上海市高等学校信息技术水平考试
数据科学技术及应用（二三级）
考试大纲（2025 年版）

一、考试性质

上海市高等学校信息技术水平考试是上海市全市高校统一的教学考试，是检测和评价高校信息技术基础教学水平和教学质量的重要依据之一。该项考试旨在规范和加强上海高校的信息技术基础教学工作，提高学生的信息技术应用能力。考试对象主要是上海市高等学校在校学生。考试每年举行一次，通常安排在当年的十一月底的星期六或星期日。凡考试成绩达到合格者或优秀者，由上海市教育委员会颁发相应的证书。

本考试由上海市教育委员会统一领导，聘请有关专家组成考试委员会，委托上海市教育考试院组织实施。

二、考试目标

考试的目标是考核学生对数据科学工作流程的理解，应用统计分析、可视化分析、建模分析等方法对数据进行处理，发现有价值信息的综合能力。考试内容涵盖相关理论知识掌和基本方法的应用实践，要求学生具有应用统计分析和机器学习方法解决数据科学实际问题、提出解决方案和决策建议的能力。

三、考试内容和要求

按照数据科学技术及应用领域知识分布和对应用能力的要求，将考试内容分为基础模块和综合模块。

基础模块考核数据处理的基本方法、数据汇总与统计、数据可视化以及机器学习数据分析的基本原理等，要求学生具备数据科学思维能力、数据组织与分析能力、基于机器学习的数据分析思维，能够使用工具软件或编程语言完成应用分析过程。

综合模块考核基于机器学习、深度学习方法分析数据的原理，多种模态数据表示分析的方法等，要求学生具备机器学习建模分析能力，模态数据分析处理能力、领域应用方案及原型设计能力，持续学习能力，能够使用编程语言完成综合应用分析过程。

模块	知识领域	知识单元	知识点	要求
基础模块	数据科学概述	数据处理的工作流程	问题描述，数据准备、数据探索、建模预测、结果可视化	理解
		大数据技术	大数据 4V 特性、大数据技术	了解
		数据分析工具	无代码数据分析工具	理解
	数据与运算	数据运算	基本算术运算、函数运算、随机生成函数	掌握
	数据汇总与统计	数据文件读写	CSV、txt、Excel 文件读写/打开与保存	掌握
		数据清洗和规整化	缺失、重复数据处理、数据合并、数据排序	掌握
		统计分析	统计的基本概念、常用统计量的意义	理解
			常用统计函数，算术运算、聚合	掌握
	可视化数据探索	常用可视化分析图形	散点图、柱形图、饼图、直方图、概率密度图、箱形图、折线图	掌握
		绘图	创建图图元设置形对象、绘制图形、绘制子图、	掌握
	机器学习建模分析	人工智能与数据智能	人工智能、数据智能	理解
		机器学习基础	有监督学习、无监督学习	理解
		回归分析	回归分析任务、线性回归的基本原理	理解
		分类分析	分类分析任务	理解
		聚类分析目标	聚类分析目标	掌握
	图文语音与时序数据处理	文本数据	自然语言处理的基本知识	掌握
		数字图像处理	数字图像表示方法	掌握
		时序数据处理	时序数据的特征提取、表示	理解
	生成式数据智能	AIGC 和大模型	AI 生成、大语言模型、编码模型	理解
		生成式 BI	生成式 BI 分析	了解

模块	知识领域	知识单元	知识点	要求
综合模块	数据科学概述	数据分析工具	Python 编程、Anaconda 集成环境	理解
	数据与运算	数据组织	创建、选择数据对象、数据条件筛选	掌握
	数据汇总与统计	数据清洗和规整化	数据组织、索引和筛选	掌握
	机器学习建模分析	机器学习基础	训练集、测试集、划分方法	掌握
		回归分析	线性回归建模、性能分析	掌握
		分类分析	逻辑回归建模，性能分析	理解
		分类分析	决策树建模、性能分析	掌握
		分类分析	朴素贝叶斯建模、性能分析	理解
		分类分析	集成学习建模、性能分析	理解
		K-Means 聚类	K-Means 聚类基本原理	掌握
		聚类分析建模	K-means 建模、性能指标、肘部原理	掌握
	神经网络和深度学习建模分析	神经网络	感知器、前馈神经网络结构	理解
		神经网络	神经网络分类与回归建模	掌握
		深度学习	深度学习基础概念	理解
		深度学习	深度学习建模数据分析	了解
	图文语音与时序数据处理	文本数据	中文文本处理步骤：分词、词性标注、特征提取、语言表示模型	理解
		文本数据数据	文本分类的基本方法	了解
		数字图像处理	图像数据存储	理解
		数字图像处理	卷积神经网络、图像分类	了解
		语音识别	语音数据表示	理解
		语音识别	语音识别的基本方法	了解
		时序数据处理	基于神经网络的时序预测	了解
	案例分析	应用实例数据组织与处理	数据科学应用场景、行业领域数据组织方式、应用实例适应性的探索性分析技术以及数据智能分析方法	综合应用
	创新应用	创新、创意应用	创新、创意应用	理解

备注：

知识与技能的考核要求分为**了解**、**理解**和**掌握**三个层次，其含义分别为：

了解：知道某原理、现象、方法或技术的存在及特点（比如一些复杂原理、新现象、新技术、新工具等）。

理解：懂得某原理、现象、方法或技术的核心知识和使用方法。

掌握：熟知并能运用某原理、方法或技术解决问题。

四、试卷结构

题号	题型	题量	分值 (基础/综合)	考核内容	考核能力
一	单选题	10 题	20（20/0）	数据处理方法 大数据概念 多维数据 表结构数据 汇总统计 可视化分析 人工智能与数据智能 图文声音序列数据处理生成式数据智能	数据组织分析能力 人工智能数据思维能力 人工智能方法处理多模态数据能力 持续学习新技术能力
二	多选题	5 题	10（8/2）	表结构数据 汇总统计 可视化分析 图文声音序列数据处理	数据组织分析能力 人工智能方法处理多模态数据能力 持续学习新技术能力
三	简答题	2 题	16（8/8）	数据汇总与统计 机器学习建模分析	数据科学思维能力
四	分析操作题	4 题	64（64/0）	根据实际问题，应用合理方法实现探索性分析	数据组织分析能力
五	综合应用题	8-10 小题	40（0/40）	根据实际应用产生的数据集和提出的分析目标，选择合理的探索性分析技术以及机器学习常用方法，编写程序，实现分析功能，并对实现结果进行解释说明	数据组织分析能力 机器学习建模分析能力 领域应用方案及原型设计能力 创新应用能力
合 计		**29-31 题**	**150 分**		

五、等第划分

1. 等第：不合格、二级合格、二级优秀、三级合格、三级优秀。
2. 等第划分：累计计算各模块得分，等第划分参考分数线如下，具体的总分分数线由考委会划定。

- 三级优秀：综合模块达到 15 分且总分达到 120 分
- 三级合格：综合模块达到 15 分且总分达到 90 分
- 二级优秀：基础模块达到 80 分或总分达到 90 分
- 二级合格：基础模块达到 60 分或总分达到 70 分
- 不合格：基础模块 59 分或总分 69 分以下

六、相关说明

1. 考试时间：150 分钟。
2. 试卷总分：150 分。
3. 考试方式：考试采用基于网络环境的无纸化上机考试。
4. 考试环境：
   - 上海市高等学校信息技术水平考试通用平台。
   - 编程环境（三级必选）：Anaconda3-2021.05（Python 3.8）以上。可选装 Pycharm、VS Code、PyScripter、Spyder 等教学中考生熟悉使用的编程调试环境。
   - 数据分析软件：Excel 2016、WPS 365 教学版（WPS 表格）、SPSS 19 及以上等工具任选。
5. 建议学时数：48-64 学时，其中实验课不少于 24 学时。
6. 参考教材：

   [1] 宋晖、刘晓强主编. 数据科学技术与应用-基于 python 实现（第 2 版）. 电子工业出版社. 2021 年.

七、题型示例

1. 单选题

【例】建模分析时，通常用于训练的样本数量_____测试集的样本数量。

A. 小于　　B. 等于　　C. 大于　　D. 小于等于

【参考答案】C

【能力目标】理解使用机器学习算法建模，性能评估的基本原则和方法，考核机器学习建模分析能力。

【知识内容】训练集、测试集、划分方法。

2. 多选题

【例】关于饼图的描述，错误的是_____。

A. 描述总体的样本值的构成比

B. 饼图每个扇形表示一类样本占总体的百分比

C. 描述总体的各样本区间的样本数量

D. 饼图反映多个总体取值之间的数量关系

【参考答案】CD

【能力目标】理解可视化图形分析的目标，考核数组分析能力。

【知识内容】常用可视化分析图形。

3. 简答题

【例】请描述自己专业领域某个具体场景所涉及的数据，给出各项数据名称，含义以及数据的类型（连续数值/可选项/文本/图像/视频/声音/时序）等。

【能力目标】从数据分析的角度了解专业领域的数据，组织专业数据，考核数据科学意识。

【知识内容】数据的多样性和数据类型。

4. 分析操作题（可选择任意分析工具或编写 Python 程序实现分析过程）

【例】文件 gold_data.csv 中存放了 2012 年至 2022 年黄金交易数据。在处理了缺失数据后，统计 2022 年金价的平均收盘价（源程序文件 fill_1.py）。具体要求如下：

（1）读取 gold_data.csv 文件中的数据，判断数据中是否有空值，并输出各列的空值情况。

（2）Volume 列缺失的值，使用 Volume 的序列均值进行填充；

（3）计算 2022 年金价的平均收盘价(Close/Last)。

```
#源程序文件fill_1.py
# -*- coding: utf-8 -*-
import pandas as pd
#设置亚洲文字显示宽度
pd.set_option("display.unicode.east_asian_width",True)
pd.set_option("display.unicode.ambiguous_as_wide",True)

#1) 读取gold_data.csv文件中的数据；
data= pd.read_csv('gold_data.csv')
print(data.head())
#2)判断数据中是否有空值，并输出各列的空值情况；

is_NAN = data.【1】.any()
print(is_NAN)
# Volume列缺失的值，使用Volume的均值进行填充；
data['Volume'].【2】
# 将Date列的数据类型转换为datetime
data['Date'] = pd.to_datetime(data['Date'])

#3)计算2022年金价的平均收盘价(Close/Last)；
jj=data.loc[data["Date"]>='01/01/2022','Close/Last']
print("2022年金价的平均收盘价：{:.2f}".format(【3】))
```

【参考答案】

1）读取 gold_data.csv 文件中的数据，判断数据中是否有空值，并输出各列的空值情况。

Excel: 逐个筛选

	A	B	C	D	E	F
1	Date	Close/L	Volume	Open	High	Low
10	10/18/202	1655.5	N/A	1655.5	1655.5	1655.5
12	10/14/202	1672.9	N/A	1672.9	1672.9	1672.9
13	10/13/202	1672.9	N/A	1672.9	1672.9	1672.9
655	03/27/202	1623.9	N/A	1623.9	1623.9	1623.9
993	11/21/201	1225.8	N/A	1225.8	1225.8	1225.8
1038	09/19/201	1202.2	N/A	1202.2	1202.2	1202.2
1726	12/25/201	1075.9	N/A	1075.9	1075.9	1075.9
1747	11/26/201	1069.7	N/A	1069.7	1069.7	1069.7
1805	09/07/201	1121.4	N/A	1121.4	1121.4	1121.4
1851	07/03/201	1163.5	N/A	1163.5	1163.5	1163.5
1880	05/25/201	1204	N/A	1204	1204	1204
1916	04/03/201	1200.9	N/A	1200.9	1200.9	1200.9
1950	02/16/201	1227.1	N/A	1227.1	1227.1	1227.1
1970	01/19/201	1276.9	N/A	1276.9	1276.9	1276.9
1982	01/01/201	1184.1	N/A	1184.1	1184.1	1184.1
1987	12/25/201	1173.5	N/A	1173.5	1173.5	1173.5
2007	11/27/201	1197.5	N/A	1197.5	1197.5	1197.5
2070	09/01/201	1287.4	N/A	1287.4	1287.4	1287.4
2111	07/04/201	1320.6	N/A	1320.6	1320.6	1320.6
2140	05/26/201	1291.9	N/A	1291.9	1291.9	1291.9
2166	04/18/201	1293.9	N/A	1293.9	1293.9	1293.9
2210	02/17/201	1318.6	N/A	1318.6	1318.6	1318.6
2230	01/20/201	1251.9	N/A	1251.9	1251.9	1251.9

Spss: 描述统计

统计

		Date	CloseLast	Volume	Open	High	Low
个案数	有效	2547	2547	2508	2547	2547	2547
	缺失	0	0	39	0	0	0

2）Volume 列缺失的值，使用 Volume 的序列均值进行填充；

Excel: 均值+替换

Date	Close/L	Volume	Open	High	Low
10/28/202	1648.3	186519	1667.2	1670.9	1640.7
10/27/202	1668.8	180599	1668.8	1674.8	1658.5
10/26/202	1669.2	183453	1657.7	1679.4	1653.8
10/25/202	1658	178706	1654.5	1666.8	1641.2
10/24/202	1654.1	167448	1662.9	1675.5	1648
10/21/202	1656.3	265985	1632.4	1663.1	1621.1
10/20/202	1636.8	159797	1634.6	1650.3	1626.3
10/19/202	1634.2	172551	1657.2	1659.8	1632.2
10/18/202	1655.5	182067.67	1655.5	1655.5	1655.5
10/17/202	1664	144374	1649.9	1674.3	1649.1
10/14/202	1672.9	182067.67	1672.9	1672.9	1672.9
10/13/202	1672.9	182067.67	1672.9	1672.9	1672.9
10/12/202	1677.5	127689	1673.4	1685.1	1668
10/11/202	1686	166065	1675.6	1691.3	1667.5
10/10/202	1675.2	152626	1703.4	1707.4	1672.5

SPSS: 替换缺失值

Date	CloseLast	Volume	Open	High	Low	Volume_1
10/28/2022	1648.3	186519	1667.2	1670.9	1640.7	186519.0
10/27/2022	1668.8	180599	1668.8	1674.8	1658.5	180599.0
10/26/2022	1669.2	183453	1657.7	1679.4	1653.8	183453.0
10/25/2022	1658.0	178706	1654.5	1666.8	1641.2	178706.0
10/24/2022	1654.1	167448	1662.9	1675.5	1648.0	167448.0
10/21/2022	1656.3	265985	1632.4	1663.1	1621.1	265985.0
10/20/2022	1636.8	159797	1634.6	1650.3	1626.3	159797.0
10/19/2022	1634.2	172551	1657.2	1659.8	1632.2	172551.0
10/18/2022	1655.5	.	1655.5	1655.5	1655.5	182067.7
10/17/2022	1664.0	144374	1649.9	1674.3	1649.1	144374.0
10/14/2022	1672.9	.	1672.9	1672.9	1672.9	182067.7
10/13/2022	1672.9	.	1672.9	1672.9	1672.9	182067.7

3）计算 2022 年金价的平均收盘价(Close/Last)。

Excel:筛选+subtotal 函数

函数库　　定义的名称

H2　=SUBTOTAL(101,B:B)

	A	B	C	D	E	F	G	H
	Date	Close/La	Volume	Open	High	Low		均值
	10/28/2022	1648.3	186519	1667.2	1670.9	1640.7		1815.003
	10/27/2022	1668.8	180599	1668.8	1674.8	1658.5		
	10/26/2022	1669.2	183453	1657.7	1679.4	1653.8		
	10/25/2022	1658	178706	1654.5	1666.8	1641.2		
	10/24/2022	1654.1	167448	1662.9	1675.5	1648		
	10/21/2022	1656.3	265985	1632.4	1663.1	1621.1		
	10/20/2022	1636.8	159797	1634.6	1650.3	1626.3		
	10/19/2022	1634.2	172551	1657.2	1659.8	1632.2		
	10/18/2022	1655.5	182067.67	1655.5	1655.5	1655.5		
	10/17/2022	1664	144374	1649.9	1674.3	1649.1		
	10/14/2022	1672.9	182067.67	1672.9	1672.9	1672.9		
	10/13/2022	1672.9	182067.67	1672.9	1672.9	1672.9		
	10/12/2022	1677.5	127689	1673.4	1685.1	1668		

Spss：个案选择+分组统计

		CloseLast
		平均值
XDATE.YEAR(Date)=2022 (FILTER)	Not Selected	.
	Selected	1815.0

Python 编程

填空：

【1】　isnull()

【2】　fillna(data["Volume"].mean(), inplace=True)

【3】　jj.mean()

结果截图：

```
         Date  Close/Last    Volume    Open    High     Low
0  10/28/2022      1648.3  186519.0  1667.2  1670.9  1640.7
1  10/27/2022      1668.8  180599.0  1668.8  1674.8  1658.5
2  10/26/2022      1669.2  183453.0  1657.7  1679.4  1653.8
3  10/25/2022      1658.0  178706.0  1654.5  1666.8  1641.2
4  10/24/2022      1654.1  167448.0  1662.9  1675.5  1648.0
Date          False
Close/Last    False
Volume         True
Open          False
High          False
Low           False
dtype: bool
2022年金价的平均收盘价：1815.00
```

【能力目标】掌握缺失数据预处理、以及简单统计分析的能力。

【知识内容】

（1）缺失数据筛选创建多维数组；

（2）缺失数据填充；

（3）样本均值统计。

5. 综合应用题

【例】动物具有多种特征，根据这些特征动物可以分为哺乳类、鸟类、爬行类、鱼类和两栖类等。animals.csv 记录了动物的名字、毛发、蛋、腿、尾巴等多种特征（具体说明见“数据集说明”文件），类型被标记为 Mammal（哺乳动物）、bird（鸟类）、others（其他）三大类。

请根据数据集（animals.csv）文件格式，正确获取数据样本进行预处理，统计分析，建立分类模型；请尝试多种算法，比较分类的性能。具体要求如下：

（1）从文件中读出所需的数据，根据分析需求将所需的数据保存到 DataFrame 中；

（2）数据清洗，判断数据集中是否有缺失数据，并采取合适的方法处理；

（3）统计各类动物的数量，并列出每类动物的名字；

（4）数据预处理，将“type”的值转换为数值类型；

（5）使用散点图矩阵分析动物类型与蛋、尾巴特征间的相关性，并计算他们之间的相关系数；

（6）选择合适的数据列作为特征和标签形成数据集用于训练分类模型，并将分为训练集和测试集；

（7）在训练集上建立分类模型，在测试集上测试模型预测的准确性。 在已学习的分类方法（决策树、支持向量机、神经网络）中试用两种算法；

（8）根据第（7）步的运行结果，说明两种算法在动物分类数据上的性能。请将结果用文字描述在程序文件给出的注释行中。

【能力目标】理解应用场景需求，选择分析方法和技术，实现分析目标，考核领域应用方案及原型设计能力。

【知识内容】

（1）数据汇总统计、可视化分析。

（2）机器学习，有监督学习，神经网络。