

2022 年首届钉钉杯大学生大数据挑战赛初赛题目

初赛 B：航班数据分析与预测

一. 问题背景

随着科技的发展，乘坐飞机出行为人们的生活带来了极大的便利，航空交通管理影响着人们的工作和生活效率。在大数据和人工智能时代的今天，各种各样的信息科学和工程技术广泛应用于航空领域，为人们的生产生活提供更高的便利性，因此提高空中资源的优化配置，一直都是计算机科学与技术、信息科学与工程、数学等领域的一门热点研究方向。当前航空延误是空中资源优化配置的一项经典课题，分析与预测航空延误有助于提高资源的优化管理，提高生产生活效率，可以为乘客提供更优质的服务。

二. 解决问题

1. 航班转机功能实现：当两个城市之间没有直飞航班或者在购买机票附件时间没有直飞航班的时候，乘客通常需要购票 APP 实现转机功能。

（一）以附件中 **2001-2003** 年航班数据作为依据，实现在 **2001-2003** 年的航班转机功能。（注：从 A 地到 B 地可以有多种转机方案，该功能应提供时间最短的方案，已知的航班延误信息也应考虑在内）。

（二）用上述转机功能，查询 2003 年 7 月 4 日出发 7 月 5 日到达，从 CVG 机场到 ANC 机场最短时间方案。（注：航班延误时间也考虑在内）

2. 迈阿密（MIA）起飞航班的延误分析

以论文形式说明：先给出自己对问题（一）和（二）航班延误的分析，再建立模型做实验得到航班是否延误的准确率，以实验结果检验自己最初的分析。

（一）以附件中的 **2001-2003** 年的航班数据作为训练集，以附件 **2004-2005** 年的航班数据作为测试集，以从迈阿密（MIA）到洛杉矶（LAX）和从迈阿密（MIA）到纽约（JFK）这两组航班数据作为研究对象，先以文字形式叙述预测航班是否延误的依据，再建立模型预测从迈阿密（MIA）**起飞**航班（从 MIA 到 LAX 和从 MIA 到 JFK）是否延误，以预测准确率和实验结果检验自己的分析。（注：测试集不能参与到训练和验证中，否则作违规处理）

提示：可以在训练、验证和预测中使用机场所在地天气情况等信息，详见附

件数据属性说明表。

（二）以附件中的 **2001-2003** 年的航班数据作为训练集，以附件 **2004-2005** 年的航班数据作为测试集，以从迈阿密（MIA）到洛杉矶（LAX）和从迈阿密（MIA）到纽约（JFK）这两组航班数据作为研究对象，先以文字形式分析航班延误的各种原因，再建立模型预测从迈阿密（MIA）**起飞**航班（从 MIA 到 LAX 和从 MIA 到 JFK）延误的原因，以预测延误原因的准确率和实验结果检验自己的分析。（注：测试集不能参与到训练和验证中，否则作违规处理）

提示：参赛队伍可以先筛选出延误航班，再对延误航班的原因进行预测与分析，详见属性说明表中航班延误原因。在训练、验证和预测中可以使用机场所在地天气情况等信息，详见附件数据属性说明表。