雲南大學
YUNNAN UNIVERSITY

智能科学与技术

机器学习实验（2025 春）课程论文

泰坦尼克号生存预测

姓名 杨德金　　学号 20221060232

摘　要

本次实验旨在通过机器学习方法预测泰坦尼克号乘客的生还情况。首先，对原始数据集进行了全面的数据清洗、特征工程等预处理步骤，以提取有效信息并构建合适的特征集。随后，分别采用了随机森林（Random Forest）、支持向量机（SVM）、K 近邻（KNN）以及多层感知机（MLP）四种经典的机器学习模型进行训练和预测。通过在测试集上评估各模型的性能指标，并进行横向对比分析，最终筛选出在泰坦尼克号生还预测任务中表现最优的模型。

目录

插图

表格

1 引言

泰坦尼克号的沉没是历史上最著名的海难之一，而其中留下来的丰富乘客数据，使其成为了数据科学和机器学习领域中一个经典的预测研究案例。通过分析乘客的年龄、性别、船票价格等特征，研究其幸存的概率，不仅具有重要的历史回顾价值，也为我们理解灾难中生存因素的复杂提供了学术视角。站在更广阔的视角，此类生还预测的研究在现实世界中具有潜在的应用价值。

本次实验的核心任务正是基于泰坦尼克号的公开数据集，运用机器学习的技术来预测乘客的生还情况。实验将首先对原始数据集进行细致的探索和分析以及预处理，这其中包括异常数值处理、缺失值填充和主要特征提取，以构建高质量的训练数据。随后，将采用多种经典的机器学习模型，如随机森林、支持向量机（SVM）、K 邻近（KNN）以及多层感知机（MLP），对处理后的数据进行训练和测试。通过比较这些数据的预测准确率、召回率等关键指标上的表现，旨在筛选出最适合此类分类任务的模型，根据实验结果对影响生还的关键因素进行分析。

2 数据分析与处理

2.1 数据集描述

本实验采用的数据集来自 Kaggle 竞赛中经典的泰坦尼克号乘客信息数据，该数据集常被用于机器学习领域的分类问题实践，特别是生还预测。数据集包含了影响乘客在 1912 年泰坦尼克号沉船事件中生还与否的多种潜在因素。下面将对数据集中各变量（特征）进行详细描述：

- `survival` **(生还情况)**：这是本实验的目标变量。它是一个二元分类变量，其中 `0` 代表未生还，`1` 代表生还。
- `pclass` **(船票等级)**：代表乘客所持船票的等级。这是一个有序分类变量，包含三个等级：`1` 代表一等舱，`2` 代表二等舱，`3` 代表三等舱。通常认为船票等级与乘客的社会经济地位相关，并可能影响其生还几率。
- `sex` **(性别)**：乘客的性别，为一个二元分类变量（通常表示为男性/女性）。历史数据表明，性别是影响生还的重要因素之一。
- `Age` **(年龄)**：乘客的年龄，以年为单位，是一个数值型连续变量。年龄在灾难中的作用复杂，不同年龄段的生还情况可能存在差异。
- `sibsp` **(船上兄弟姐妹/配偶数量)**：乘客在泰坦尼克号上的兄弟姐妹或配偶的数量。这是一个数值型离散变量，反映了乘客是否与近亲同行。
- `parch` **(船上父母/子女数量)**：乘客在泰坦尼克号上的父母或子女的数量。这也是一个数值型离散变量，与 `sibsp` 共同构成了乘客的家庭规模信息。
- `ticket` **(船票号码)**：乘客的船票号码。这是一个字符串类型的分类变量，通常具有较高的唯一性，可能包含一些隐含信息，但直接使用较为困难，需要进一步处理。
- `fare` **(乘客票价)**：乘客支付的船票费用。这是一个数值型连续变量，通常与船票等级（`pclass`）和客舱位置（`cabin`）相关。
- `cabin` **(客舱号码)**：乘客所在的客舱号码。这是一个字符串类型的分类变量。该特征通常存在大量缺失值，并且其格式可能包含船舱区域信息。

- **embarked（登船港口）**：乘客登船的港口。这是一个分类变量，包含三个主要港口：`C` 代表瑟堡 (Cherbourg)，`Q` 代表皇后镇 (Queenstown)，`S` 代表南安普敦 (Southampton)。

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

图 1: 数据样例

2.2 数据预处理

2.2.1 划分训练集与测试集

为了确保模型的训练和测试具有良好的泛化能力，同时验证模型能够在不同的数据集上保持一致性，我们将数据集按照 8:2 的比例划分为训练集和测试集。其中，划分过程中保证目标特征 `Survived` 以及主要特征 `Pclass` 和 `Sex` 的分布比例在训练集和测试集中一致（即采用分层抽样的方法进行划分）。

在完成数据划分后，为了验证特征 `Survived`、`Pclass` 和 `Sex` 在训练集和测试集中的分布一致性，绘制了它们的分布直方图，三种不同的颜色代表三个特征，如图 2 所示。

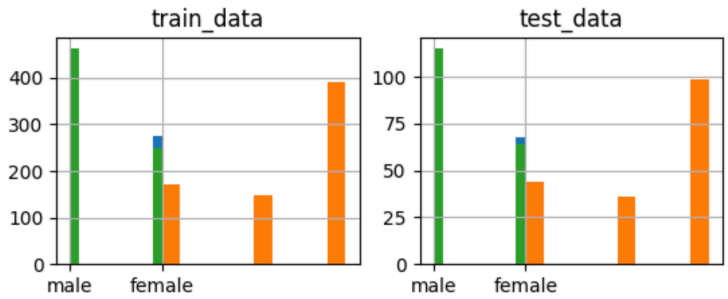

图 2: 关键特征分布

2.2.2 关键特征选取

为了分析数据中实数字类型特征与目标变量 `Survived` 的相关性，我们计算了所有实数类型特征之间的皮尔逊相关系数 (Pearson Correlation Coefficient)，并绘制了相关性矩阵热力图，如图3所示。通过观察热力图，可以直观地了解各个特征之间的相关性强弱，从而识别出与 `Survived` 最相关的特征。

分析矩阵热力图之后发现，与目标变量 `Survived` 相关性高的特征包括 `Age`、`Pclass`（船票等级）、`SibSp`（船上兄弟姐妹/配偶数量）、`Parch`（船上父母/子女数量）和 `Fare`（票价）。此外，`PassengerId` 也表现出一定的相关性，但是其本质上只是用于标识乘客的序列号，不包含与任何与乘客生还可能性的信息，并且可能引入一些潜在的序列偏差，进而影响模型的泛化能力。因此，在

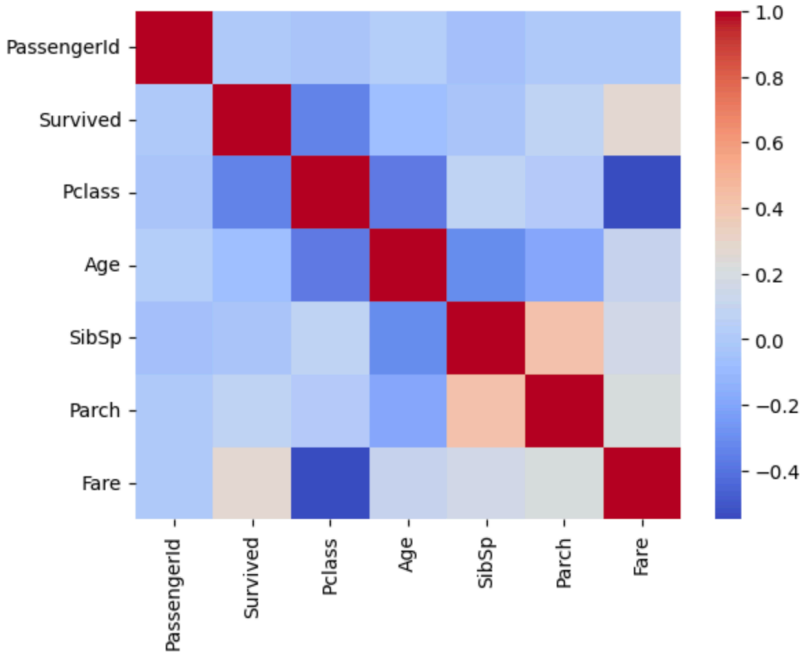

图 3: 相关性热力图

接下来的模型训练中，我们将忽略 Survived 特征，以免模型对无意义的特征进行学习，从而提高实验结果的可靠性和稳定性。

除了实数类型的特征外，我们还将对数据集中剩余的字符串特征 Sex、Ticket、Cabin 和 Embarked 进行了分析与处理。性别 (Sex) 是灾难中影响生还的重要特征之一，已有的研究和统计显示女性乘客的生还率显著高于男性，因此将此特征作为分类变量进行保留，并在后续模型中进行数值化处理。船票号码 (Ticket) 是乘客的标识符之一，其值为无序字符串，很难提取与乘客生还的相关信息，故应舍去此特征。此外还有一个特征客舱号码 (Cabin)，虽然包含乘客的空间位置信息，但是该特征存在大量的缺失值，其值格式也比较复杂，不易处理，因此也选择丢弃。登船港口 (Embarked) 是一个分类变量，间接反映乘客的社会经济地位或空间位置。该特征仅有少量缺失值，且取值范围较小（C 表示瑟堡港，Q 表示皇后镇，S 表示南安普敦），具有一定的预测价值，因此将其保留并进行数值化编码。

通过以上的数据特征分析与提取，我们最终选取了 Age、Pclass、SibSp、Parch、Fare、Sex 和 Embarked 特征，这些特征涵盖了乘客的基本人口统计信息、家庭结构，以及票价和船舱登录地点。这些特征不仅与目标变量 Survived 具有高度的相关性，同时也具有明确的解释意义，能够从多个维度反映乘客在灾难中的生还可能性。

2.2.3 缺失值处理策略

在数据预处理中，缺失值的处理是确保数据完整性和模型稳定性的关键步骤。常见的缺失值处理策略包括删除缺失样本、基于统计量（如均值、众数或中位数）填充、插值法填充，以及基于模型预测的填补等。根据特征的类型与数据分布特性，我们选择了适合的缺失值处理方法。

对于实数类型的特征（如 Age），我们采用均值填充（mean）的策略。这种方法能够有效保留数据的整体分布特性，避免因缺失值导致的样本丢失，同时减少模型训练中的偏差。此外，均值填充在数据分布较为均匀且无显著偏态的情况下，可以确保数据分布的连续性和一致性，从而对模型的性能起到积极作用。

对于字符串类型的特征（如 Embarked），我们采用缺失值标记法（NaN 填充）。该方法通过保留缺失值信息，避免在填补过程中引入额外的偏差。具体来说，对于分类变量，缺失值标记可以视为一个独立类别，从而帮助模型学习到缺失值可能隐含的潜在模式。此外，由于 Embarked 特征的缺失比例较低，采用缺失值标记法对整体数据分布的影响可以忽略不计，同时保留了特征的完整性和解释性。

2.2.4 字符串特征的数字化处理

在数据预处理中，为了使模型能够有效地利用字符串类型特征，需要将其转换为数值形式的特征表示。在实验中，我们采用了 One-Hot 编码 (One-Hot Encoding) 方法对字符串特征进行数值化处理。One-Hot 编码是一种将分类变量转换为数值表示的常用方法，其基本原理是为每个类别创建一个独立的二进制变量（列），用 1 表示该类别存在，用 0 表示该类别不存在。通过采用 One-Hot 编码，我们成功地将字符串类型的特征数值化，使其能够被机器学习模型直接利用。同时，该方法在保留特征原始信息的基础上，避免了分类变量的大小关系假设，为后续模型训练提供了高质量的输入特征。

- **Sex 特征**：Sex 是一个二分类变量，仅包含 male 和 female 两个类别。在 One-Hot 编码后，该特征被转换为两个二进制变量 Sex_male 和 Sex_female，用于表示性别信息。
- **Embarked 特征**：Embarked 是一个多分类变量，包含三个类别值：C（瑟堡港）、Q（皇后镇）和 S（南安普敦）。通过 One-Hot 编码，该特征被转换为三列二进制变量 Embarked_C、Embarked_Q 和 Embarked_S。此外，为了处理少量的缺失值（NaN），我们将缺失值单独归为一类，并生成一个额外的二进制变量 Embarked_NaN，以保留缺失信息对模型潜在的影响。

3 实验模型

3.1 实验面临的挑战

泰坦尼克号生还预测问题是一个经典的二分类问题，其核心挑战在于类别不平衡和特征多样性。数据中幸存者和未幸存者的数量分布不均，可能导致模型更倾向于预测多数类。此外，数据包含数值型、类别型和文本型等多种特征，这要求模型能够有效处理异构数据并捕获潜在的复杂模式。由于生还概率受多个特征的交互影响，模型还需要具备处理非线性决策边界的能力。因此，一个理想的模型应具备良好的鲁棒性，能够在处理噪声和缺失值时保持稳定；同时具备强泛化能力，能够避免过拟合；在某些场景下，还需要提供一定的可解释性，以便理解特征对预测结果的影响。

基于上述需求，本实验选择了四种具有代表性的机器学习模型进行实验和对比分析。

3.2 随机森林

随机森林作为 Bagging 集成学习方法的典型代表，通过构建多个决策树并结合其预测结果来提高模型的泛化能力和稳定性。该算法的核心思想是利用 Bootstrap 抽样和随机特征选择来增加基学习器之间的差异性，从而降低模型方差并提高预测精度。

随机森林作为一种集成学习方法，通过构建大量决策树并对其结果进行集成，有效提升了模型的泛化能力和鲁棒性。这种特性使得随机森林能够很好地应对数据中的噪声和异常值，尤其适用于处理包含缺失信息的历史数据集。在泰坦尼克号数据集中，乘客信息往往存在缺失或不完整的情况，而随机森林凭借其随机采样和特征选择机制，能够最大限度地利用已有信息，减少数据缺陷带来的影响。

Algorithm 1: 随机森林算法

Input: 训练集 D, 树的数量 T, 特征子集大小 m

Output: 随机森林模型 $\{h_1, h_2, \ldots, h_T\}$

for $t = 1$ **to** T **do**
  通过 Bootstrap 抽样生成训练子集 D_t;
  初始化决策树 h_t;
  Function `BuildTree`(D_t, *node*)**:**
    if 停止条件满足 **then**
      设置叶节点标签;
      return;
    end
    随机选择 m 个特征构成候选集 F_{cand};
    从 F_{cand} 中选择最优分割特征和阈值;
    根据分割条件划分数据;
    递归构建左右子树;
  $h_t \leftarrow$ `BuildTree`(D_t, *root*);
end
return $\{h_1, h_2, \ldots, h_T\}$;

随机森林算法在处理特征类型方面具有天然的优势。它不仅能够处理乘客年龄、票价等数值型特征，还能同时处理性别、船舱等级等类别型特征，为模型构建提供了极大的灵活性。更为重要的是，随机森林能够对各特征的重要性进行评估，帮助我们直观地理解不同因素对生还概率的影响。这一特性为灾难分析和应急管理领域的研究与决策提供了有力的理论支持和实践参考，有助于制订更为科学和有效的应急响应策略。

3.3 SVM

支持向量机（Support Vector Machine, SVM）基于统计学习理论中的结构风险最小化原则，通过寻找最优分离超平面来实现分类。该算法的核心思想是最大化不同类别样本之间的间隔，从而获得良好的泛化性能。

对于线性可分情况，SVM 的目标是找到分离超平面 $\mathbf{w}^T\mathbf{x} + b = 0$，使得间隔 $\frac{2}{\|\mathbf{w}\|}$ 最大化。这等价于求解如下优化问题：

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2$$
$$\text{s.t. } y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1, \quad i = 1, 2, \ldots, n$$

通过拉格朗日乘数法，可将上述问题转化为对偶问题：

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j$$
$$\text{s.t. } \sum_{i=1}^{n} \alpha_i y_i = 0, \quad \alpha_i \geq 0$$

对于非线性可分情况，引入松弛变量 ξ_i 和核函数 $K(\mathbf{x}_i, \mathbf{x}_j)$，目标函数变为：

$$\min_{\mathbf{w},b,\boldsymbol{\xi}} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i$$

相应的对偶问题为：

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j K(\mathbf{x}_i,\mathbf{x}_j)$$

最终的决策函数为：

$$f(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^{n}\alpha_i y_i K(\mathbf{x}_i,\mathbf{x}) + b\right)$$

泰坦尼克号生还预测数据集样本数量有限且数据分布未知，在这样的情况下，SVM 基于间隔最大化

Algorithm 2: SVM 算法（SMO 实现）

Input: 训练集 D, 惩罚参数 C, 核函数 K, 容忍度 ε

Output: SVM 模型参数 $\boldsymbol{\alpha}, b$

初始化 $\boldsymbol{\alpha} = \mathbf{0}, b = 0$;
repeat
  $numChanged = 0$;
  for $i = 1$ **to** n **do**
    计算 $E_i = f(\mathbf{x}_i) - y_i$;
    if 违反 KKT 条件 **then**
      启发式选择第二个变量 α_j;
      计算 $E_j = f(\mathbf{x}_j) - y_j$;
      计算边界 L, H;
      计算 $\eta = K(\mathbf{x}_i,\mathbf{x}_i) + K(\mathbf{x}_j,\mathbf{x}_j) - 2K(\mathbf{x}_i,\mathbf{x}_j)$;
      if $\eta > 0$ **then**
        更新 $\alpha_j^{new} = \alpha_j^{old} + \frac{y_j(E_i - E_j)}{\eta}$;
        裁剪 α_j^{new} 到 $[L, H]$;
        更新 $\alpha_i^{new} = \alpha_i^{old} + y_i y_j(\alpha_j^{old} - \alpha_j^{new})$;
        更新偏置 b;
        $numChanged = numChanged + 1$;
      end
    end
  end
until $numChanged = 0$;
return $\boldsymbol{\alpha}, b$;

的理论基础，在有限样本条件下依然能够获得良好的泛化性能，这对于样本容量有限且数据分布未知的历史数据集来说尤为重要。通过最大化类别间的间隔，SVM 不仅提升了模型的判别能力，还有效降低了过拟合的风险。在实验过程中，为了精准映射并捕获乘客性别、年龄、船舱等级等特征间的非线性关系，SVM 可选择使用径向基函数（RBF 核），可以更准确地反映各类特征组合对生还概率的复杂影响。

SVM 的稀疏性使其在面对噪声数据时表现出较强的鲁棒性。模型训练过程中，仅有少数支持向量决定最终的分类边界，这意味着即使数据中存在部分异常值或记录错误，也不会显著影响模型的

整体性能。此外，SVM 在高维特征空间中依然能够保持良好的表现，尤其适用于经过特征工程处理后的复合特征数据。这一特性使得 SVM 能够充分挖掘和利用数据中潜在的信息，提高预测的准确率，因此 SVM 模型很适合泰坦尼克号生还概率预测实验。

3.4 KNN

K 近邻（K-Nearest Neighbors, KNN）算法是一种基于实例的惰性学习方法，其核心思想是” 相似的样本具有相似的标签”。该算法通过计算测试样本与训练样本之间的距离，选择最近的 K 个邻居，并基于这些邻居的标签进行预测。

Algorithm 3: KNN 分类算法

Input: 训练集 D, 测试样本 $\mathbf{x}_q$, 邻居数 K, 距离函数 $d(\cdot,\cdot)$

Output: 预测标签 $\hat{y}$

初始化距离数组 $distances = []$;
for $i = 1$ **to** n **do**
    计算 $dist_i = d(\mathbf{x}_q, \mathbf{x}_i)$;
    $distances.append((dist_i, y_i))$;
end
按距离升序排序 $distances$;
选择前 K 个最近邻居 $N_K = \{(dist_j, y_j)\}_{j=1}^{K}$;
初始化类别计数 $counts = \{\}$;
for $(dist_j, y_j)$ *in* N_K **do**
    if 使用加权投票 **then**
        $weight = \frac{1}{dist_j + \varepsilon}$;
        $counts[y_j] += weight$;
    end
    else
        $counts[y_j] += 1$;
    end
end
$\hat{y} = \arg\max_c counts[c]$;
return $\hat{y}$;

KNN 算法作为一种非参数方法，KNN 无需对数据的分布形式做出任何假设，能够灵活适应数据本身的结构特点。这一优势对于包含多种复杂社会经济因素的乘客数据来说尤为重要，使得模型能够更真实地反映数据间的本质联系。KNN 的直观性也为其预测结果带来了良好的可解释性——通过分析与目标样本距离最近的邻居样本，可以直观理解模型做出某一预测的依据，有助于提升模型的透明度和信任度。

在处理局部模式识别方面，KNN 能够有效捕捉特定乘客群体的生还模式，例如针对特定年龄段的女性乘客或某一船舱等级乘客的生还规律进行识别和分析。此外，KNN 算法结构简单，易于实现，成为理想的基准模型，为后续引入更为复杂的算法提供了重要的性能参考标准。因此，KNN 算法也比较适合泰坦尼克号生还预测。

3.5 MLP

多层感知机（Multi-Layer Perceptron, MLP）是前馈神经网络的基本形式，通过多个隐藏层的非线性变换实现复杂函数的逼近。该算法采用反向传播（Backpropagation）算法进行参数学习，能够自动学习特征表示并处理高度非线性的分类问题。

Algorithm 4: 多层感知机训练算法

Input: 训练集 D, 网络结构 $\{n_1, n_2, \ldots, n_L\}$, 学习率 α, 最大迭代次数 T

Output: 训练好的 MLP 模型参数 $\{\mathbf{W}^{(l)}, \mathbf{b}^{(l)}\}_{l=1}^{L}$

随机初始化所有权重和偏置;
for $t = 1$ **to** T **do**
  for 每个训练样本 $(\mathbf{x}_i, y_i)$ **do**
    // 前向传播
    $\mathbf{a}^{(0)} = \mathbf{x}_i$;
    for $l = 1$ **to** L **do**
      $\mathbf{z}^{(l)} = \mathbf{W}^{(l)}\mathbf{a}^{(l-1)} + \mathbf{b}^{(l)}$;
      $\mathbf{a}^{(l)} = \sigma(\mathbf{z}^{(l)})$;
    end
    // 反向传播
    计算输出层误差 $\boldsymbol{\delta}^{(L)} = \frac{\partial \mathcal{L}}{\partial \mathbf{z}^{(L)}}$;
    for $l = L - 1$ **to** 1 **do**
      $\boldsymbol{\delta}^{(l)} = ((\mathbf{W}^{(l+1)})^T \boldsymbol{\delta}^{(l+1)}) \odot \sigma'(\mathbf{z}^{(l)})$;
    end
    // 参数更新
    for $l = 1$ **to** L **do**
      $\mathbf{W}^{(l)} := \mathbf{W}^{(l)} - \alpha \boldsymbol{\delta}^{(l)} (\mathbf{a}^{(l-1)})^T$;
      $\mathbf{b}^{(l)} := \mathbf{b}^{(l)} - \alpha \boldsymbol{\delta}^{(l)}$;
    end
  end
end
return $\{\mathbf{W}^{(l)}, \mathbf{b}^{(l)}\}_{l=1}^{L}$;

多层感知机（MLP）拥有强大的非线性建模能力，数据通过多层的神经网络，MLP 能够自动学习输入特征之间的高阶交互关系，可以发现传统的线性模型所不能捕捉的复杂结构。在本次泰坦尼克生还预测实验中，乘客的年龄、性别、船舱等级以及家庭关系等因素之间，往往潜存着非线性且难以预测的深层次的特征关系。MLP 可以通过隐藏层的逐级抽象和表示学习，MLP 能够有效识别并利用这些深层次的特征关系，可以更好的预测乘客生还的概率，这使得 MLP 在面对高度复杂和异质性强的数据时候，能够建模出一个表现优异的模型。

MLP 端到端的学习特质，使得其在模型的训练过程中能够同时完成对特征的学习和分类器的训练。通过灵活的调整网络结构的深度与宽度，MLP 可以适应不同的数据规模，同时避免过拟合的风险，基于以上 MLP 的优势，也是一个很适合本次实验的模型之一。

4 实验结果与分析

4.1 评估指标

在机器学习模型评估中，选择合适的评估指标对于衡量模型性能至关重要。一个表现优异的模型不仅应在整体性能上达到较高水平，还应在不同类别的预测能力上保持均衡。本实验中，针对泰坦尼克号生还预测问题的特点，选用了五个具有代表性的指标来综合评价模型的性能，分别是准确率（Accuracy）、精确率（Precision）、召回率（Recall）、F1 分数（F1-Score）以及 AUC-ROC 曲线。以下对每个指标的定义及其适用性进行详细阐述。

准确率 (Accuracy) 是最常用的分类性能指标之一，定义为模型预测正确的样本数量占总样本数量的比例。其计算公式为：

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

其中，TP 表示真正例，TN 表示真负例，FP 表示假正例，FN 表示假负例。尽管准确率能够反映模型的整体预测能力，但是在极端情况下（正负样本比例悬殊），高准确率不能够精准的反映模型的性能，此时单一的评价指标比较片面，不具有说服力。因此，仅使用准确率作为评估标准可能会导致对模型性能的误判。

精确率 (Precision) 用于衡量模型在预测为正类的样本中，实际为正类的比例，其计算公式为：

$$Precision = \frac{TP}{TP + FP}$$

精确率的意义在于评估模型在正类预测中的可靠性，特别是在关注降低误报（False Positive）影响的场景下具有重要作用。

召回率 (Recall)，也称为敏感度（Sensitivity）或真阳性率（True Positive Rate, TPR），衡量了实际正类样本中被正确预测为正类的比例，其计算公式为：

$$Recall = \frac{TP}{TP + FN}$$

召回率特别适用于关注漏报（False Negative）的任务场景，例如灾难预警或医疗诊断中，漏报可能带来严重后果的应用。

F1 分数 (F1-Score) 是精确率和召回率的调和平均值，用于综合评价模型的精确性和召回能力，其计算公式为：

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

在类别分布不平衡的情况下，F1 分数能够更全面地反映模型的性能，避免单独依赖精确率或召回率可能导致的偏颇。

AUC-ROC 曲线 是一种基于分类阈值变化的性能评估指标，衡量模型在不同分类阈值下的分类能力。ROC 曲线通过绘制真阳性率（TPR）与假阳性率（FPR）之间的关系来描述模型性能，而 AUC 值表示 ROC 曲线下的面积。AUC 的取值范围为 $[0, 1]$，值越接近 1，说明模型的分类能力越强。AUC-ROC 指标的优势在于其不依赖于具体的分类阈值，能够全面反映模型的整体分类表现。

4.2 各个模型性能对比

我们使用统一处理过的数据集在各个模型上先后进行了多轮训练，每次训练之后依据测试集的评估指标微调了相关的模型参数，直到训练出效果最好的模型，其最终评估结果已在表 1 中展现，可以发现不同模型在各评估指标上的表现存在显著差异，接下来我们将针对不同模型的评价指标进行深入地分析。

表 1: 各模型的性能评估结果

模型	准确率	精确率	召回率	F1 分数	AUC-ROC
随机森林 (Random Forest)	0.8492	0.8516	0.8492	0.8454	0.8779
K 近邻 (KNN)	0.8156	0.8140	0.8156	0.8142	0.8538
多层感知机 (MLP)	0.8268	0.8295	0.8268	0.8216	0.8616
支持向量机 (SVM)	0.8212	0.8294	0.8212	0.8132	0.8279

随机森林 (Random Forest) 在所有评估指标上表现优异，尤其在 AUC-ROC 指标上达到了 0.8779，表明其综合分类能力较强。这主要得益于随机森林模型能够通过集成多棵决策树捕获数据的复杂模式，同时具有较强的抗过拟合能力。此外，其精确率和召回率均较为平衡，分别为 0.8516 和 0.8492，使其在实际应用中具有较高的可靠性。

K 近邻 (KNN) 的表现相对较弱，其准确率为 0.8156，AUC-ROC 为 0.8538。由于 KNN 是一种基于实例的非参数模型，其性能对数据分布和特征缩放较为敏感。在本实验中，数据维度较高且分布复杂，可能导致 KNN 在捕获决策边界时表现欠佳。但是其精确率和召回率基本一致（分别为 0.8140 和 0.8156），说明其预测结果具有一定的稳定性。

多层感知机 (MLP) 展现出了较强的非线性建模能力。其 F1 分数达到了 0.8216，且在 AUC-ROC 上取得了 0.8616 的较好表现，说明 MLP 能够有效捕获数据的高阶特征。然而，与随机森林相比，MLP 的训练过程对超参数的依赖较强，且可能需要更多的计算资源。

支持向量机 (SVM) 的性能略低于随机森林和 MLP，其 AUC-ROC 为 0.8279，召回率为 0.8212。SVM 在小规模数据集上的表现通常较好，但在处理较大规模的高维数据时，其计算复杂度较高。此外，SVM 的性能较大程度上依赖于核函数的选择与参数调优，这可能在本实验中对其结果造成了一定限制。

4.3 实验模型的确定以及优势分析

经过多模型多维度的对比，我们确定了随机森林在准确率、精确率、召回率、F1 分数和 AUC-ROC 等多个指标上均表现优异，是本任务中最具竞争力的模型。随机森林在所有重要评估指标上均表现稳定且优越。其准确率达到了 0.8492，表明模型对整体数据的预测能力较强；精确率和召回率分别为 0.8516 和 0.8492，显示出该模型在减少误报和捕获正类样本方面的均衡能力；随机森林的 F1 分数为 0.8454，进一步表明其在处理类别分布不平衡问题时依然能实现较好的综合性能。最为显著的是，其 AUC-ROC 达到了 0.8779，表明在不同分类阈值下，随机森林能够有效区分正负样本，展现出卓越的分类能力。

相比之下，其他模型在性能上存在一定的局限性。K 近邻模型由于对数据分布的敏感性，其准确率和 AUC-ROC 分别为 0.8156 和 0.8538，未能充分捕获数据的复杂模式；多层感知机尽管在非线性建模方面具备一定优势，但其对超参数的依赖较高，导致模型在实验中未能达到随机森林的整体表现；支持向量机的性能略低于随机森林，其 AUC-ROC 为 0.8279，召回率为 0.8212，可能受到核函数选择和计算复杂度的限制。

随机森林之所以能够成为本任务中表现最优的模型，主要有以下优势：首先随机森林通过集成多棵决策树，能够有效降低单一模型可能产生的过拟合风险，同时提升模型的泛化能力，而且该模型具有内置的特征重要性评估机制，能够识别出对分类任务最具贡献的特征，为后续特征工程优化提供了科学依据。最重要的优势是随机森林对噪声数据具有较强的鲁棒性，即使在存在缺失值或异常值的情况下，依然可以保持稳定的性能表现。

基于实验结果的全面对比和分析，随机森林以其出色的性能和稳健的表现被确定为本任务的最终模型。未来的研究中，可以进一步结合随机森林的特征重要性分析结果，优化特征工程流程，从而进一步提升模型的性能。

5 实验结论与展望

本实验通过对泰坦尼克号生还预测的二分类任务进行系统性实验分析，比较了随机森林、K 近邻、多层感知机和支持向量机等经典机器学习模型的性能表现。在实验过程中，我们采用了多维评估指标，包括准确率、精确率、召回率、F1 分数和 AUC-ROC，以全面衡量各模型的优劣。相比之下，其他模型在某些特定指标上虽有亮点，但整体表现均逊色于随机森林。最终结果表明，随机森林模型在综合性能上表现最佳。

通过这次的实验，我不仅提升了对机器学习模型的实际应用能力，也让我意识到数据处理对实验结果产出的重要性。在模型选择的过程中，不仅需要对模型结构进行深入的学习，同时也需要对数据的结构有着深刻的理解。未来，我将继续探索更有效的特征工程方法、更先进的模型架构以及更高效的实验设计，为解决实际问题提供更科学的解决方案。