



BRICS
Business Council



2023

金砖国家职业技能大赛 (金砖国家未来技能挑战赛)

技术描述 TD(仅供省级选拔赛参考)

BRICS-FS-02_机器学习与大数据

2023 年 5 月

目录

一、 项目简介	2
(一) 项目描述	2
(二) 竞赛目的	3
(三) 参赛对象	3
二、 选手应具备的能力	4
三、 竞赛项目	10
(一) 竞赛模块	10
(二) 模块简述	10
(三) 命题方式	11
(四) 命题方案	11
四、 评分规则	12
(一) 评价方法	12
(二) 评分程序	13
(三) 成绩计算	14
(四) 裁判组构成和分组	15
五、 竞赛相关设施设备	16
(一) 竞赛技术平台标准	16
(二) 环境要求	16
(三) 设备清单	17
六、 竞赛须知	17
七、 赛场布局要求	18
八、 健康安全和绿色环保	18
(一) 比赛环境	18
(二) 安全要求	19
(三) 环境保护	20

一、 项目简介

（一） 项目描述

机器学习与大数据是针对庞大数据集分析和洞察的领域，涉及到大数据环境、数据处理、模型构建和算法优化等关键任务。从业人员需熟悉主流大数据操作环境，数据分析和机器学习的应用分析包，掌握分布式计算和统计分析的基本理论和基本技能，具备运用工程化方法和工具完成软件编码和测试的能力，完成数据的信息挖掘。

从业人员的专业能力具体要求包括：明确需求、大数据环境准备、数据获取、数据清洗与存储、数据分析、数据挖掘、数据可视化等，从业人员能够：通过项目需求分析了解面向的用户群体的目标和需求，确定所需数据集和数据收集方法；其次，处理缺失值、异常值和重复数据，确保数据质量和一致性。然后，提取、选择和转换数据特征，以获得适合机器学习模型的输入。接着，选择合适的机器学习算法和模型架构，并进行交叉验证和调优，以提高模型的性能和泛化能力。模型训练完成后，进行模型评估，使用测试数据集评估模型的准确性和效果，并进行必要的调整和改进。最后，他们将训练好的模型应用于实际数据，并进行预测、分类或推荐等任务。另外从业人员还应该具备其他通用能力，例如专业英语阅读能力、解决问题的能力、组织与沟通能力等。

（二）竞赛目的

为继续落实金砖国家《厦门宣言》《约翰内斯堡宣言》《巴西利亚宣言》《莫斯科宣言》和《北京宣言》中关于技能发展工作、技术创新工作的相关精神，通过成员国之间的交流合作，在金砖五国和一带一路范围内，促进机器学习与大数据技术的发展，推动机器学习与大数据技术和产业应用结合。赛项体现了行、校、企联合设计思路，促进校企合作协同育人，对接产业发展，实现行业资源、企业资源与教学资源的有机融合，针对机器学习与大数据行业典型工作岗位，培养学生机器学习与大数据，达到以赛促教，以赛促学的目的，使学校能够更好地建设专业，提高教学质量，创新教学模式。

（三）参赛对象

本赛项采取个人赛方式，16-35 岁均可参赛，参赛选手需在规定时间内完成所有竞赛模块。每支参赛队由 1 名选手组成，每支参赛队可配备 1 名专家。

二、 选手应具备的能力

相关要求		权重
1	工作组织、管理	5%
	个人需要知道和理解： ● 确保高效团队合作的原则和技能； ● 系统的原则和行为； ● 高产品稳定性和环境安全性的系统方面； ● 分析和评估从各种来源获得的信息。	
	个人应能够： ● 考虑时间限制和最后期限； ● 调试和处理错误； ● 使用计算机或设备和一系列软件包； ● 应用研究技术和技能，以保持最新的行业指南； ● 根据可用时间计划每天的生产计划； ● 使用英文版操作系统和软件，按照任务要求完成英文版作品； ● 掌握丰富的专业英语词汇、具备英文阅读能力。	
2	沟通和人际关系技能	5%
	个人需要知道和理解：	

	● 记录决策的重要性； ● 解决问题的重要性； ● 书面和口头沟通技巧的重要性； ● 详细记录已开发解决方案的重要性； ● 在准备文件方面表现出专业性。	
	个人应能够： ● 阅读理解规则文档； ● 遵循所提供手册中的书面说明； ● 了解工作场所组织的说明和其他技术文件； ● 对需求的解释和理解； ● 了解最新的行业建议； ● 讨论和提出数据； ● 提出建议和最终决定； ● 使用书面沟通技巧； ● 开发用户文档； ● 处理英文技术文档； ● 在 Jupyter Notebook 或类似环境中编写有关数据分析的交互式报告； ● 任务的优先级和调度； ● 任务之间的资源分配。	
3	解决问题、创新、创造力	5%

	个人需要知道和理解： ● 开发数据分析解决方案时可能出现的一般问题类型； ● 商业组织中可能出现的一般类型的问题； ● 解决问题的诊断方法； ● 行业趋势和发展，包括新技术、方法、语言、惯例和技术技能。	
	个人应能够： ● 复杂或异构信息的分析和综合； ● 定义平凡和非平凡的数据依赖关系； ● 独立查询问题； ● 及时发现和解决问题； ● 正确收集和分析信息； ● 制定决策替代方案，选择最合适的选项并实施必要的解决方案。	
4	大数据搭建与运维	25%
	个人需要知道和理解： ● 大数据处理框架； ● 数据管理和存储； ● 分布式计算原理； ● 数据处理和分析；	

	● 数据安全和隐私保护。	
	个人应能够： ● Hadoop 环境搭建； ● Hadoop 相关组件安装配置； ● 大数据集群故障排查； ● HDFS 的基础操作； ● MapReduce 计算； ● Sqoop 数据迁移； ● 数据仓库构建； ● Hive 数据查询； ● Spark 分布式计算； ● Flume 数据采集； ● Kafka 消息队列； ● Flink 流数据处理；	
5	数据分析与处理	25%
	个人需要知道和理解： ● 各种数据分析； ● 数据分析技术； ● 提出并检验关于数据依赖性的假设； ● 数据收集方法； ● 图形数据的工作原则；	

	● 处理文本数据的原则。	
	个人应能够： ● 使用开发工具软件完成开发要求； ● 分析时间序列； ● 识别数据中的异常和偏差； ● 结构数据； ● 进行数据的规范化和准备； ● 突出数据中对象的特征、属性和特性； ● 聚类数据； ● 搜索数据相关性； ● 利用大数据开展业务； ● 识别数据中的模式； ● 对数据进行可视化分析； ● 创建模式； ● 使用软件工具进行研究和数据处理。	
6	构建机器学习模型	35%
	个人需要知道和理解： ● 机器学习的各种方法和算法； ● 机器学习模型的质量标准； ● 机器学习模型的发展顺序； ● 根据源数据应用哪些机器学习方法；	

	● 如何处理各种数据样本； ● 如何使用各种软件工具来开发和改进模型。	
	个人应能够： ● 使用开发工具软件完成开发要求； ● 无监督学习（降维、搜索规则、聚类）； ● 监督学习（回归、分类）； ● 应用强化学习算法； ● 遗传算法； ● Q-学习算法； 应用机器学习的集成方法： ○ stacking； ○ bagging； ○ boosting； ● 深度学习方法和神经网络： ○ 感知器； ○ 卷积神经网络； ○ 递归网络； ○ 自动编码器	

三、 竞赛项目

（一） 竞赛模块

模块	考核模块	时间分配	权重
A	大数据	2.5 小时	30%
B	数据分析	2.5 小时	30%
C	机器学习	2.5 小时	40%
比赛总用时		7.5 小时	

（二） 模块简述

1. 模块 A： 大数据

参赛选手依据任务要求，对 HADOOP 平台进行操作。主要内容包括但不限于：HADOOP 集群搭建、HDFS 常用操作命令（查询文件类别、上传、删除文件、查询 HDFS 基本统计信息等）、FLUME 数据采集、SQOOP 数据导入、MAPREDUCE 批处理操作、HIVE 数据分析、SPARK 数据挖掘、KAFKA 流式处理、FLINK 实时计算、HUE 可视化分析、HADOOP 集群的负载均衡等。

2. 模块 B： 数据分析

参赛选手根据任务所提供的数据源，结合数据业务背景明确分析目的和思路；将不同格式的数据通过整理、清洗、转化、合并等过程，形成完整的单一有效数据源；接着使用

数据处理技术将各种原始数据加工为数据分析所需要的样式，通过多种数据分析方法进行数据分析最后考察对数据可视化美学的理解和应用能力，如图表设计、颜色搭配、字体选择、布局等。

3. 模块 C：机器学习

理解实际问题、数据获取、数据探查、数据处理、特征工程、模型训练、诊断与调优、模型融合，考察选手在大数据环境下，采用分布式技术将问题模型化的能力，复杂数据结构化的能力，理解特征并可以有效转换特征的能力，选择恰当算法的能力，优化损失函数的能力，模型训练与模型融合的能力。

（三）命题方式

竞赛进行技能实操，涉及大数据、数据分析、机器学习 3 个模块，根据赛题要求进行设计和编码最终达到所需结果和目的。

本项竞赛为闭卷，技术文件公布后，即刻发布考核思路、命题方向；赛前 14 天公布样题。涉密部分在正式比赛时公布。

（四）命题方案

本次竞赛以大数据环境为基础，实现数据分析与挖掘，涵盖与该领域相关的核心技能和要点，旨在考察参赛选手在机器学习和大数据应用方面的管理与分析能力。竞赛尽可能

保留企业级机器学习与大数据项目的技术实际需求，以检验参赛选手能够运用所学知识和技术，分析并实现数据价值的能力。

四、 评分规则

本项目评分标准为测量和评价两类。凡可采用客观数据表述的评判称为测量；凡需要采用主观描述进行的评判称为评价。

（一） 评价方法

1. 评价分（主观）

评价分（Judgement）打分方式：3 名裁判为一组，各自单独评分，计算出平均权重分，除以 3 后再乘以该子项的分值计算出实际得分。裁判相互间分差必须小于等于 1 分，否则需要给出确切理由并在小组长或裁判长的监督下进行调分。

权重表如下：

权重分值	要求描述
0 分	各方面均低于行业标准，包括“未做尝试”
1 分	达到行业标准
2 分	达到行业标准，且某些方面超过标准
3 分	达到行业期待的优秀水平

2. 测量分（客观）

测量分（Measurement）打分方式：按模块设置若干个评分组，每组由 3 名及以上裁判构成。每个组所有裁判一起商议，在对该选手在该项中的实际得分达成一致后最终只给出一个分值。若裁判数量较多，也可以另定分组模式。

3. 主观评估与客观评估的运用

主观评估样例：

权重分值	要求描述
0 分	未实现本模块
1 分	实现了本模块的部分内容
2 分	实现了本模块的内容
3 分	实现了本模块的内容，设计效果完美，达到了达到行业优秀水平

客观评估准则样例表：

类型	示例	最高分 值	正确分 值	不正确分 值
满分或零分	计算并保存结果正确	1.00	1.00	0

（二）评分程序

所有模块均为赛后成果评分，选手需要在规定时间内，将当前模块的作品提交至指定地方，裁判员从答案存放点拷

贝作品并进行评分。

评价分和测量分将分开进行评判，同一评判组有可能既要负责评价分又要负责测量分的评判工作，根据裁判长的实际安排依次开展评判工作。

评分过程评分表不得涂改，任何涂改需要小组所有成员签名确认，同时任意一环评分工作结束，均需要对自己所参与负责的评分表每一页进行检查和签名确认。

（三）成绩计算

1. 抽检复核

为保障成绩评判的准确性，监督组对赛项总成绩排名前30%的所有参赛队伍的成绩进行复核；其余成绩进行抽检复核，抽检覆盖率不低15%。监督组在复检中发现错误，需以书面形式及时告知裁判长，由裁判长更正成绩并签字确认。如复核、抽检错误率超过5%，裁判组需对所有成绩进行复核。

2. 统分方法

赛项采取分步得分、累计总分的计分方式进行评分。各模块间分别计算得分，模块间错误不传递。竞赛项目和竞赛总分均按照百分制计分。

3. 成绩并列

如选手决赛成绩出现同分情况的，按照C>A>B模块的顺序计算排名顺序。

（四） 裁判组构成和分组

1. 裁判组

评分前，由裁判进行随机抽签分组，杜绝主观意愿组队，各裁判组按竞赛模块独立进行评分，评分后统计总分，确保成绩评定客观、严谨、准确。

2. 裁判建议

裁判员听从裁判长的安排，必须经过现场培训，裁判长对所有的裁判员进行分组、分工，裁判员在评判时，公平公正的评判和记录竞赛成绩，不得帮助选手完成工作任务，个人不得私自修改选手的作品。

3. 评判中的纪律和要求

裁判员出入赛场要佩戴胸牌，衣着整齐，举止大方，不大声喧哗，听从指挥，服从裁判长的安排。

遵守保密规定，保证公开、公平、公正原则。

裁判员和选手，在竞赛现场一律不允许带入或带出任何通讯设备、智能设备、存储设备；竞赛期间不泄露任何竞赛信息。

裁判员要注意自身的安全，操作符合各项规范，竞赛时不得随意进入选手工作区。

五、 竞赛相关设施设备

（一） 竞赛技术平台标准

本竞赛平台基于分布式计算思想，按照企业机器学习与大数据相关工程师人才标准，指导学生完成大数据分析与挖掘项目。参赛选手在平台初始环境 V0.1 基础上，通过项目实训进行需求分析、Hadoop 环境搭建与管理、编码查询、数据分析和可视化展示等，实现项目 V1.0 版本。通过项目实训提高选手机器学习与大数据综合技能。

（二） 环境要求

选手竞赛环境清单：

名称	技术规格
Hadoop	3. 3. x
Sqoop	1. 4. x
Hive	2. 3. x
Zookeeper	3. 6. x
Flume	1. 10. x
Spark	3. 3. x
Flink	1. 15. x
JDK	11
Scala	2. 13. x
Python	3. 7. x

（三）设备清单

1. 技术平台

赛场提供统一的云计算平台。

平台包含数据准备、数据整理、数据分析、报表生成等比赛环境，融合机器学习与大数据的技术知识点，内容包括：Hadoop 平台搭建、HDFS 数据存储、Flume 数据采集、Sqoop 数据导入、MapReduce 批处理操作、Hive 数据查询、大数据可视化、数据整合、数据读取、数据探查、数据处理、数据建模、模型评价。

2. 硬件规格

类别	配置
选手电脑	操作系统：Windows 10（64 位） 处理器：Intel 9 代 i5 及以上 内存：16G 及以上 固态硬盘：128GB 或以上 显示器：23.8 寸 LED

六、 竞赛须知

赛场内除指定的裁判、工作人员外，其他与会人员须经组委会同意或在组委会负责人陪同下，佩带相应的标志方可进入赛场内。

允许进入赛场的人员，只可在安全区内观摩竞赛，不得

使用录像设备长时间拍摄选手工位、屏幕。

允许进入赛场的人员，应遵守赛场规则，不得与选手交谈，不得妨碍、干扰选手竞赛。

允许进入赛场的人员，不得在场内吸烟、喧哗。

七、 赛场布局要求



根据本项目的特点，比赛场地包含了选手竞赛区、裁判休息区、录分区及裁判长工作区，本图仅供参考。

八、 健康安全和绿色环保

（一） 比赛环境

竞赛场地应为通风、明亮的室内场地，场地净高应不低于 3.5m，应保证赛场采光(大于 500lux)、照明和通风良好。

每个比赛工位上标明赛位编号，同时配备 1 个工位和 1 个设备位，用于配备竞赛平台和技术工作要求的软、硬件，每个比赛间配置工作台，用于摆放计算机、显示器、工具等。

竞赛场地每个工位内设有操作平台并配备 220 伏电源，工位内的电缆线应符合安全要求。

竞赛场地设置裁判区，并配置计算机等统计工具，记录各参赛队的比赛全过程。

竞赛场地设置服务区，提供维修服务、医疗、生活补给等服务保障。

技术支持区为参赛选手提供 PC、竞赛备用平台等竞赛相关设备。

（二）安全要求

1. 赛场人员安全要求：

现场裁判、选手、工作人员在竞赛期间应该遵守执委会的安全规定和要求。

参赛选手进入竞赛场地后，须听从并尊重裁判人员的管理，文明参赛。

参赛选手必须在确保人身安全和设备安全的前提下开始竞赛，发现或发生有关安全问题，应立即向裁判报告。

参赛选手严禁在赛场区域内吸烟和私自动用明火，严禁携带易燃易爆物品。

2. 设施设备安全操作要求

. 禁止选手及所有参加赛事的人员携带任何有毒有害物品进入竞赛现场。

赛点单位应设置专门的安全防卫组，负责竞赛期间健康和安​​全事务。主要包括检查竞赛场地、与会人员居住地、车辆交通及其周围环境的安全防卫；制定紧急应对方案；监督与会人员食品安全与卫生；分析和处理安全突发事件等工作。

3. 场消防安全要求

消防设施、器材和消防安全标志全都在位且功能完整。

消防安全重点部位人员正常在岗工作。

4. 安全标识张贴要求

安全出口、疏散通道保证畅通，安全疏散指示标志、应急照明完好无损，竞赛场地安全疏散通道禁止被占用。

（三）环境保护

赛场严格遵守我国环境保护法。

赛场所有废弃物应有效分类并处理，尽可能地回收利用。

2023

金砖国家职业技能大赛 (金砖国家未来技能挑战赛)



金砖职赛微信号