2022 年中国高校大数据挑战赛

赛题 A　工业机械设备故障预测

制造业是国民经济的主体，近十年来，嫦娥探月、祝融探火、北斗组网，一大批重大标志性创新成果引领中国制造业不断攀上新高度。作为制造业的核心，机械设备在工业生产的各个环节都扮演着不可或缺的重要角色。但是，在机械设备运转过程中会产生不可避免的磨损、老化等问题，随着损耗的增加，会导致各种故障的发生，影响生产质量和效率。

实际生产中，若能根据机械设备的使用情况，提前预测潜在的故障风险，精准地进行检修维护，维持机械设备稳定运转，不但能够确保整体工业环境运行具备稳定性，也能切实帮助企业提高经济效益。

某企业机械设备的使用情况及故障发生情况数据见“train data.xlsx”，用于设备故障预测及故障主要相关因素的探究。数据包含 9000 行，每一行数据记录了机械设备对应的运转及故障发生情况记录。因机械设备在使用环境以及工作强度上存在较大差异，其所需的维护频率和检修问题也通常有所不同。

数据提供了实际生产中常见的机械设备使用环境和工作强度等指标，包含不同设备所处厂房的室温（单位为开尔文K），其工作时的机器温度（单位为开尔文K）、转速（单位为每分钟的旋转次数rpm）、扭矩（单位为牛米Nm）及机器运转时长（单位为分钟min）。除此之外，还提供了机械设备的统一规范代码、质量等级及在该企业中的机器编号，其中质量等级分为高、中、低（H\M\L）三个等级。对于机械设备的故障情况，数据提供了两列数据描述——“是否发生故障”和“具体故障类别”。其中“是否发生故障”取值为 0/1，0 代表设备正常运转，1 代表设备发生故障；“具体故障类别”包含 6 种情况，分别是NORMAL、TWF、HDF、PWF、OSF、RNF，其中，NORMAL代表设别正常运转（与是否发生故障”为 0 相对应），其余代码代表的是发生故障的类别，包含 5 种，其中TWF代表磨损故障，HDF代表散热故障，PWF代表电力故障，OSF代表过载故障，RNF代表其他故障。

基于赛题提供的数据，自主查阅资料，选择合适的方法完成如下任务：

任务 1：观察数据“train data.xlsx”，自主进行数据预处理，选择合适的指标用于机械设备故障的预测并说明原因。

任务 2：设计开发模型用于判别机械设备是否发生故障，自主选取评价方式和评价指标评估模型表现。

任务 3：设计开发模型用于判别机械设备发生故障的具体类别（TWF/HDF/PWF/OSF/RNF），自主选取评价方式和评价指标评估模型表现。

任务 4：利用任务 2 和任务 3 开发的模型预测“forecast.xlsx”中是否发生故障以及故障类别。数据“forecast.xlsx”。与数据“train data.xlsx”格式类似，要求在“forecast.xlsx”第 8 列说明设备是否发生故障（0 或 1），在第 9 列标识出具体的故障类型（TWF/HDF/PWF/OSF/RNF）。

任务 5：探究每类故障（TWF/HDF/PWF/OSF/RNF）的主要成因，找出与其相关的特征属性，进行量化分析，挖掘可能存在的模式/规则。

补充说明：

1. 开发语言不限，推荐使用python 3.7 及以上版本或Java 8 进行开发。

2.允许使用公开模型/开源代码，但需要在文档中注明出处。

3.除论文报告外，还需提供完整的程序代码、运行说明（包括依赖包、版本号）、预测结果文件“forecast.xlsx”及其他必要的佐证材料，以压缩包的形式提交。**注意：forecast.xlsx不要改名，便于评审专家检测。**

4.提交的支撑材料不得超过 20Mb。