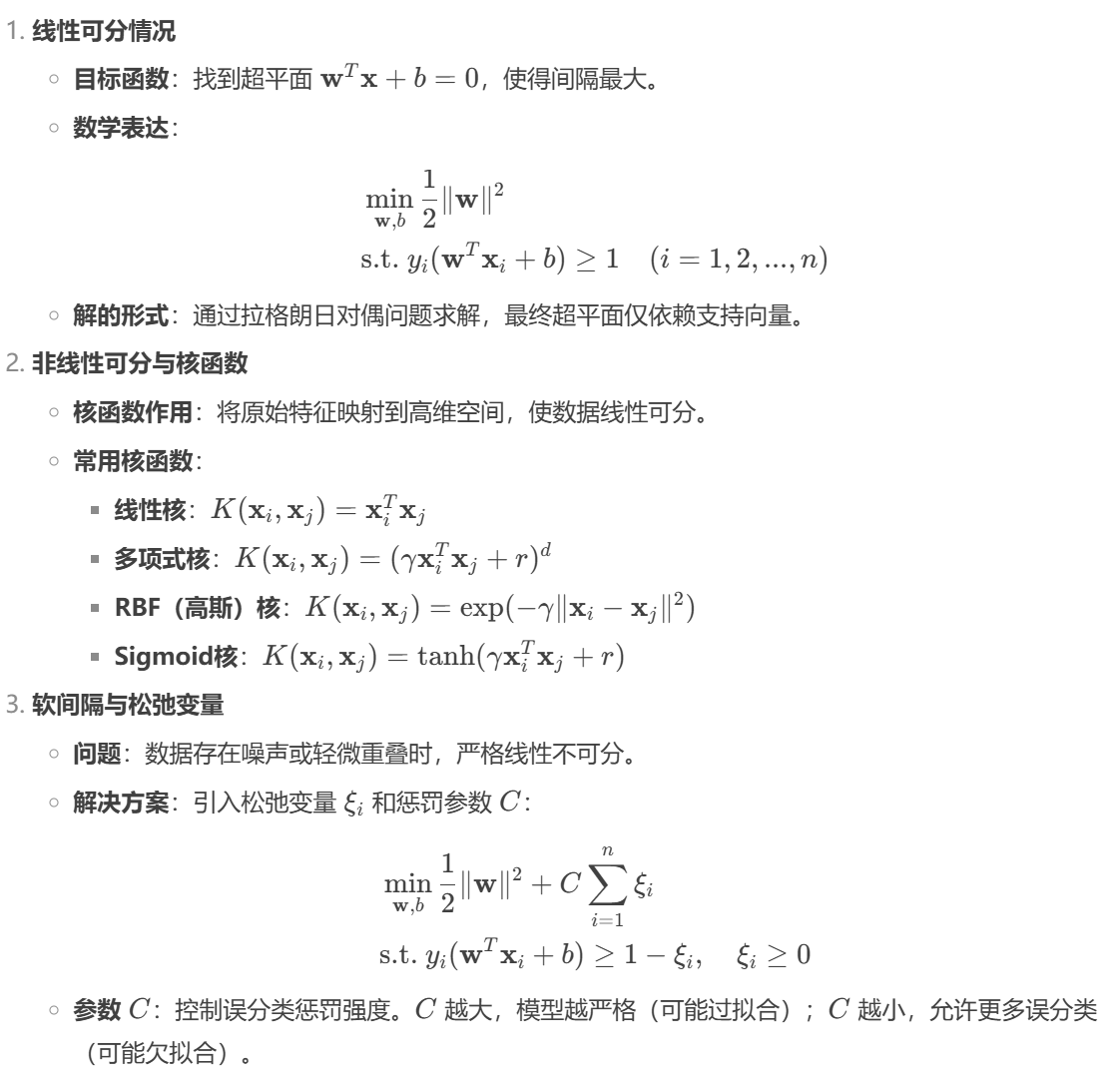
1. **AdaBoost（Adaptive Boosting）**：是一种集成学习算法，通过组合多个弱分类器（如决策树桩）构建强分类器。其核心思想是**迭代调整样本权重**与**分类器权重**，逐步聚焦于难以正确分类的样本，最终通过加权投票提升整体性能。



本次作业集成50个弱分类器（决策树max\_depth=5），通过动态调整样本权重提升性能。迭代过程中，错误分类样本权重增加，增强模型鲁棒性。

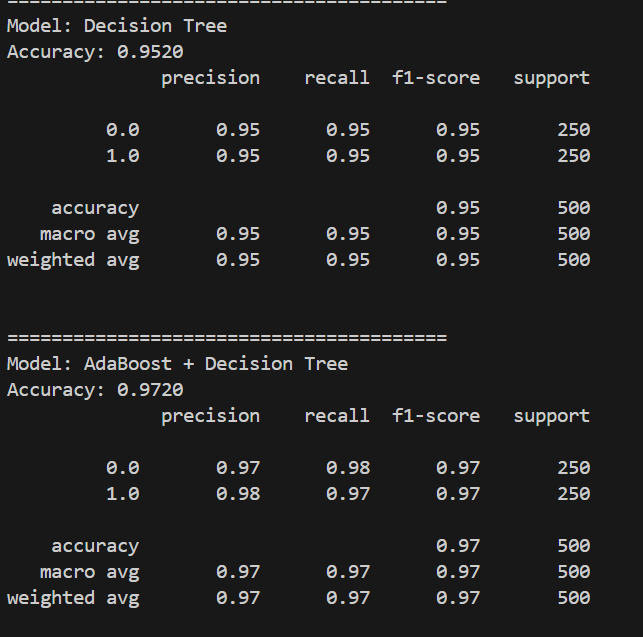
1. **支持向量机（SVM）**：是一种监督学习算法，主要用于分类和回归任务，核心思想是寻找一个最优超平面，将不同类别的数据分开，同时最大化两类数据到超平面的最小距离（称为间隔）。其关键特点包括：
2. 间隔最大化：通过几何间隔定义分类边界，提升泛化能力。
3. 核技巧（Kernel Trick）：将数据映射到高维空间，解决非线性可分问题。
4. 支持向量：超平面的位置仅由距离最近的样本点（即支持向量）决定。

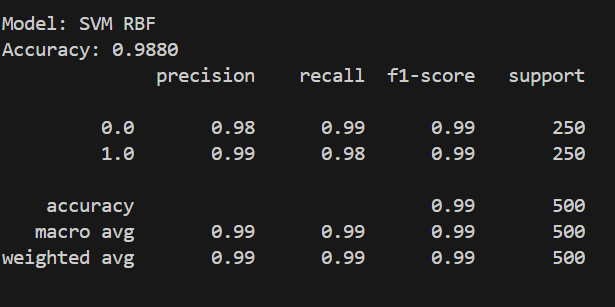
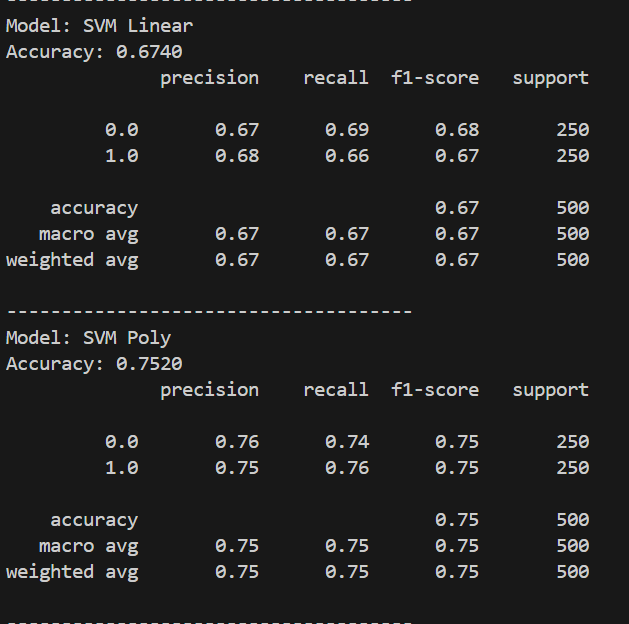
核心原理

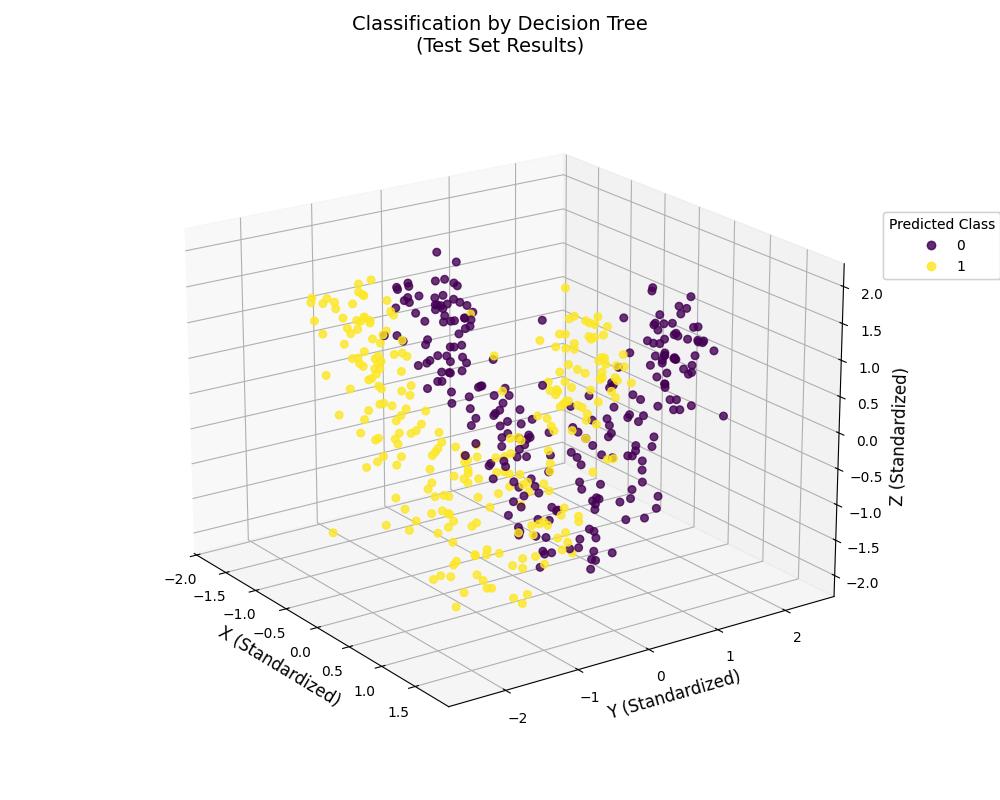


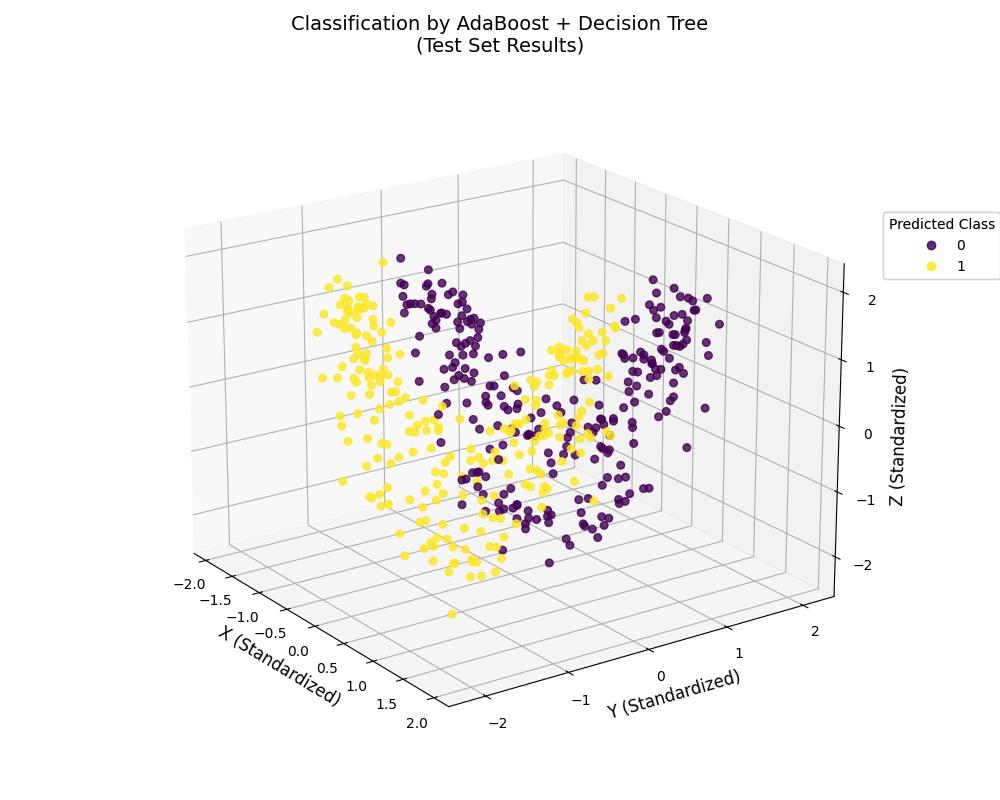
1. **实验结果**

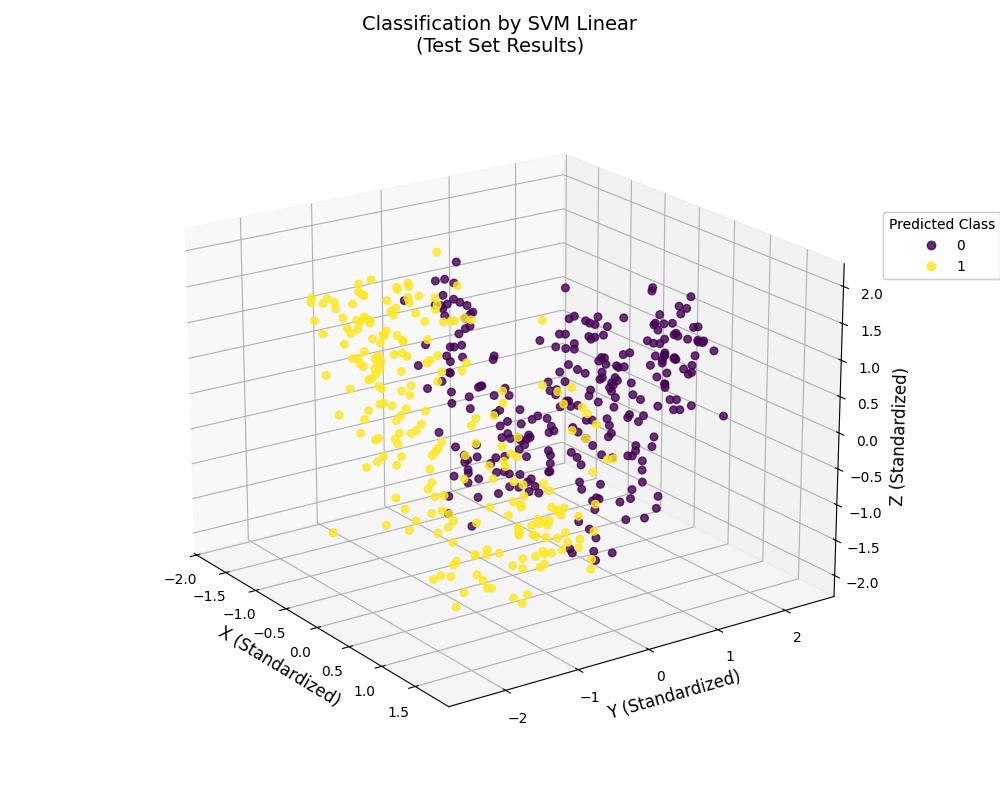
运行程序得到如下结果：

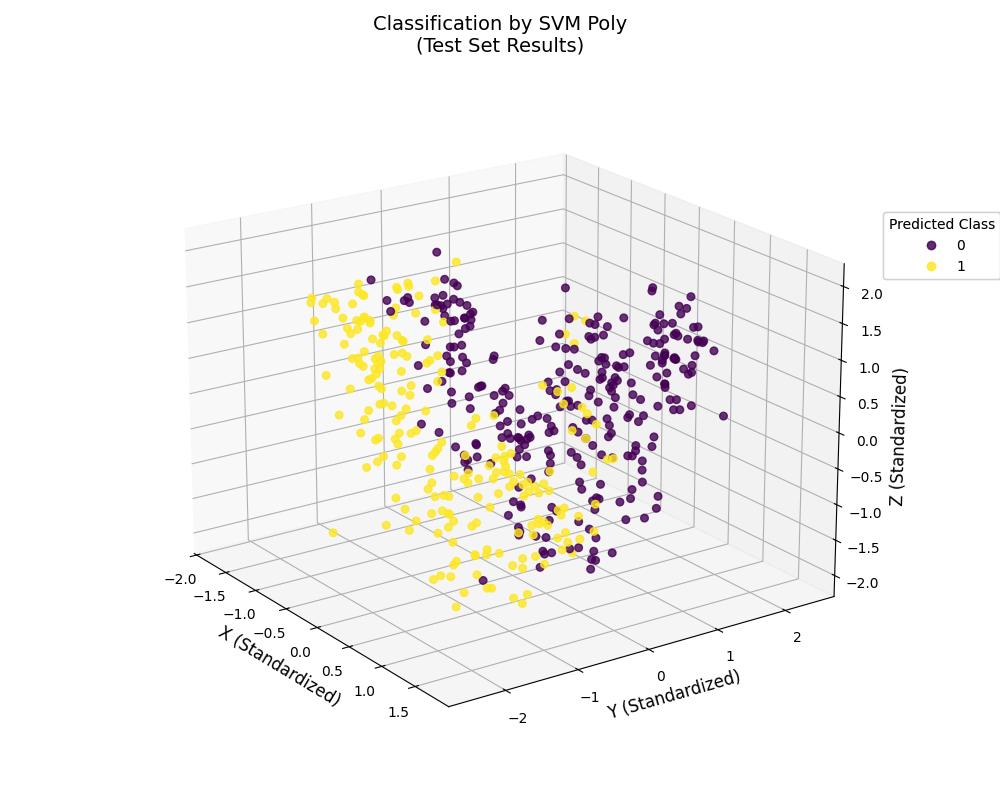


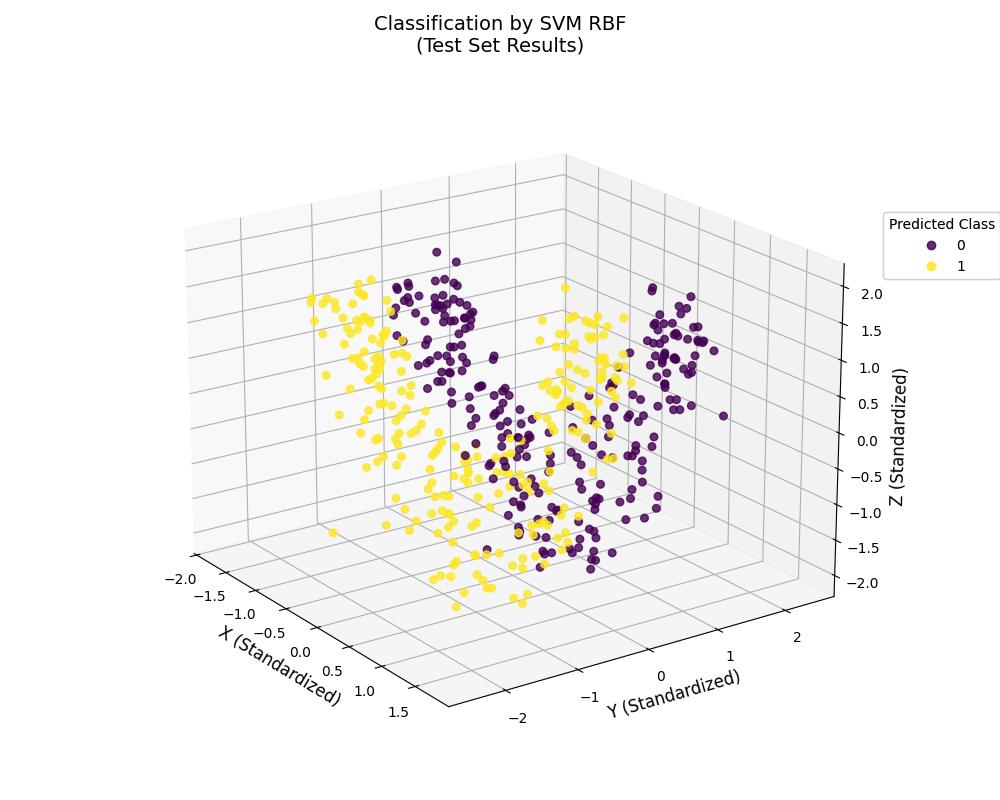












1. **性能差异分析**
2. 决策树：

表现：准确率较高（95.2%）。由于数据中存在噪声，决策树容易过拟合训练数据中的局部模式，导致泛化能力下降。

原因：尽管数据在z轴上有明显的分界（正类z值多为正，负类多为负），但噪声导致决策树需要复杂的分支，增加了错误分割的风险。

1. AdaBoost + 决策树：

表现：准确率显著提升至97.2%。通过集成多个弱分类器（决策树桩），AdaBoost减少了方差，提升了鲁棒性。

原因：弱分类器的组合能够有效捕捉数据的全局结构，避免单棵决策树对噪声的敏感。

1. SVM：

线性核：准确率67.4%。由于数据在3D空间中接近线性可分（z轴提供关键区分），线性核能部分分离类别，但噪声和非线性残余结构限制了性能。

多项式核：准确率75.2%。3次多项式核捕捉了部分非线性关系，但参数未调优（如degree可能不匹配真实数据分布），表现中等。

RBF核：准确率最高（98.8%）。RBF核通过非线性映射处理复杂边界，完美适应数据的螺旋结构，噪声容忍度高。

**五、结论**

1. RBF核SVM表现最优，因其能建模复杂的非线性决策边界。
2. AdaBoost通过集成提升了决策树的泛化能力，表现比决策树更优。
3. 线性SVM和多项式核SVM受限于模型假设，未能充分捕捉数据结构，表现较差。