泰坦尼克号案例数据分析流程

1. 问题定义

1. **明确目标**：
    - 任务是**分类**问题。具体来说，是根据乘客的个人信息（如年龄、性别、船舱等级等）来预测其是否在泰坦尼克号沉船事件中**幸存** (Survived)。这是一个二分类问题（幸存 vs. 遇难）。
2. **确定评估指标**：
    - 首要的评估指标是**准确率 (Accuracy)**，即模型正确预测的乘客比例。
    - 同时，考虑到这可能是一个类别不平衡的问题（例如，遇难人数可能远多于幸存人数），也会关注如**精确率 (Precision)**、**召回率 (Recall)**、**F1分数 (F1-Score)** 以及**ROC曲线和AUC值**，以更全面地评估模型性能。

2. 数据收集与理解

1. **数据来源**：
    - 数据来源于Kaggle竞赛提供的两个CSV文件：`train.csv` (包含乘客信息及是否幸存的标签) 和 `test.csv` (包含乘客信息，用于预测其是否幸存)。
2. **数据探索**：
    - **统计分布分析**：
        - 查看各个特征（如年龄 `Age`、票价 `Fare`、船舱等级 `Pclass`、兄弟姐妹配偶数 `SibSp`、父母子女数 `Parch`）的描述性统计信息，如均值、中位数、标准差、最大最小值、四分位数等。
        - 了解目标变量 `Survived` 的分布，例如幸存者和遇难者的比例。
        - 分析分类特征（如性别 `Sex`、登船港口 `Embarked`）的取值频次。
    - **可视化分析**：
        - 使用直方图或密度图观察数值特征（如 `Age`, `Fare`）的分布情况，以及它们在幸存者和遇难者之间的分布差异。
        - 使用条形图分析分类特征（如 `Pclass`, `Sex`, `Embarked`）与幸存率的关系。
        - 利用箱线图或小提琴图比较不同类别下数值特征的分布。
        - 通过散点图或热力图探索特征之间的相关性。
    - **缺失值、异常值检测**：
        - 检查每个特征的缺失值数量和比例。在泰坦尼克号数据中，`Age`、`Cabin` 和 `Embarked` (训练集) 以及 `Fare` (测试集) 存在缺失值。`Cabin` 的缺失比例非常高。
        - 通过统计方法（如Z-score、IQR）或可视化（如箱线图）初步识别数值特征中的潜在异常值（例如，极高或极低的票价 `Fare`）。

3. 数据预处理

1. **数据清洗**:
    - **处理缺失值**:
        - 对于 `Age`：由于年龄是重要特征，不能简单删除。可以考虑使用中位数、均值（可能按 `Pclass` 和 `Sex` 分组后的中位数/均值）进行插值，或者更复杂的模型预测方法。
        - 对于 `Cabin`：由于缺失值过多，且信息难以恢复，可能会选择删除该特征，或者将其转换为一个表示是否有客舱信息的二元特征。
        - 对于 `Embarked`：缺失值较少，可以使用众数进行填充。
        - 对于 `Fare` (测试集)：缺失值极少，可以使用中位数或均值填充。
    - **处理重复值**：检查是否存在完全重复的乘客记录（在此案例中通常不太可能）。
    - **处理异常值**:
        - 对于 `Fare` 中的高价票，可以考虑进行截断（设置上限）或分箱处理，以减少极端值对模型的影响。
2. **数据转化**:
    - **标准化/归一化**：对于数值特征如 `Age` 和 `Fare`，如果选用的模型对特征尺度敏感（如SVM、逻辑回归的某些实现），可能需要进行标准化（转换为均值为0，标准差为1）或归一化（缩放到0-1范围）。
    - **类型编码**:
        - **标签编码 (Label Encoding)**：将分类文本转换为数字。例如，`Sex` 特征 ('male', 'female') 可以转换为 (0, 1)。`Embarked` ('S', 'C', 'Q') 可以转换为 (0, 1, 2)。
        - **独热编码 (One-Hot Encoding)**：对于名义分类特征（类别间没有顺序关系），如果类别较少，可以考虑使用独热编码，以避免引入错误的顺序关系。例如，`Embarked` 如果不适合标签编码，可以转换为三个二元特征。
3. **特征工程**:
    - **特征构造**:
        - 从 `Name` 中提取称谓 (`Title`，如 Mr., Mrs., Miss., Master.)，这可能与社会地位和幸存率相关。
        - 合并 `SibSp` 和 `Parch` 创建家庭成员数量 `FamilySize` (`SibSp + Parch + 1`)。
        - 基于 `FamilySize` 创建一个是否独自一人 (`IsAlone`) 的二元特征。
        - 考虑将连续特征如 `Age` 和 `Fare` 进行分箱 (Binning)，将其转换为有序分类特征，这有助于捕捉非线性关系并减少噪声影响。例如，将年龄分为“儿童”、“青年”、“中年”、“老年”等。
        - 可能会尝试创建交互特征，如 `Age * Pclass`，但需谨慎评估其有效性。
    - **特征选择**:
        - 删除明显无用的特征，如 `PassengerId` (对于训练模型而言)、`Ticket` (票号通常复杂且难以提取通用模式，除非进行深度挖掘)、以及处理后的 `Name` (如果已提取 `Title`) 和 `Cabin` (如果因缺失过多而决定放弃)。

        - 可以通过统计检验（如卡方检验评估分类特征与目标变量的关联性）、相关性分析或基于模型的特征重要性（如决策树或随机森林提供的特征重要性）来选择对预测幸存最有用的特征子集。
    - **降维**：在此案例中，特征数量不算特别多，可能不会优先考虑PCA或LDA等降维方法，除非在特征工程后产生了大量稀疏特征（如大量独热编码特征）。

4. 模型选择与训练

1. **划分数据集**：
    - `train.csv` 文件已经是标记好的训练数据。通常会从这个训练数据中再划分出一部分作为**验证集 (Validation Set)**，用于模型调优和评估，以避免在测试集上过早评估导致过拟合。`test.csv` 作为最终的**测试集 (Test Set)**，用于提交预测结果。常见的划分比例可能是训练集70-80%，验证集20-30%。
2. **选择模型**：
    - 由于是二分类问题，且数据集规模中等，可以尝试多种经典的分类模型：
        - **逻辑回归 (Logistic Regression)**：作为基线模型。
        - **K最近邻 (KNN)**
        - **支持向量机 (SVM)**
        - **朴素贝叶斯 (Naive Bayes)**
        - **决策树 (Decision Tree)**
        - **集成模型**：如**随机森林 (Random Forest)**、梯度提升树 (如 XGBoost, LightGBM, AdaBoost) 通常在此类问题上表现较好。
3. **训练与验证**：
    - 使用训练集对选定的模型进行训练。
    - 使用验证集进行**交叉验证 (Cross-Validation)**，如K折交叉验证，以获得更稳健的模型性能评估，并减少因单次划分验证集带来的偶然性。
    - 进行**超参数调优 (Hyperparameter Tuning)**，例如使用网格搜索 (Grid Search) 或随机搜索 (Random Search) 结合交叉验证，为每个模型找到最佳的超参数组合。

5. 模型评估

1. **分类问题评估指标**：
    - **准确率 (Accuracy)**：计算在验证集（或测试集）上正确预测的比例。
    - **精确率 (Precision)**：在所有预测为“幸存”的乘客中，真正幸存的比例。
    - **召回率 (Recall)**：在所有真正幸存的乘客中，被模型成功预测为“幸存”的比例。
    - **F1分数 (F1-Score)**：精确率和召回率的调和平均数，综合衡量两者。
    - **ROC曲线与AUC值**：绘制ROC曲线（以假正例率为横轴，真正例率为纵轴），计算曲线下面积AUC。AUC值越接近1，模型区分正负样本的能力越强，性能越好。ROC曲线越靠近左上角越好。
2. **模型比较与选择**：
    - 比较不同模型在验证集上的各项评估指标。
    - 选择综合表现最好的模型作为最终模型。在泰坦尼克号案例中，集成学习方法如随机森林通常能取得较好的效果。

- 分析最终选定模型的预测错误案例，尝试理解模型在哪些情况下表现不佳，是否还有改进空间。
- 最终使用选定的、调优后的模型在完整的训练集上重新训练（如果之前划分了验证集），然后对 `test.csv` 中的数据进行预测，并按比赛要求格式提交结果。

通过以上流程，可以系统地对泰坦尼克号数据进行分析和建模，以预测乘客的生还情况。