



**《Python玩转大数据》课程大作业**

**设**

**计**

**方**

**案**

|  |  |
| --- | --- |
| **项目名称：** | **足球五大联赛球员伤病数据分析及预测模型** |
| **学生学号：** | **2022216590** |
| **学生姓名：** | **吴秉康** |
| **所属班级：** | **精细化工22-2** |

**二O二五年五月**

# 足球五大联赛球员伤病数据分析及预测模型

设计方案

# 1.项目背景：

**随着足球的发展，现代足球比赛强度愈发激烈、战术对于球员的要求越来越高，同时赛程也变得更加密集，球员受伤风险显著提高。就算是足坛中出名的“铁人”，也避免不了受伤病困扰。但幸运的是现代足球又有着更先进的运动科学体系与运动康复体系，来预防和治疗运动员的伤病。**

**本项目通过爬虫获取了五大联赛（英超、西甲、德甲、意甲、法甲）截至2025年5月份所有球员的相关数据，并根据数据建立相关模型，并且将其可视化。以进行更直观的数据分析。在此基础上尝试利用大模型工具进行预测。**

# 2.项目条件及资源：

**操作系统：Windows 11**

**Python环境：Anaconda 2024.10**

**开发工具：PyCharm、Jupyter Notebook**

**数据来源：**www.transfermarkt.com

# 3.项目实施的步骤及可能的结果：

****步骤一：数据获取****

**经过多次爬取测试制定合适的爬取策略，将爬取结果保存为相应数据结构，并保存为文件。**

**其中，爬取策略的制定步骤：首先浏览各足球数据网站，寻找所需数据，并用浏览器的DevTool分析网站的架构；然后用测试网站反爬机制，设置连续访问间隔时间；利用BeatifulSoup4进行解析，并设计自动爬取的流程。**

**数据存储结构：先通过联赛界面得到所有球队，再通过球队界面得到所有球员；再将数据源分为“球员基本信息”与“球员伤病信息”两个部分保存。**

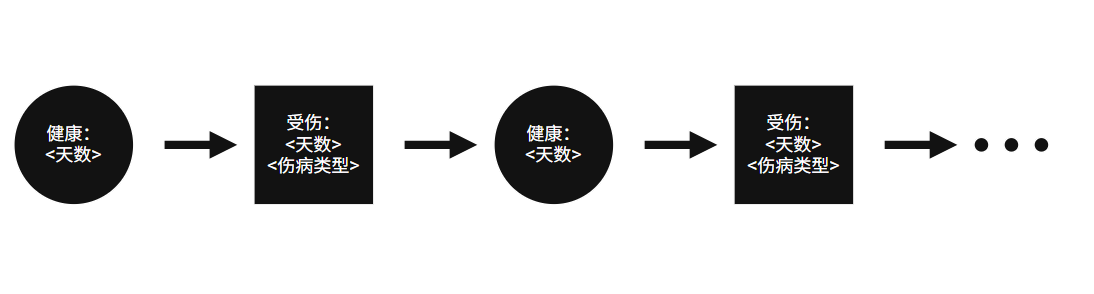
**步骤二：数据用途分析**

**这一步为项目的关键，承上启下，将项目的目标具体化。具体问题具体分析，我们得到的数据能够解答什么问题？解决问题又需要什么样的数据？两者是同时思考同时解决的。**

**项目主题是对伤病的分析和预测，为构建相应的模型，我们需要将球员和球员的伤病抽象化。**

**球员抽象：位置，身高，体重，打法，场均出场时间，场均跑动量。**

**伤病行为抽象：用一个单向的线性结构表示。这个结构可以看作为一个随时间变化的函数，也可以看作成一个数列。**

****

球员伤病行为抽象

**接下来就是实现相应的数据结构，并通过技术实际进行优化，设计相应的接口。**

**步骤三：数据预处理**

**使用Pandas将数据加载成相应结构。**

**由于数据是自己爬取的，数据清洗的工作量并不大，完成标准的重复值、缺失值、异常值检查以及数据类型转换就可以。**

**再根据下面两部，进行所需要的操作。例如数据的标准化，数据集的划分等。**

**步骤四：数据可视化**

**这个步骤分为两个部分。**

**第一部分，单纯伤病类型的统计分析。将球员剥离开，仅分析不同伤病类型出现的频率、康复天数以及对球员的影响。可以画出相关的频率分布图，词云图等。另外还可以画出不同位置、不同联赛的统计数据。**

**第二部分，球员伤病历史的波纹图。可以将球员伤病行为抽象可视化为以时间为横轴的波动图像，可以直观表现出单个球员受伤病影响的程度和特征。另外还可以依照次波纹将球员分类。**

**步骤五：机器学习模型**

**这一步实现伤病的预测。**

**首先是模型的选用。伤病预测可以看作为一个序列建模问题，输入包含两个部分：**

**1. 球员抽象：固定维度特征（x1, x2, x3）**

**2. 球员伤病行为抽象：可变长度序列（a1, a2, ... , an）**

**目标是预测数列的后两项 (an+1, an+2)**

**由于存在可变长序列，一般的回归模型较难做到。具体策略需要进一步的分析与实践。**

**模型选取与训练之后，再设计简单的接口方便输入输出即可。**

# 4.可行性分析：

**德转网站 （www.transfermarkt.com）的数据可信、完整、多维，一个网站可以完成所有的数据需求。而且经过实验发现该网站有着较低的反爬强度，使用request库就可以很容易实现自动化的爬取。并且球员伤病信息只有该网站有详细的、系统化的数据，所以可以说，利用该网站是项目成功的关键。**

**经爬取得五大联赛球员数量为2591个，这个规模的样本容量可以满足描述性统计分析、可视化与简单机器学习模型的训练。**

**另一个对项目起决定性作用的是对伤病模型的构建与机器学习模型的选取。抽象模型已经构建了，技术上的实现还需要进一步的研究与实践。若在当前条件下，目前技术上无法实现，则简化模型处理。**

# 5.未来展望：

**该项目能够让我们对足球运动员的伤病，通过数据分析，有着更清晰的认知。伤病虽然残酷，但仍是竞技运动的一部分。我们需要理性地看待它，科学地预防它。**

**足球是及其复杂的系统，但是一般来说对其的评价体系是纯经验主义的，缺乏科学性的。**

**该项目虽说有一定量的数据和技术支撑，但由于各种限制，其分析与模型仍是偏经验的。对伤病专业的预测需要更精细的数据和模型。球员的身体运动模式、肌肉骨骼结构，以及心理因素、场外因素都是我们无法获取同时也是关键的数据。所以说该项目存在局限性，从而导致其一定的非专业性。**

**通过该项目的完成，我们有了可靠的数据来源与较为成熟的处理流程，可以更轻松得完成足球相关的其他类型的数据分析与可视化，如球队战术风格分析、球员身价及状态分析、球员与球队适配度分析等等。**

**该项目也可以进一步扩大规模，获取其他联赛的球员数据。甚至经过简单修改，进行其他体育项目运动员的伤病分析及预测。另外还可以设计GUI方便输入输出，甚至封装为产品。**