运动员（用户）画像系统

研究内容：

在当今大数据时代，人们越来越关注如何从海量的数据中挖掘更有价值的信息，从而加以整合，优化用户体验，因此用户画像应运而生。通过原始数据抽象出一个用户的信息全貌，是用户画像最原始的想法，也是企业应用大数据的根基。尽管用户画像的构建形式和方法趋于相同，但不同领域的用户画像，标签设计以及应用分析却不尽相同。就本项目所关注的赛艇运动而言，构建运动员（用户）画像系统主要包括以下几部分内容：

将采集到的赛艇运动员数据信息进行标签化，并按基础、竞技、训练、体能、机能、康复等指标进行分类，构成完善的标签体系；通过各个维度数据的枚举，对运动员的各项特征属性进行细粒度刻画，进而挖掘潜在信息，抽象出运动员的画像标签；依赖于标签体系和运动员画像标签，按照条件筛选，对运动员进行分群，统计分析运动员特征，构建群体画像，同时按照特定条件勾勒出运动员个体画像；整合个体画像，从中归纳出具有赛艇项目夺冠条件的冠军模型，以此为参照，为运动员个性化制定精准且合理的饮食、训练、康复等计划，以提升竞技水平。

标签体系作为整个运动员画像系统的基础，可以看作构建系统的地基。其主要工作是将原始数据进行归类，并根据人为定义的指标标签化。标签体系要求便于扩展，以及便于理解和使用。标签体系比画像标签更全面、丰富，也正因此，需要对其进行分级。鉴于赛艇运动领域性较强的特点，系统对标签进行了详细的分类。总体上，根据指标分为六大类一级标签：基础类、竞技类、训练类、体能类、机能类、康复类。每个类别下再进行细分得到二级标签，例如，基础类包括：个人信息、身体形态等，其中个人信息可以再细分为三级标签：年龄、性别等，身体形态可以细分为身高、体重、肩宽等；竞技类包括：专项表现等，专项表现可以再细分为专项名、成绩等；训练类包括：训练计划、专项技术等，其中训练计划可以再细分为训练频次、训练量等，专项技术可以细分为专项动作等；体能类包括：基础体能、专项体能等，其中基础体能可以再细分为卧拉、预摆跳等，专项体能可以细分为单人双桨、双人双桨等；机能类包括：总体健康、实时状态等，两者都可细分为肺活量、心功能指数、最大摄氧量等；康复类包括：伤病史、治疗状况等，其中伤病史又可以细分为伤病时间、部位等，治疗状况可以细分为治疗单位、治疗时长等。

标签体系只是将结构化或非结构化的数据进行标签化，在构建完成标签体系的基础上，为了进一步对运动员各项数据进行分析，需要根据业务不同，将其抽象为画像标签。画像标签按不同方式又可分为统计类、规则类以及预测类标签。对于统计类而言，即使用统计学方法对原始数据进行清洗等处理，得到的诸如年龄分布、身高分布等标签。而规则类标签是将统计值进行业务规则转换，相比统计类标签，其拥有更具针对性的业务属性，并且更加直观。以标签体系中的体能类为例，在卧拉、预摆跳、负重深蹲等训练项目成绩基础上，统计全体数据后，人为定义成绩区间，并将其数值化，从而得出运动员的速度、耐力、力量、爆发力等规则类标签。（举例来说，假设卧拉成绩对应力量标签，运动员们的平均成绩为75kg，即设置成绩75kg为及格线60分，若某运动员测试成绩为80kg，按每高出1kg加5分计算，该运动员的该项成绩标签即为85分）预测类标签是基于统计和规则无法得出的复杂标签，需要借助机器学习算法，如决策树、贝叶斯等进行数据挖掘和训练，从而得到标签预测的结果。例如，根据运动员以往的训练数据，参照冠军运动员数值，预测运动员是否达到取得冠军的标准。

基于构建好的标签体系和画像标签，就可以通过制定约束条件和洞察分析对运动员进行分群，进而获得运动员个体画像和群体画像。类似于电商领域的RFM模型，在赛艇领域同样可以定义阈值对运动员进行分群。例如，根据运动员最近一次训练的时间，来衡量运动员比赛表现的因果关系；根据运动员训练频次和每次的训练量，衡量运动员的机能保持和竞技状态。进而可以将运动员分为重点培养、一般发展等群体，用以未来人才储备和培养计划的调整。运动员个体则可以通过处理好的画像标签进行集中整合，对其特征进行勾勒，以词云等形式生成其细粒度的个体画像。

运动员的个体画像和群体画像都是为冠军模型服务的。本项目是为备战奥运选拔出具有夺金潜力的运动员，因此需要从生成的优秀运动员画像中整合得到冠军画像。不同领域对于“冠军模型”的定义不尽相同，针对赛艇运动而言，冠军模型应当具备几个必要条件（标签）：身体形态、力量、耐力、技巧等。通过与冠军模型进行实时对比，可以对重点培养的运动员进行训练重点和训练计划等的变更及调整，从而使其更具备夺冠可能。

运动员画像系统是整个智能分析和决策辅助系统的基础之一，对于如何调整运动员训练计划，提升其竞技水平，进而提高夺金概率都有着重要的意义。

关键技术：

①构建动态画像。就运动员画像而言，可以分为静态画像和动态画像，鉴于赛艇运动的实时性和变化性，需要构建动态的运动员画像，根据运动员实时的饮食、训练及康复等状态变更，将结构化结果传递给运动员与教练员。

②扩展特征维度。画像标签是由数据标签经过分析、加工处理，形成的更加抽象、易于理解的复合标签。本项目主要为备战奥运培养夺金人才，因此对运动员各方面的要求都要达到世界级水准，因而需要对运动员画像标签的特征维度进行充分的扩展，以满足夺冠可能。

③融合特征处理方法。对于原始数据的处理往往建立在提取的特征上，主要分为连续类和离散类两大类特征。针对连续类特征，如运动员年龄、训练时长等，可以采取归一化、离散化以及增加非线性等方式，对于运动员训练数据、属性标签以及ID等离散型特征，则可以采取one-hot等经典方法。

④优化预测模型。针对预测类标签，传统的机器学习方法固然有其优势，但近来深度学习发展迅速，因而需要对预测模型做相应的优化和改进，提升预测效果。

⑤细化标签权重。对赛艇运动而言，运动员各项数据的实时性较强，因而基于时间衰减的权重尤为关键。单一的权重可能无法得出很好的变更计划，因此需要详细对权重进行枚举和设计，使其真实适应赛艇运动。

⑥离线自动更新。对于赛艇运动员海量的数据而言，使用人工进行数据录入、生成画像等操作需要花费大量时间和人力资源，因此整个运动员画像系统需要实现上传运动员数据后的自动化更新，同时满足离线分析计算，力求达到快速响应。

⑦可视化展示平台。整个画像系统以网站为主体，搭建具备信息可视化能力的平台，从而使数据信息更加形象，更具说服力。画像标签中的统计类、规则类标签等可以通过Echarts图表等形式进行展示，同时为了更好地将运动员画像用以训练等计划制定和变更，可使用词云形式进行可视化展示，冠军模型则可以借助六维图等形式，方便教练员对运动员专项技术进行查漏补缺。

⑧智能计划推荐。基于构建好的运动员画像平台，在对运动员各项指标进行综合评估后，系统需要根据人为制定的一系列规则，智能生成一套饮食、训练等计划，供运动员和教练员进行参考，进一步实现集计算分析、可视化、智能推荐为一体的画像平台。

预期效果：

研究内容三的总体目标是构建基于运动员画像和知识图谱的智能分析和决策辅助系统。就运动员画像平台而言，预期效果为，①在提供了运动员各项原始数据的基础上，处理为统一格式（如csv），通过平台上传，提取运动员各项数据，自动构建运动员特征库，实现离线分析和计算，对已有标签体系进行扩充和更新。②运动员、教练员注册账号，绑定个人信息，运动员可以查看个人数据，也可在站内进行搜索，查看其他运动员数据；教练员可以查看各运动员数据信息。③通过选择标签，实现对符合相应条件运动员的查询。④通过选择给定条件，可视化运动员分群，并根据分群结果提供智能化建议。⑤通过选择标签权重，自动调用模型，计算出预测结果。⑥通过可视化功能模块，基于数据库中运动员个人数据，生成运动员个人数据词云作为参考。⑦基于运动员分群结果，结合训练、竞技等标签，智能推荐训练计划，绑定运动员和教练员账号，实时传递提供参考。⑧根据一定时期内对运动员数据的跟踪分析，定期向运动员推送消息，提供饮食、康复等指导建议。⑨其他可视化功能模块。

风险分析：

就运动员画像系统而言，可能存在以下几类风险：①数据风险。对于上传数据中的重复值、缺失值和异常值，若未能有效处理，会导致画像分析结果不准确；由于训练数据涉及运动员隐私，而数据存储不可避免地存在泄露和被攻击的风险；②账号和设备风险。网站系统存在同一设备多账号登录、同一账号在多设备上登录等风险；③标签风险。规则类标签可能存在分类标准不明确或规则定义模糊，导致查询结果不准确问题；预测类标签由于模型选择或模型效果不佳，可能导致预测结果不准确，进而影响计划推荐。

实践主题：面向奥运备战的运动员画像分析系统开发

学习目标：

（1）了解企业处理大数据流程，学习基本的用户画像分析方法。

（2）设计运动员画像分析方案，依托已有数据开展全方位分析，并向运动员提供基本推荐。

（3）针对赛艇运动的固有特点，构建合理的冠军模型，以此为依据，提供专项计划建议。

学习内容：

（1）学习在大数据时代，如何从海量的数据中挖掘更有价值的信息，并加以整合，通过生成用户画像，优化用户体验。通过原始数据抽象出一个用户的信息全貌，从而应用于企业大数据分析。针对赛艇运动领域，构建运动员画像，设计标签并应用分析。

（2）采集赛艇运动员数据信息进行标签化，并按不同指标进行分类，构成完善的标签体系。基于标签体系和画像标签，通过制定约束条件和洞察分析对运动员进行分群，进而获得运动员个体画像和群体画像。从生成的优秀运动员画像中整合得到冠军画像，并对重点培养的运动员进行训练重点和训练计划等的变更及调整。

进度安排：

本项目计划以6个月为周期进行时间安排，主要分为初期、中期和末期三个阶段。每个阶段会有相应内容产出，初期阶段主要学习基本的大数据处理和分析方法，明确系统业务需求，提出合理的解决方案，并生成可行性分析文档；中期阶段主要针对提出的功能模块进行实现，包括标签体系构建、运动员分群、冠军模型生成等，并在此基础上提供直观的可视化界面；末期阶段主要对中期产物进行整合封装，形成完整可运行的系统。具体进度安排如下：

（1）初期: 2022年9月-10月：学习基本的大数据处理和分析方法，明确系统业务需求，提出合理的解决方案，并生成可行性分析文档。

（2）中期: 2022年10月-2023年1月：针对提出的功能模块进行实现，包括标签体系构建、运动员分群、冠军模型生成等，并在此基础上提供直观的可视化界面。该过程需要很强的工程能力，需要确保系统开发涉及技术的学习相对扎实，例如Hadoop、Spark等基本的大数据处理分析技术，在此基础上再进行功能模块实现。过程中需要及时和业务人员沟通，以确保业务顺利进行。

（3）末期: 2023年1月-2月：对中期产物进行整合封装，形成完整可运行的系统。在有余力的情况下，可以考虑引入深度学习推荐算法，并提供更加直观、准确的文字和图片形式的多模态推荐结果。