实验二

背景

在借贷交易中，银行和其他金融机构通常提供资金给借款人，期望借款人能够按时还款本金和利息。然而，由于各种原因，有时借款人可能无法按照合同规定的方式履行还款义务，从而导致贷款违约。本次实验以银行贷款违约为背景，选取了约30万条贷款信息 ，包含在application_data.csv文件中，数据描述包含在columns_description.csv文件夹中。

数据来源：https://www.kaggle.com/datasets/mishra5001/credit-card/data

任务

任务一

编写MapReduce程序，统计数据集中违约和非违约的数量，按照标签**TARGET**进行输出，即1代表有违约的情况出现，0代表其他情况。

输出格式：

```
<标签><交易数量>
```

例：

```
1 100
```

任务二

编写MapReduce程序，统计一周当中每天申请贷款的交易数**WEEKDAY_APPR_PROCESS_START**，并按照交易数从大到小进行排序。

输出格式：

```
<weekday><交易数量>
```

例：

```
Sunday 16000
```

任务三

根据application_data.csv中的数据，基于MapReduce建立贷款违约检测模型，并评估实验结果的准确率。

说明：

1、该任务可视为一个“二分类”任务，因为数据集只存在两种情况，违约（Class=1）和其他（Class=0）。

2、可根据时间特征的先后顺序按照8：2的比例将数据集application_data.csv拆分成训练集和测试集，时间小的为训练集，其余为测试集；也可以按照8：2的比例随机拆分数据集。最后评估模型的性能，评估指标可以为accuracy、f1-score等。

3、基于数据集application_data.csv，可以自由选择特征属性的组合，自行选用分类算法对目标属性**TARGET**进行预测。

提交方式

提交git仓库地址或者相关文件的zip包。实验报告应包括设计思路、运行结果和可能的改进之处等。