**1. 问题描述​​**

本项目基于Kaggle经典竞赛数据集"Titanic: Machine Learning from Disaster"，目标是通过乘客特征（如年龄、性别、船舱等级等）​**​预测其生存概率​**​。该问题属于二分类任务，核心挑战包括：

* 数据不完整（如年龄缺失、部分乘客无舱位记录）
* 特征间相关性复杂（如性别与生存率高度相关）
* 类别不平衡（约38%乘客幸存）

通过此项目，将掌握数据清洗、特征工程和分类模型构建的全流程数据分析技能。

**2. 数据来源​​**

使用Kaggle提供的泰坦尼克号数据集（[Titanic - Machine Learning from Disaster | Kaggle](https://www.kaggle.com/c/titanic/data)），包含以下两个文件：

1. ​**​train.csv​**​（891条样本，含生存标签）
   * 特征：PassengerId, Survived, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, Embarked
2. ​**​test.csv​**​（418条样本，无生存标签）

数据集已公开且无需额外授权，可直接用于学术研究。

**3. 分析方法​​**

**​​3.1 数据预处理​​**

* ​**​缺失值处理​**​：对Age用中位数填充，Cabin用"Unknown"标记，Embarked用众数填充
* ​**​特征转换​**​：将Sex转为数值型（0/1），对Fare分箱处理减少离群值影响
* ​**​新特征生成​**​：从Name提取称呼（如Mr/Mrs），从SibSp+Parch生成家庭规模

**​​3.2 探索性分析（EDA）​​**

* 绘制生存率与性别、船舱等级的堆叠条形图
* 使用箱线图分析Age和Fare的分布差异
* 计算特征间的Pearson相关系数

**​​3.3 建模与评估​​**

* ​**​基线模型​**​：逻辑回归（Logistic Regression）
* ​**​对比模型​**​：随机森林（Random Forest）、XGBoost
* ​**​评估指标​**​：准确率（Accuracy）、F1-score（解决类别不平衡）
* ​**​工具​**​：Pandas、Scikit-learn、Matplotlib/Seaborn

**​​4. 预期成果​​**

1. ​**​技术报告​**​：完整的数据预处理流程、特征重要性分析
2. ​**​可视化图表​**​：至少5种图表（如生存率热力图、年龄分布直方图）
3. ​**​模型性能​**​：在测试集上达到​**​>80%准确率​**​
4. ​**​可复用代码​**​：模块化的Jupyter Notebook代码（含详细注释）。