基于数据分析的员工离职预测及影响因素研究

一、数据背景领域

在人机共生的数字化转型浪潮中，人力资源管理正从传统事务性工作向战略决策支持升级，70% 的企业已在 HR 管理中应用 AI 与大数据技术。员工离职率过高会导致企业面临工作断层、招聘成本激增、团队绩效下滑等问题，尤其在技术密集型团队中，核心成员的不可替代性使得离职损失更为显著。

传统 HR 管理依赖经验判断离职风险，难以精准识别关键影响因素。而数据科学技术的应用为离职预测提供了新路径：通过分析员工特征数据（如满意度、薪资水平、工作时长等），可构建量化模型提前锁定离职高风险群体。大连理工大学等研究机构已证实，基于协作强度、技能匹配度等数据的算法模型，能有效预测离职可能性并优化人员替换策略，这为本次研究提供了技术可行性支撑。

本项目聚焦企业员工离职现象，结合 GitHub 公开的 HR 数据集，探究离职行为的关键驱动因素，为企业制定留存策略提供数据参考。

二、数据获取方法

1. **核心数据集来源**：通过指定 GitHub 仓库（https://github.com/ZhangWei214/Predicting_employee_left_HR/blob/main/README.md）获取员工离职相关原始数据集（文件名为 HR_datascience.csv），包含员工满意度水平、最后评估分数、参与项目数等 10 个关键字段（含目标变量 “是否离职”），数据结构清晰且为结构化格式。
2. **补充数据渠道**：
    - Kaggle 平台公开 HR 分析数据集（如 “Employee Turnover Dataset”），用于数据对比与模型验证；
    - 智联招聘《2025 年人力资源发展趋势》报告，提取行业基准离职率及影响因素统计数据；

1. **数据预处理说明**：采用 Python（pandas 库）进行数据处理，执行去重、缺失值处理、分类变量独热编码等操作，同时优化数据类型以减少内存占用，为后续分析提供高效标准化数据。

三、准备实现的分析目标

1. 标准化数据预处理：基于 hr_analysis.py 脚本逻辑，完成原始数据去重、缺失值处理，优化数据类型以降低内存消耗，对 “Department”“salary” 等分类变量执行独热编码，生成可直接用于建模的标准化数据集。
2. 多维度探索性数据分析：结合 matplotlib/seaborn/scienceplots 工具，生成指定可视化图表 —— 包括员工离职分布饼图、薪资水平分布饼图、各数值变量（满意度、工作时长等）分布箱型图、特征相关性热力图及特征间关系配对图等，直观呈现数据分布特征与变量关联模式。
3. 机器学习模型构建与优化：基于 scikit-learn 框架，实现逻辑回归与随机森林两种核心模型；采用随机搜索替代网格搜索进行超参数调优，结合 SelectKBest 方法完成特征选择优化，同时利用 joblib 库实现并行处理，提升模型训练效率。
4. 模型性能评估与可视化：通过混淆矩阵、随机森林模型 ROC 曲线评估模型预测性能，借助特征重要性条形图，量化 “满意度”“晋升情况” 等变量对离职预测的贡献度，筛选核心影响因素。
5. 计算性能优化落地：依托脚本中多核 CPU 并行处理逻辑，优化大数据集计算速度；对高维度数据进行采样处理，确保在保证分析精度的前提下，提升整体流程运行效率，匹配 hr_analysis.py 的性能设计目标。
6. 实用化离职管理支撑：基于模型输出的核心影响因素与预测结果，为企业提供可落地的决策建议 —— 包括识别高离职风险员工群体、定位关键影响因素（如低满意度、无晋升机会），助力制定针对性留存策略，提升员工满意度与留存率。

四、小组成员

卓越软件 2302 张辰远 231310227