# 工业和信息化人才培养工程培训课程标准

## 大数据分析师

（试行版）



工业和信息化部教育与考试中心

二〇二一年十二月

# 说　明

为推动工业和信息化人才培养工程高质量发展，加快培养大批高素质高技术技能人才，工业和信息化部教育与考试中心依据当前软件、制造业行业人才发展实际需要，积极整合行业教育资源优势，组织行业专家、教育专家持续研发《工业和信息化人才培养工程培训课程标准》（以下简称“标准”），用于指导工业和信息化人才培养工程相关培训课程建设。

《标准》以客观反映现阶段行业的水平和对从业人员的要求为目标，在充分考虑经济发展、科技进步和产业结构变化对本课程影响的基础上，对本课程的等级、培训内容要求、技能要求、知识水平和考核权重都作了明确说明。

《标准》的组编遵循了有关技术规程的要求，既保证了《标准》体例的规范化，又体现了以专业活动为导向、以专业技术技能为核心的特点，同时也使其具有根据科技发展进行调整的灵活性和实用性，符合培训工作的需要。

《标准》编制工作由工业和信息化部教育与考试中心具体组织实施。参与标准编制单位有北京大学、北京理工大学、中国科学院深圳先进技术研究院、承德石油高等专科学校、北京盛久盈天科技有限公司、上海栀博教育科技有限公司、北京东方金信科技有限公司、北京大唐高鸿数据网络技术股份有限公司、国信优易数据有限公司。参与编制人有王慧、严冬宇、王腾蛟、牛振东、喻之斌、陈薇、孔小利、刘敏、刘国文、李悦、童金浩、王伟哲、李耀华、王圣魁。龚玉涵和严冬宇完成汇编与校稿工作。

# 工业和信息化人才培养工程

## 培训课程标准

### 1　课程概况

#### 1.1　课程名称

大数据分析师

#### 1.2　课程定义

本课程面向信息技术行业数据分析从业人员，培养其精通大数据分析方法和大数据分析工具，能从业务理解、数据理解、数据准备、建立模型、模型评估、模型优化等多个操作环节中挖掘数据潜在价值，能够帮助企业更清晰地了解内部现状和外部竞争环境、了解目标客户，从而做出风险评判和决策，提高企业盈利的能力。

#### 1.3　课程等级

本课程共设三个等级，分别为：初级、中级、高级。

#### 1.4　能力要求

具有较强的学习能力、研究分析能力；具有一定的理解、判断和表达能力；具有较强的分析解决问题的能力和沟通能力。

#### 1.5　普通受教育程度

高中及以上文化程度（或同等学历）。

#### 1.6　课程培训要求

##### 1.6.1　培训期限

初级课程不少于线上或线下 80 标准学时；中级课程不少于线上或线下 120 标准学时；高级课程不少于线上或线下 160 标准学时。

##### 1.6.2　培训教师

承担初级、中级理论知识或专业能力培训任务人员，应具有相关课程培训经验 1-3 年。

承担高级理论知识或专业能力培训任务人员，应具有相关课程培训经验 5 年以上，或具有相关职业高级专业技术等级、相关专业高级职称二者之一。

##### 1.6.3　培训场所设备

理论知识培训应有可容纳 30 人以上学员的教室，并配有满足教学需要的网

络环境和学习软件、设施等。培训所需软件：Excel、MySQL、Power BI 等。

## 2　基本要求

### 2.1　专业守则

（1）遵纪守法，爱岗敬业

（2）精益求精，勇于创新

（3）诚实守信，恪守职责

（4）遵守规程，安全操作

（5）认真严谨，忠于职守

### 2.2　基础知识

#### 2.2.1　基础理论知识

（1）操作系统基本知识

（2）计算机网络基本知识

（3）编程基础知识

（4）数据结构与算法基本知识

（5）数据库基本知识

（6）软件工程基本知识

（7）大数据基本知识

#### 2.2.2　技术基础知识

（1）大数据系统环境安装、配置和调试知识

（2）大数据平台架构知识

（3）软件应用开发知识

（4）接口开发与功能模块设计知识

（5）数据采集与数据预处理知识

（6）数据计算与数据指标知识

（7）常用数据分析与挖掘方法

（8）常用数据报表与可视化技术方法

（9）数据管理知识

（10）数据运营及技术指导知识

## 3　课程内容要求

本标准对初级、中级、高级大数据分析师的专业能力要求依次递进，高级别涵盖低级别的要求。

### 3.1　初级

| 课程模块 | 培训内容 | 专业能力要求 | 相关知识要求 |
| --- | --- | --- | --- |
| 1.数据分析基础 | 1.1 数据分析概要 | 1.1.1 能掌握数据分析基本概念<br>1.1.2 能掌握数据分析流程<br>1.1.3 能掌握数据分析应用场景 | 1.1.1 数据分析基本概念<br>1.1.2 数据分析流程<br>1.1.3 数据分析应用场景 |
| | 1.2 Excel 数据分析技术 | 1.2.1 能使用 Excel 获取外部数据<br>1.2.2 能使用 Excel 进行数据处理<br>1.2.3 能使用 Excel 函数进行数据处理<br>1.2.4 能使用 Excel 数据透视表和透视图进行数据统计分析<br>1.2.5 能使用 Excel 进行数据可视化 | 1.2.1 Excel 基本概念<br>1.2.2 获取文本数据<br>1.2.3 从数据库获取数据<br>1.2.4 排序、筛选与分类汇总<br>1.2.5 数组公式<br>1.2.6 日期和时间函数<br>1.2.7 数学函数、统计函数、文本函数、逻辑函数<br>1.2.8 透视表、透视图操作<br>1.2.9 对比分析、趋势分析<br>1.2.10 数据可视化 |
| | 1.3 MySQL 数据库基础 | 1.3.1 能安装配置 MySQL<br>1.3.2 能操作 MySQL 数据库、表、关系等<br>1.3.3 能使用 SQL 语句完成数据的增、删、改、查操作 | 1.3.1 关系型数据库<br>1.3.2 MySQL 安装配置<br>1.3.3 表结构及 DDL<br>1.3.4 数据类型及 DML<br>1.3.5 复杂查询及实际案例 |
| 2.数据可视化分析 | 2.1 认识数据可视化 | 2.1.1 能理解数据可视化基本概念<br>2.1.2 能掌握数据可视化工具 | 2.1.1 数据可视化基本概念<br>2.1.2 数据可视化工具 |
| | 2.2 Power BI 数据分析 | 2.2.1 能使用 Power BI 进行数据获取、集成<br>2.2.2 能使用 Power BI 进行数据清洗、规约、变换<br>2.2.3 能使用 DAX 语言进行数据建模 | 2.2.1 数据获取<br>2.2.2 数据集成<br>2.2.3 数据清洗<br>2.2.4 数据规约<br>2.2.5 数据变换<br>2.2.6 DAX 语言数据建模 |

| | | | |
|---|---|---|---|
| | 2.3 Power BI数据可视化 | 2.3.1 能使用条形、柱状、雷达和漏斗图进行对比分析<br>2.3.2 能使用饼状、环形、瀑布和树状图进行结构分析<br>2.3.3 能使用散点和折线图进行相关分析<br>2.3.4 能使用表、箱线图进行描述性分析<br>2.3.5 能使用仪表、KPI Indicator、子弹图进行KPI分析 | 2.3.1 可视化设计概念<br>2.3.2 对比分析（条形图、柱形图、雷达图、漏斗图）<br>2.3.3 结构分析（饼图、环形图、瀑布图、树状图）<br>2.3.4 相关分析（散点图、折线图）<br>2.3.5 描述性分析（表、箱线图）<br>2.3.6 KPI分析（仪表、KPI Indicator、子弹图） |
| | 2.4 Power BI数据分析报表 | 2.4.1 能完成Power BI数据分析报表<br>2.4.2 能完成数据报表部署 | 2.4.1 Power BI数据分析报表的基本概念、类型、原则、结构<br>2.4.2 数据分析的背景与目的<br>2.4.3 Power BI报表整合 |
| 3.数据分析实战 | 3.1 综合项目实战 | 3.1.1 能用Excel、Power BI等工具连接不同数据源<br>3.1.2 能用Excel、Power BI等工具进行数据预处理<br>3.1.3 能用Excel、Power BI等工具进行业务数据分析<br>3.1.4 能用Excel、Power BI等工具完成最终可视化展示 | 3.1.1 连接数据源<br>3.1.2 数据预处理<br>3.1.3 数据可视化分析<br>3.1.4 数据报表制作与部署 |



## 3.2 中级

| 课程模块 | 培训内容 | 专业能力要求 | 相关知识要求 |
|---|---|---|---|
| 1.数据采集 | 1.1 Python编程基础 | 1.1.1 能掌握Python基本数据类型<br>1.1.2 能掌握Python数据结构<br>1.1.3 能编写Python流程控制<br>1.1.4 能编写自定义函数<br>1.1.5 能编写面向对象的类和对象<br>1.1.6 能读写文件和编写常用操作 | 1.1.1 Python基础知识<br>1.1.2 Python数据结构<br>1.1.3 程序流程控制语句<br>1.1.4 函数<br>1.1.5 面向对象编程<br>1.1.6 文件基础 |
| | 1.2 Python数据采集 | 1.2.1 能使用Python获取静态网页数据<br>1.2.2 能使用Python获取动态网 | 1.2.1 Python爬虫环境与爬虫<br>1.2.2 网页前端基础 |

| | | | |
|---|---|---|---|
| | | 页数据<br>1.2.3 能使用Python进行模拟登录<br>1.2.4 能使用Scrapy框架获取数据 | 1.2.3 静态网页爬取<br>1.2.4 HTTP 请求与解析网页<br>1.2.5 动态网页<br>1.2.6 逆向分析爬取动态网页<br>1.2.7 Selenium 库爬取动态网页<br>1.2.8 登录（表单、Cookie、Selenium）<br>1.2.9 Scrapy 爬虫 |
| 2.大数据分析与挖掘 | 2.1 数据处理 | 2.1.1 能使用Python识别与处理数据异常值、缺失值和重复值<br>2.1.2 能使用Python完成数据变换、标准化、离散化等操作<br>2.1.3 能使用Python完成数据维规约、数值规约、特征构造等操作<br>2.1.4 能使用Python完成数据检索、数据排序等操作 | 2.1.1 读写不同数据源数据<br>2.1.2 DataFrame 常用操作<br>2.1.3 转换与处理时间序列数据<br>2.1.4 使用分组聚合进行组内计算<br>2.1.5 创建透视表与交叉表<br>2.1.6 合并数据<br>2.1.7 清洗数据（重复值、异常值、缺失值）<br>2.1.8 标准化数据<br>2.1.9 转换数据 |
| | 2.2 数据分析 | 2.2.1 能使用Python完成数据质量分析、关联分析、特征分析<br>2.2.2 能使用Python进行线性回归、岭回归<br>2.2.3 能使用 Python 进行决策树、逻辑回归、支持向量机、贝叶斯等分类预测<br>2.2.4 能使用Python进行K均值聚类、密度聚类、期望最大化聚类等聚类分析<br>2.2.5 能使用Python进行回归模型、分类模型、聚类模型的模型评估及参数调优 | 2.2.1 数据质量分析、关联分析、特征分析<br>2.2.2 线性回归、岭回归等回归预测<br>2.2.3 决策树、逻辑回归、支持向量机、贝叶斯等分类预测<br>2.2.4 K 均值聚类、密度聚类、期望最大化聚类等聚类分析<br>2.2.5 回归模型、分类模型、聚类模型的模型评估及参数调优 |
| | 2.3 数据可视化 | 2.3.1 能使用 Matplotlib、Seaborn、pyecharts 等绘图模块或工具绘制柱状图、散点图、饼图<br>2.3.2 能根据业务需求调整绘图参数<br>2.3.3 能根据业务需求和数据可视化结果，撰写相应的数据分析报 | 2.3.1 Matplotlib 数据可视化基础<br>2.3.2 Matplotlib 绘图基础语法与常用参数<br>2.3.3 常见图表类型实现（柱状图、散点图、饼图、箱线图） |

| | | | |
|---|---|---|---|
| | | 告 | 2.3.4 Seaborn 数据可视化进阶<br>2.3.5 Seaborn 绘图基础<br>2.3.6 pyecharts 交互式数据可视化<br>2.3.7 pyecharts 绘图逻辑 |
| 3.数据分析实战 | 3.1 综合项目实战 | 3.1.1 能用Python连接不同数据源<br>3.1.2 能用Python进行数据预处理<br>3.1.3 能用Python进行业务数据建模<br>3.1.4 能用Python完成最终可视化展示 | 3.1.1 数据建模<br>3.1.2 模型调优 |

### 3.3 高级

| 课程模块 | 培训内容 | 专业能力要求 | 相关知识要求 |
|---|---|---|---|
| 1.平台管理 | 1.1 软件安装 | 1.1.1 能实现 Linux 系统集群搭建与基础配置<br>1.1.2 能配置Hadoop相关组件或框架<br>1.1.3 能配置Python集成开发环境 | 1.1.1 Linux 系统集群搭建与基础配置<br>1.1.2 Hadoop 相关组件或框架配置<br>1.1.3 Python 集成开发环境配置 |
| | 1.2 架构管理 | 1.2.1 能实现架构选型<br>1.2.2 能实现架构设计与优化<br>1.2.3 能实现大数据平台到业务系统的端到端解决方案 | 1.2.1 架构选型<br>1.2.2 架构设计与优化<br>1.2.3 大数据平台到业务系统的端到端解决方案 |
| 2.大数据分析与挖掘 | 2.1 数据处理 | 2.1.1 能根据业务需求基于Python 进行数据清洗、变换、合并、校验、特征分析等操作<br>2.1.2 能根据业务需求基于Python 正则表达式处理文本<br>2.1.3 能根据业务需求基于Python 实现文本分词、去停用词、词性标注与命名实体识别<br>2.1.4 能根据业务需求基于Python 实现文本向量化、文本特征计算、文本特征标准化<br>2.1.5 能根据业务需求基于Python 完成网络在线语料库获取 | 2.1.1 数据预处理（数据清洗、变换、合并、校验、特征分析等）<br>2.1.2 正则表达式<br>2.1.3 中文自然语言处理基础<br>2.1.4 文本分词<br>2.1.5 去停用词<br>2.1.6 词性标注<br>2.1.7 命名实体识别<br>2.1.8 文本向量化<br>2.1.9 文本特征计算<br>2.1.10 文本特征标准化 |

| | | | |
|---|---|---|---|
| | | | 2.1.11 网络在线语料库获取 |
| | 2.2 模型构建 | 2.2.1 能掌握线性模型、神经网络等分类与回归算法原理，并根据业务需求基于 Python 构建相应模型<br>2.2.2 能掌握 K 均值聚类、密度聚类等聚类算法原理，并根据业务需求基于 Python 构建相应模型<br>2.2.3 能掌握关联规则的算法原理，并根据业务需求基于 Python 构建相应模型<br>2.2.4 能掌握智能推荐、时序模式等算法原理，并根据业务需求基于 Python 构建相应模型<br>2.2.5 能根据业务需求基于 Python 完成文本分类、聚类等文本挖掘任务<br>2.2.6 能根据业务需求基于 Python 的词典、主题模型等方法完成文本情感分析任务 | 2.2.1 线性模型、神经网络等分类与回归算法原理与 Python 实现<br>2.2.2 K 均值聚类、密度聚类等聚类算法原理与 Python 实现<br>2.2.3 关联规则的算法原理与 Python 实现<br>2.2.4 智能推荐、时序模式等算法原理与 Python 实现<br>2.2.5 Python 实现文本分类<br>2.2.6 Python 实现文本聚类<br>2.2.7 基于 Python 的词典、主题模型的文本情感分析实现 |
| | 2.3 模型优化与部署 | 2.3.1 能使用 Python 实现模型评价与优化<br>2.3.2 能够实现模型部署<br>2.3.3 能够实现模型维护 | 2.3.1 模型评价<br>2.3.2 模型优化 |
| 3.深度学习实战 | 3.1 数据处理 | 3.1.1 能使用 Python 实现语音、图像、文本等数据的探索分析<br>3.1.2 能使用 Python 实现语音、图像、文本等数据的清洗<br>3.1.3 能使用 Python 实现语音、图像、文本等数据的标准化<br>3.1.4 能使用 Python 实现语音、图像、文本等数据的特征选择与构造 | 3.1.1 语音、图像、文本等数据的探索分析<br>3.1.2 语音、图像、文本等数据清洗<br>3.1.3 语音、图像、文本等数据标准化<br>3.1.4 语音、图像、文本等数据的特征选择与构造 |
| | 3.2 模型构建 | 3.2.1 能熟悉卷积神经网络、循环神经网络、长短时记忆网络等算法流程并能够根据业务需求基于 TensorFlow 实现相应模型构建<br>3.2.2 能使用 TensorFlow 完成图像处理、图像识别、语音识别、自然语言处理等<br>3.2.3 能使用 TensorFlow 搭建长 | 3.2.1 深度学习概念<br>3.2.2 常见神经网络原理与实现<br>3.2.3 卷积神经网络<br>3.2.4 循环神经网络<br>3.2.5 长短时记忆网络<br>3.2.6 深度学习常见任务实现 |

| | | | |
|---|---|---|---|
| | | 短时记忆网络模型，完成文本分类、情感分析等<br>3.2.4 能使用TensorFlow搭建序列到序列模型完成语音翻译、中英文翻译等 | 3.2.7 图像处理<br>3.2.8 图像识别、语音识别<br>3.2.9 文本分类<br>3.2.10 情感分析<br>3.2.11 语音翻译、中英文翻译 |
| | 3.3 模型优化与部署 | 3.3.1 能使用TensorFlow等深度学习框架实现模型搭建<br>3.3.2 能使用TensorFlow等深度学习框架实现模型评价与优化<br>3.3.3 能实现业务系统的模型部署与维护 | 3.3.1 模型部署<br>3.3.2 模型维护 |
| 4.数据分析实战 | 4.1 综合项目实战 | 4.1.1 能使用TensorFlow连结不同数据源<br>4.1.2 能使用TensorFlow进行数据预处理<br>4.1.3 能使用TensorFlow进行业务数据建模 | 4.1.1 数据建模与模型调优<br>4.1.2 分析文档撰写 |

## 4　考核权重表

### 4.1　理论知识权重表

| 级别 / 课程模块 | | 初级（%） | 中级（%） | 高级（%） |
|---|---|---|---|---|
| 基本要求 | 专业道德 | 5 | 5 | 5 |
| | 基础知识 | 15 | 10 | 10 |
| 理论知识要求 | 数据分析基础 | 30 | - | - |
| | 数据可视化分析 | 30 | - | - |
| | 数据采集 | - | 30 | - |
| | 大数据分析与挖掘 | - | 30 | 40 |
| | 平台管理 | - | - | 20 |
| | 深度学习实战 | | | 15 |
| | 数据分析实战 | 20 | 25 | 10 |
| 合计 | | 100 | 100 | 100 |



### 4.2　实操能力权重表

| 级别 / 课程模块 | | 初级（%） | 中级（%） | 高级（%） |
|---|---|---|---|---|
| 实操能力要求 | 数据分析基础 | 30 | - | - |
| | 数据可视化分析 | 30 | - | - |
| | 数据采集 | - | 25 | - |
| | 大数据分析与挖掘 | - | 30 | 35 |
| | 平台管理 | - | - | 15 |
| | 深度学习实战 | - | - | 25 |
| | 数据分析实战 | 40 | 45 | 25 |
| 合计 | | 100 | 100 | 100 |

## 附录

### 1 术语和定义

国家、行业标准界定的以及下列术语和定义适用于本文件。

(1) **数据** data

信息的可再解释的形式化表示，以适用于通信、解释或处理。

[GB/T5271.1-2000, 定义 01.01.02]

(2) **大数据** big data

具有体量巨大、来源多样、生成极快、且多变等特征并且难以用传统数据体系结构有效处理的包含大量数据集的数据。

[GB/T 35295-2017, 定义 2.1.1]

(3) **关系数据库** relational database

数据按关系模型来组织的数据库。

[GB/T5271.17-2010, 定义 17.04.05]

(4) **机器学习** machine learning

功能单位通过获取新知识或技能，或通过整理已有的知识或技能来改进其性能的过程。

[GB/T5271.31-2006, 定义 31.01.02]

(5) **数据处理** data processing

数据操作的系统执行。

[GB/T5271.1-2000, 定义 01.01.06]

(6) **数据管理** data management

在数据处理系统中，提供对数据的访问、执行或监视数据的存储，以及控制输入输出操作等功能。

[GB/T5271.1-2000, 定义 01.08.02]

(7) **分析** analytics

根据信息合成知识的过程。

[GB/T 35295-2017, 定义 2.1.48]

(8) **数据挖掘** data mining

从大量的数据中通过算法搜索隐藏于其中信息的过程。

[GB/T 33745-2017, 定义 2.5.3]

(9) **可视化（用于计算机图形）** visualization (in computer graphics)

为帮助人们理解，采用计算机图形和图像处理技术来表现各个过程或对象的模型或特性的做法。

[GB/T 5271.13-2008, 定义 13.01.07]

(10) **操作系统** operating system

控制程序执行的软件，它能提供诸如资源分配、目录调度、输入输出控制及数据管理的服务。

[GB/T5271.1-2000, 定义 01.04.08]

(11) **算法** algorithm

为解决问题严格定义的有限的有序规则集。

[GB/T5271.1-2000, 定义 01.05.05]

(12) **深度学习** deep learning

深度学习是机器学习的分支，是一种以人工神经网络为架构，对数据进行表征学习的算法。

(13) **自然语言** natural language

一种其规则是基于当前的用法且无需特别规定的语言。

[GB/T5271.1-2000, 定义 01.05.08]

## 2 参考文献

[1] GB/T 35589-2017《信息技术 大数据 技术参考模型》相关知识

[2] GB/T 35295-2017 《信息技术 大数据 术语》相关知识

[3] GB/T 38673-2020《信息技术 大数据 大数据系统基本要求》相关知识

[4] GB/T 37721-2019《信息技术 大数据分析系统功能要求》相关知识

[5] GB/T 37722-2019 《信息技术 大数据存储与处理系统功能要求》相关知识

[6] GB/T 36073-2018 《数据管理能力成熟度评估模型》相关知识