



2024

金砖国家职业技能大赛（金砖国家未来技能和技术挑战赛）

机器学习与大数据

BRICS-FS-02-RU

技术规程（全国选拔赛暨国内决赛）

2024 年 02 月



目录

1 简介 1

1.1 技能竞赛名称及说明..... 1

1.2 本文件的相关性和重要性 2

2 技能标准 3

2.1 技能标准的一般说明 3

2.2 技能标准..... 3

3 评分方案 10

3.1 评分方法 10

3.2 评分规则 10

3.3 评测依据..... 10

4 竞赛赛题 12

4.1 常见注意事项 12

4.2 竞赛赛题格式/框架 12

4.3 竞赛赛题时间分配及分值权重 13

4.4 各模块作业内容及要求 13

4.5 竞赛样题公布 15

4.6 竞赛赛题修订 15

5 技能管理与沟通 16

5.1 专家组..... 16

5.2 讨论交流	16
6 安全要求	17
6.1 组织机构	17
6.2 赛项安全管理.....	17
6.3 比赛环境安全管理.....	17
6.4 生活条件保障.....	19
6.5 参赛队职责	19
6.6 应急处理	19
6.7 处罚措施	20
7 材料和设备	21
7.1 基础设施列表.....	21
7.2 参赛选手的工具箱.....	21
7.3 竞赛设备清单	21
7.4 在技能区域内禁止使用的材料和设备	22
7.5 建议的比赛区域和工作站布局	22
8 技能特定的规则	23

1 简介

1.1 技能竞赛名称及说明

1.1.1 技能竞赛的名称

2024 金砖国家职业技能大赛（金砖国家未来技能挑战赛）机器学习与大数据（Machine Learning and Big Data）。赛项编号:BRICS-FS-02-RU。

1.1.2 技能竞赛描述

本次机器学习与大数据竞赛项目集中于应用机器学习和大数据技术来解决现实问题，特别强调大数据处理、数据分析和机器学习模型的构建与优化。此竞赛不仅测试选手的技术技能，更通过模拟实际业务场景的挑战，考察他们的创新思维、项目管理和解决问题的能力。

机器学习与大数据技术应用包括几个方面：大数据、数据仓库、数据分析、机器学习，参赛选手需要具备以下的工作技能：

大数据：选手能够搭建和配置 Hadoop 基础环境，构建和管理数据仓库，并能够迅速诊断并解决任何系统故障，保证数据处理和存储任务的顺畅执行，还能实现大数据离线计算和实时数据统计与分析等任务。

数据仓库：选手能够设计、构建和维护数据仓库；掌握数据仓库的基础架构和数据模型；实现数据的提取、转换和加载（ETL）过程；并通过数据仓库支持数据分析和业务决策。还需要具备数据仓库的性能优化和安全管理能力。

数据分析：选手能够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；并通过可视化展示数据。

机器学习：选手能够在理解数据的基础上，应用机器学习算法构建模型以预测或分类，进而解决实际问题。选手能展示对不同算法的掌握，如回归分析、分

类分析、神经网络等，以及对模型进行调参和优化的能力。

1.2 本文件的相关性和重要性

本文件包含本次技能竞赛所需的标准，以及管理竞赛的评测原则、方法和程序的信息。

每位专家和选手都必须了解和理解本技术说明。

如果不同语言的技术说明之间有任何冲突，以英文版本为准。

2 技能标准

2.1 技能标准的一般说明

技能标准规定了知识、理解和特定技能，这些技能是国际上在技术和职业表现方面的最佳实践。它将反映全球对相关工作角色或职业在工业和企业中代表什么的全球共识。

技能竞赛旨在反映该技能标准所描述的国际最佳实践，以及它所能达到的程度。因此，该标准是技能竞赛所需培训和准备的指南。

该标准分为不同的带有标题和参考编号的部分。

每个部分被分配总分的百分比，以表明其在标准中的相对重要性。这通常被称为“权重”。所有百分比的总和分值为 100。权重决定在评分标准中分值的分配。

通过竞赛赛题，评分方案只对标准中列举的技能进行评测。他们将在技能竞赛的约束下尽可能全面地反映标准。

评分方案将在实际可能的范围内按照标准中分配的分值进行。允许有 5% 的变动，但不得改变标准规范分配的权重。

2.2 技能标准

部分		相对重要性 (%)
1	工作组织、管理	
	选手需要知道和理解： – 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。	

	选手应能够： - 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	
	选手需要知道和理解： - 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	
	选手需要知道和理解： - 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。	

	选手应能够： - 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	20
	选手需要知道和理解： - 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。	
	选手应能够： - 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； - Spark 分布式计算； - Flume 数据采集；	

	- Kafka 消息队列； - Flink 流数据处理。	
5	数据仓库构建与维护	20
	选手需要知道和理解： - Sqoop 命令行工具的操作方法； - HiveQL 高级查询功能和优化技术； - 数据验证的 SQL 脚本设计； - 数据表结构及其业务应用； - 技术环境的详细配置和工具。	
	选手应能够： - 执行 Sqoop 全量和增量数据迁移； - 数据仓库构建； - 利用 Hive 和 Spark 处理数据，执行 ETL 任务； - 编写 SQL 验证脚本； - 开发和调试 Scala, Python 数据处理脚本； - 解决数据迁移和处理中的技术问题。	
5	数据分析与处理	30

	选手需要知道和理解： - 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。	
	选手应能够： - 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	30

	选手需要知道和理解： - 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。	
	选手需要了解 and 知道： - 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	

3 评分方案

3.1 评分方法

本次竞赛评分由评分裁判线下现场完成评分。如果选手在比赛过程中存在作弊或其他违规行为，裁判组将根据选手的违规情况进行处理，情节严重者取消成绩。

3.2 评分规则

1. 总成绩高者名次在前；
2. 总成绩相同者，按照模块 D、模块 C、模块 B、模块 A 的次序，模块成绩高者名次在前，各模块内容详见本文 4.2。

3.3 评测依据

在赛项设计过程中，将通过评分方案和竞赛赛题来决定标准和评测方法的选择。评测依据，包括但不限于：

- Hadoop 集群搭建的理解程度与管理能力
- HDFS 操作命令的熟练度与准确性
- Flume 数据采集的效率与稳定性
- MapReduce 批处理操作的技术熟练度与执行效率
- Hive 数据仓库的使用技能与数据处理能力
- Spark 分布式计算的实现效率与准确性
- Kafka 消息队列的配置与管理能力
- Sqoop 数据抽取的准确性与效率
- Flink 实时计算的应用能力与性能优化
- Hadoop 常用组件的安装准确性与效率
- Hadoop 集群故障处理的响应速度与问题解决能力

- Hadoop 集群的负载均衡配置与优化效果
- 分析目的的明确性与逻辑性
- 数据整合的技术规范与正确性
- 数据处理的方法与准确性
- 数据分析的深度与透彻性
- 数据可视化的表现力与清晰度
- 重点突出的能力与效果
- 逻辑表达的清晰性与条理性
- 创新点的实用性与原创性
- 对实际问题的理解程度与模型化能力
- 数据探查的深入性与详尽程度
- 数据处理的技术规范与准确性
- 模型训练的效率与有效性
- 诊断与调优的技能与结果改进
- 模型融合的技术应用与效果提升
- 模型预测的准确性与实用性

4 竞赛赛题

4.1 常见注意事项

无论是单个模块或者是一系列独立的或相关联的模块，竞赛赛题可以对标准（Skill Specification）中定义的知识、技能和行为的应用情况进行评测。

结合评分方案，竞赛赛题的目的是为针对标准的评测和评分提供全面的、均衡的及真实的机会。竞赛赛题和评分方案与标准之间的关系将是质量的一个关键指标，就如同标准和实际工作表现的关系一样。

竞赛赛题不包括标准以外的方面，也不影响标准内评分的平衡。

竞赛赛题对知识和理解的评测，仅通过实际工作中对其应用而进行的。

操作过程中需要及时按照答题要求保存相关结果。竞赛结束后，所有设备保持运行状态，评判以最后提交的成果为最终依据。

竞赛完成后，竞赛设备、软件和赛题请保留在座位上，禁止将竞赛所用的所有物品（包括试卷等）带离赛场。

禁止在提交资料上填写与竞赛无关的标记，如违反规定，可视为 0 分。

4.2 竞赛赛题格式/框架

竞赛赛题是四个相对独立和联系的模块组成：

模块 A：大数据

模块 B：数据仓库

模块 C：数据分析

模块 D：机器学习

4.3 竞赛赛题时间分配及分值权重

模块	时长（min）	分值权重（%）
模块 A：大数据	120	20
模块 B：数据仓库	120	20
模块 C：数据分析	120	30
模块 D：机器学习	120	30
合计	480	100

4.4 各模块作业内容及要求

竞赛内容涵盖大数据、数据仓库、数据分析、机器学习等内容，综合考查参赛选手机器学习与大数据综合能力。

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 集群的负载均衡等等。

模块 B 数据仓库：参赛选手需依据任务要求，完成数据仓库的构建和管理。主要内容包括但不限于：数据仓库的搭建、使用 Hive 进行数据查询与分析、使用 Sqoop 进行数据抽取与加载，以及配置与管理 Kafka 和 Flink 实现数据的实时处理。此外，参赛选手还需要展示对数据仓库安全和维护操作的能力。

模块 C 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；并通过可视化展示数据。

模块 D 机器学习：选手需要理解实际问题、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。

模块编号	模块名称	考核范围
A	大数据	包括但不限于：Hadoop 环境搭建、Hadoop 平台基本配置、Hadoop 组件安装、MapReduce 批计算、HDFS 常用操作命令、Yarn 常见管理操作、Hive 数据仓库、Spark 分布式计算、Flume 数据采集、Kafaka 消息队列、Flink 实时计算、Hadoop 组件应用、Hadoop 集群故障处理、Hadoop 集群的负载均衡等。
B	数据仓库	包括但不限于：数据仓库架构设计、数据模型的构建、数据仓库的优化与维护、使用 Hive 进行高级数据分析与查询、Sqoop 数据抽取与数据仓库整合、数据仓库中的数据治理与安全策略、数据仓库的备份与恢复等。
C	数据分析	包括但不限于：数据抽取、数据整合、数据迁移、数据清洗、数据存储、数据探查、数据预处理、数据分析、数据可视化等。
D	机器学习	包括但不限于数据获取、数据探查、数据处理、特征工程、数据建模、模型训练、模型选择、模

		型调参、模型融合、模型验证等。
--	--	-----------------

4.5 竞赛样题公布

竞赛样题将会通过网站（<http://www.brskills.com/jzzy/index.html>）公布。

4.6 竞赛赛题修订

经由技能专家组审核，正式赛题的考点或考试范围与考试样题有 70%相同，本赛项正式赛题不公开。

5 技能管理与沟通

5.1 专家组

技能专家组由首席专家、副首席专家和专家成员组成，负责共同进一步修订本赛项竞赛技术文件以及日常技能管理。

5.2 讨论交流

比赛前如有关于软硬件准备、考试环境部署等相关疑问，参赛者可通过讨论交流群提出反馈。本赛项安排了赛前培训会和赛前领队会，这两个会议将对比赛的规则和纪律进行详细说明，并公布比赛相关的其他重要事宜，也会对竞赛相关问题现场答疑。通过这些活动，参赛者能够获得必要的信息和指导，以确保比赛期间各方面的顺利进行。

6 安全要求

6.1 组织机构

1. 设置比赛安全保障组，组长由赛项执委会主任担任。成员由各赛场安全责任人担任。每一赛场制定一名安全责任人，对本赛场的安全负全责，在发生意外情况时负责调集救援队伍和专业救援人员，安排场内人员疏散。

2. 建立与公安、消防、司法行政、交通、卫生、食品、质检等相关部门的协调机制，保证比赛安全，制定应急预案，及时处置突发事件。设置医护人员、消防人员和保安人员的专线联系，确定对方联系人，由场地安全负责人对口联系。比赛场地布置和器材使用严格依照安全施工条例进行。场地布置划分区域，按安全要求设定疏散通道，并在墙面显著位置张贴安全疏散通道和路线示意图。

6.2 赛项安全管理

1. 比赛设备和设施安装严格按照安全施工标准施工，电源布线、电器安装按规范施工。

2. 按防火安全要求安置灭火器，并指定责任人在紧急时候使用。

3. 赛项竞赛规程中明确国家（或行业）相关职业岗位安全的规范、条例和资格证书要求等内容。

4. 赛项执委会在赛前对本赛项全体裁判员、工作人员进行安全培训。根据《中华人民共和国劳动法》等法律法规，建立完善的安全事故防范制度，在赛前对选手进行培训，避免发生人身伤害事故。

5. 赛项执委会将建立专门方案保证比赛命题、赛题保管、发放、回收和评判过程的安全。

6.3 比赛环境安全管理

1. 赛项执委会赛前组织专人对比赛现场、住宿场所和交通保障进行考察，并对安全工作提出明确要求。赛场的布置，赛场内的器材、设备符合国家有关安

BRICS-FS-02-RU_机器学习与大数据_技术描述 TD

全规定。并进行赛场仿真模拟测试，以发现可能出现的问题。承办单位赛前按照赛项执委会要求排除安全隐患。

2. 赛场周围设立警戒线，防止无关人员进入，发生意外事件。比赛现场内参照相关职业岗位要求为选手提供必要的劳动保护。在具有危险性的操作环节，比赛前裁判员要检查、确认设备正常，比赛过程中严防选手出现错误操作。

为了确保本次大赛的顺利进行，承办学院建立大赛期间相应的安全保障制度，同时由安全保卫、校园环境及卫生医疗保障组执行。

（1）比赛期间所有进入赛区车辆、人员需凭证入内，并主动向工作人员出示。

（2）在比赛开始前，选手要认真阅读场地内张贴的《入场须知》和应急疏散图。

（3）赛场由裁判员监督完成用电设备通电前的检查全过程，对出现的操作隐患及时提醒和制止。

（4）每台竞赛设备使用独立的电源，保障安全。使用选手在进行计算机编程时要及时存盘，避免突然停电造成数据丢失。

（5）比赛过程中，参赛选手应严格遵守安全操作规程，遇有紧急情况，应立即切断电源，在工作人员安排下有序退场。

（6）各类人员须严格遵守赛场规则，严禁携带比赛禁止的物品入内。

（7）安保人员发现安全隐患及时通报赛场负责人员。

（8）比赛场馆严禁吸烟，安保人员不得将证件转借他人。

（9）如果出现安全问题，在安保人员指挥下，迅速按紧急疏散路线撤离现场。

4. 赛项执委会会同承办单位在赛场人员密集、车流人流交错的区域，除了设置齐全的指示标志、增加引导人员外，还需开辟备用通道。

5. 大赛期间，赛项承办单位在赛场管理的关键岗位，增加力量，并建立安全管理日志。

6. 在参赛选手进入赛位，赛项裁判工作人员进入工作场所时，赛项承办单位须提醒、督促参赛选手、赛项裁判工作人员严禁携带通讯、照相摄录设备，禁止携带未经许可的记录用具，并安检设备，对进入赛场重要区域的人员进行安检。

6.4 生活条件保障

比赛期间，由赛事承办单位统一安排参赛选手和指导教师食宿。承办单位须尊重少数民族参赛人员的宗教信仰及文化习俗，根据国家相关的民族、宗教政策，安排好少数民族参赛选手和教师的饮食起居。

比赛期间安排的住宿地要求具有宾馆、住宿经营许可资质。

大赛期间有组织的参观和观摩活动的交通安全由赛区执委会负责。赛项执委会和承办单位须保证比赛期间选手、指导教师和裁判员、工作人员的交通安全。

除必要的安全隔离措施外，严格遵守国家相关法律法规，保护个人隐私和人身自由。

6.5 参赛队职责

各参赛单位在组织参赛队时，须安排为参赛选手购买大赛期间的人身意外伤害保险。

各单位参赛队组成后，须制定相关管理制度，并对所有参赛选手、指导教师进行安全教育。

各参赛队伍须加强参与比赛人员的安全管理，并与赛场安全管理对接。

参赛队如有车辆，一律凭大赛执委会核发的证件出入赛场，并按指定线路行驶，按指定地点停放。

6.6 应急处理

比赛期间发生意外事故时，发现者应第一时间报告赛项执委会，同时采取措

BRICS-FS-02-RU_机器学习与大数据_技术描述 TD

施，避免事态扩大。赛项执委会应立即启动预案予以解决。出现重大安全问题的赛项由赛区执委会决定是否停赛。事后，赛区执委会应出具详细报告情况。

6.7 处罚措施

1. 赛项出现重大安全事故的，停止承办单位的赛项承办资格。
2. 因参赛队伍原因造成重大安全事故的，取消其评奖资格。
3. 参赛队伍发生重大安全事故隐患，经赛场工作人员提示、警告无效的，取消其继续比赛的资格。
4. 赛事工作人员违规的，按照相应的制度追究责任。情节恶劣并造成重大安全事故的，由司法机关追究相应法律责任。

7 材料和设备

7.1 基础设施列表

基础设施清单详细列出了承办方需准备的所有设备和设施，见“2024 金砖国家职业技能大赛线下竞赛机器学习与大数据基础设施清单”。

7.2 参赛选手的工具箱

参赛者由竞赛组织者提供

7.3 竞赛设备清单

7.3.1 技术平台

序号	平台名称	数量
1	在线竞赛云计算平台	1

7.3.2 规格参数

名称	参数规格
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. 9. x

7.4 在技能区域内禁止使用的材料和设备

参赛者携带的任何材料和设备应向专家申报（出示）。专家可禁止使用与执行任务无关或可能给竞争对手带来不公平优势的任何物品。

7.5 建议的比赛区域和工作站布局



8 技能特定的规则

技能特定的规则不能与比赛规则相矛盾或优先于比赛规则。它们将提供不同方面的具体细节和清楚说明，这些方面因技能竞赛而异。它们包括但不限于个人计算设备、数据存储设备、互联网访问、工作程序以及文档管理和分发。

专题/任务	技能专用的规则
使用技术：个人笔记本电脑、平板电脑和手机	专家和口译人员可以使用个人笔记本电脑、平板电脑和手机。 参赛者不得将个人笔记本电脑、平板电脑或手机带入场地。
使用技术 - 个人相机	只有在竞赛赛题完成后或经首席专家同意后，参赛者、专家和口译人员才可以在场地使用个人拍照和录像设备。
互联网访问	参赛者在比赛过程中禁止使用互联网查询资料。
客户端切出	参赛者在比赛过程中未经裁判允许不得切出比赛客户端界面。
参赛者在完成任务期间出现技术问题	如果在竞赛赛题的实施过程中出现技术问题（不是由于参赛者的过错），参赛者将获得额外的时间，该时间等于从发现缺陷到完全消除缺陷的时间。 如果发现技术问题是由于参赛者的过错引起的，参赛者将不会获得额外的时间。



2024金砖国家职业技能大赛 (金砖国家未来技能和技术挑战赛)

