美赛各题型必备高频算法汇总表

常见题型	高频算法	基本原理	适合场景	模型检验
数据预处理	缺失值处理-均值/中位数填充	利用数据的统计中心特性，用全部有效数据的均值（适用于正态分布）或中位数（适用于偏态分布）替代缺失值，快速补全数据缺口	各类含缺失数据的赛题；特征：缺失值占比低（＜10%）、数据分布相对均匀、无强关联性的单维度数据	1. 正态性检验：Shapiro-Wilk 检验（判断是否适配均值填充）；2. 填充合理性检验：填充后数据与原始数据分布一致性分析；3. 缺失机制检验：判断数据为随机缺失（MAR）或完全随机缺失（MCAR）
	缺失值处理-K 近邻（KNN）填充	基于欧式距离/曼哈顿距离，筛选缺失值样本的 K 个最相似邻居样本，用邻居样本对应特征的均值/中位数填补缺失值	多维度关联数据赛题；特征：缺失值占比中等（10%-20%）、变量间关联性强、样本量充足	1. K 值合理性检验：通过交叉验证选择最优 K 值；2. 距离度量有效性检验：对比不同距离函数（欧式/曼哈顿）的填充误差；3. 稳定性检验：不同样本子集填充结果的一致性
	异常值处理-3σ 原则	基于正态分布假设，若数据点偏离均值超过 3 倍标准差（即$\|x-\mu\| > 3\sigma$），判定为异常值，可选择剔除或修正	连续型数值数据赛题；特征：数据服从正态分布、异常值为极端离群点、无明显局部异常	1. 正态性检验：Q-Q 图+K-S 检验（验证数据是否符合正态分布）；2. 异常值有效性检验：判断异常值为离群点或数据错误；3. 修正后检验：修正后数据的标准差、均值稳定性
	异常值处理-箱型图法	通过四分位数（Q1、Q2、Q3）计算四分位距 IQR，将超出[Q1-1.5IQR, Q3+1.5IQR]范围的数据判定为异常值，鲁棒性更强	非正态分布/偏态数据赛题；特征：数据存在偏态、含局部极端值、适用于任意数值型数据	1. 偏态性检验：偏度系数（$\|偏度\| > 1$ 为强偏态）；2. 异常值比例检验：控制异常值占比在合理范围（一般＜5%）；3. 鲁棒性验证：对比剔除异常值前后数据统计特征变化

常见题型	高频算法	基本原理	适合场景	模型检验
	数据标准化-Z-score 标准化	将数据转换为均值=0、标准差=1 的标准正态分布，公式为 x'=(x-μ)/σ，保留数据原始分布趋势	机器学习、回归分析、聚类分析前置处理；特征：数据无明确上下限、存在极端值、需消除量纲影响	1. 标准化效果检验：标准化后均值≈0、标准差≈1；2. 分布一致性检验：标准化前后数据分布形态相似度；3. 极端值影响检验：极端值对标准化结果的扰动分析
	数据标准化-最小-最大归一化	通过线性映射将数据压缩至[0,1]区间，公式为 x'=(x-min)/(max-min)，保留数据相对大小关系	多指标综合评价、神经网络输入预处理；特征：数据有明确取值范围、需突出数据相对差异	1. 边界值稳定性检验：验证最大值、最小值是否为异常值（避免边界失真）；2. 归一化一致性检验：不同样本子集归一化结果对比；3. 量纲消除检验：归一化后各指标方差合理性
相关性分析	皮尔逊（Pearson）相关系数	衡量两个连续型变量的线性相关程度，取值范围[-1,1]，绝对值越接近 1 线性相关性越强，需满足双变量正态分布假设	探究变量线性关联的赛题；特征：变量为连续型、数据服从正态分布、无明显非线性趋势	1. 双变量正态性检验：联合正态性 Q-Q 图+K-S 检验；2. 相关性显著性检验：t 检验（$P < 0.05$ 为显著相关）；3. 线性假设检验：残差散点图（验证线性关系）
	斯皮尔曼（Spearman）秩相关系数	将变量值转换为秩次后计算相关度，不依赖数据分布类型，衡量变量间的单调相关关系	非正态/非线性数据关联分析；特征：变量为连续型/有序分类型、数据偏态分布、变量呈单调变化趋势	1. 单调趋势检验：散点图趋势分析；2. 相关性显著性检验：t 检验或卡方检验；3. 秩次合理性检验：变量秩次转换无异常偏差
	肯德尔（Kendall）和谐系数	基于秩次的协同相关分析，核心衡量多个评价主体对同一对象评价结果的一致性程度	多维度评价、主观打分类赛题；特征：存在多个评价者/评价指标、数据为秩次或打分型、需验证评价一致性	1. 一致性检验：卡方检验（验证和谐系数显著性）；2. 评价者信度检验：克朗巴赫α系数（α≥0.7 为高信度）；3. 异常评价检验：识别偏离整体趋势的评价主体

常见题型	高频算法	基本原理	适合场景	模型检验
评价类	层次分析法（AHP）	将复杂评价问题按“目标层-准则层-方案层”逐级分解，通过两两比较构建判断矩阵，计算指标权重并进行一致性检验，将主观定性判断转化为客观定量数据	多专家参与的方案评估、政策实施效果打分、含主观判断的多指标评价（如项目立项优先级评估）；特征：评价指标少（＜10 个）、需结合专家经验、无大量实测数据支撑	1. 一致性检验：CI（一致性指标）+RI（随机一致性指标），CR=CI/RI＜0.1 为通过；2. 专家判断一致性检验：肯德尔和谐系数（W≥0.6）；3. 权重稳定性检验：不同判断矩阵下权重波动分析
	TOPSIS 法	先对评价指标标准化处理，确定正理想解（各指标最优值）和负理想解（各指标最劣值），计算待评方案与两类理想解的欧氏距离，通过相对贴近度排序筛选最优方案，支持主观/客观赋权结合	区域发展水平排名、多方案可行性优选、供应链风险评级、产品质量多维度评价；特征：指标数多、有完整定量数据、需明确方案优劣排序	1. 标准化有效性检验：不同标准化方法（Z-score/归一化）结果对比；2. 权重合理性检验：主观与客观权重融合后的一致性；3. 敏感性检验：指标权重扰动±10%后排名稳定性
	模糊综合评价法	运用模糊集合理论，将“优秀”“良好”等定性指标转化为定量隶属度，构建模糊评判矩阵，结合指标权重计算综合隶属度，量化模糊性评价问题	服务质量评价、环境风险等级评估、模糊指标主导的系统评价（如生态环境质量评级）；特征：评价标准模糊、含定性描述指标、需划分明确的评价等级	1. 隶属度合理性检验：专家打分法验证隶属度矩阵；2. 模糊矩阵一致性检验：一致性比率验证；3. 结果可信度检验：不同评价等级阈值下结果稳定性
	熵权-TOPSIS 组合评价法	先通过熵权法（基于数据信息熵）计算客观权重，避免主观赋权偏差，再结合 TOPSIS 法进行方案排序，兼顾数据客观性与评价合理性	需弱化主观干扰的综合评价赛题；特征：有大量实测数据、指标区分度较强、追求评价结果客观公正	1. 熵权有效性检验：信息熵大小（熵值越小指标区分度越高）；2. 组合效果检验：对比单一熵权法、TOPSIS 法结果相关性；3. 稳定性检验：Bootstrap 抽样（1000 次）验证排名一致性
预测类	ARIMA 模型	由自回归（AR）、差分（I）、移动平均（MA）三部分组成，通过差分运算使非平稳时间序列平稳化，利用自回归捕捉历史数据关联，移动平均修	经济指标预测（如 GDP、物价指数）、季节性商品销量预测、常规时间序列数据的短期/中期预测；特征：大样本（n≥30）、数据为线	1. 平稳性检验：ADF 检验（P＜0.05）+自相关图（ACF）/偏自相关图（PACF）；2. 模型定阶检验：AIC、BIC 准则选择最优 p、q 阶；

常见题型	高频算法	基本原理	适合场景	模型检验
		正随机误差，精准捕捉线性时序趋势与周期性	性时间序列、无剧烈随机波动	3. 残差检验：残差正态性（Shapiro-Wilk）、无自相关性（LM 检验）
	SARIMA 模型	在 ARIMA 基础上加入季节项（S），通过季节差分处理捕捉数据的季节周期性规律，适配含季节波动的时序数据	含季节变化的时序预测赛题（如月度用电量、季度农产品产量）；特征：时间序列数据、存在明显季节/周期波动、样本覆盖至少 2 个完整周期	1. 季节平稳性检验：季节差分后 ADF 检验；2. 季节周期检验：自相关图峰值对应的滞后阶数；3. 精度检验：RMSE、MAE、MAPE（≤10%最优）
	灰色 GM(1,1)模型	针对小样本、贫信息数据，通过累加生成处理弱化原始数据的随机性，构建一阶线性微分方程，拟合数据内在变化趋势，无需大量历史数据，对非线性、不确定性数据适配性强	小众行业产量预测、稀有灾害损失预估、数据匮乏场景的短期预测（如新型产品市场需求预测）；特征：小样本（$n < 20$）、贫信息、数据无明显规律	1. 精度检验：后验差检验（$C < 0.35$、$P > 0.95$）、平均相对误差（≤5%）；2. 残差检验：残差序列随机性（游程检验）；3. 拟合优度检验：可决系数 R^2（≥0.8）
	LSTM 神经网络	属于循环神经网络（RNN）的改进版，通过输入门、遗忘门、输出门的门控机制，选择性记忆和遗忘长序列历史信息，解决传统 RNN 的梯度消失问题，可捕捉非线性时间序列的复杂关联	交通流量时空预测、污染物浓度变化预测、疫情传播趋势模拟、多因素影响的复杂时序预测；特征：高维数据、非线性关联、长序列时序数据、多影响因素	1. 收敛性检验：训练损失曲线（验证是否收敛、无过拟合）；2. 精度检验：RMSE、MAE 对比传统模型；3. 泛化性检验：测试集与训练集精度差异（差异过小避免过拟合）
	BP 神经网络	基于误差反向传播的多层前馈神经网络，通过正向传播计算预测值，反向传播调整神经元权重，拟合输入与输出间的复杂非线性映射关系	多因素非线性预测赛题（如环境质量综合预测、企业营收多维影响预测）；特征：大样本、高维数据、变量间无明确线性关系	1. 过拟合检验：正则化（L1/L2）后精度变化、训练集与测试集误差对比；2. 权重合理性检验：神经元权重分布无极端值；3. 收敛性检验：迭代次数与误差收敛速度
优化类	线性规划（LP）	在目标函数和约束条件均为线性的前提下，求解目标函数的最大值或最小值，适配连续变量优化问题，核心是资源的最优	资源分配优化、生产计划调度、物资运输规划、成本最小化/收益最大化问题；特征：决策变量连续、目标与约束均为线性关系、无复杂非	1. 可行性检验：约束条件是否存在可行解（无矛盾约束）；2. 最优性检验：对偶理论验证最优解；3. 灵敏度检验：约束系数、目标

常见题型	高频算法	基本原理	适合场景	模型检验
		配置	线性交互	函数系数变化对最优解的影响
	整数规划（IP）	在 LP 基础上限制决策变量为整数（纯整数/0-1 整数），适配离散决策变量场景，解决“计数型”优化问题	设施选址规划、设备采购台数优化、任务分配问题、0-1 决策类场景（如是否投资某项目）；特征：决策变量为离散整数、含二元选择或计数约束	1. 可行性检验：整数约束下是否存在可行解；2. 最优性检验：对比松弛问题（去掉整数约束）最优解，验证整数解合理性；3. 边界检验：决策变量取值是否在合理边界内
	粒子群优化（PSO）	模拟鸟群觅食行为，将每个优化解视为“粒子”，所有粒子构成“粒子群”，粒子在解空间中迭代搜索，通过跟踪自身最优解和群体最优解更新位置与速度，无梯度依赖，寻优速度快	工业参数调优、神经网络超参数优化、非线性约束下的多目标优化（如成本-效率平衡）；特征：目标函数复杂、无明确解析解、需快速寻优、多约束条件	1. 收敛性检验：迭代曲线（群体最优值是否稳定）；2. 参数合理性检验：惯性权重、学习因子对寻优效果的影响；3. 鲁棒性检验：不同初始粒子位置下最优解一致性
	遗传算法（GA）	基于生物进化理论，将优化问题编码为“染色体”，通过选择（保留优解）、交叉（基因重组）、变异（随机突变）操作，迭代进化生成最优解，能跳出局部最优，适配复杂非线性优化场景	物流路径规划、复杂组合优化问题、多约束条件下的全局寻优（如供应链网络优化）；特征：非线性、多局部最优解、组合优化或全局寻优需求	1. 收敛性检验：迭代过程中最优解收敛速度与稳定性；2. 遗传算子有效性检验：对比不同交叉率、变异率的寻优效果；3. 全局最优检验：避免陷入局部最优（多次迭代结果对比）
	非线性规划（NLP）	目标函数或约束条件中含非线性项，通过梯度下降、牛顿法等迭代算法求解最优解，适配复杂非线性优化场景	非线性成本优化、参数非线性约束问题（如化工反应参数优化）；特征：目标/约束含非线性关系、需高精度寻优	1. 最优性检验：KKT 条件（验证是否为最优解）；2. 收敛性检验：迭代步长与目标函数值变化（趋于稳定）；3. 初值敏感性检验：不同初始值对最优解的影响
机理分析类	系统动力学（SD）	以反馈控制理论为核心，梳理系统内各变量的因果关系，构建反馈回路和流图，通过微分方程量化变量间的相互作用，动态	城市扩张机理研究、生态系统演化模拟、经济系统反馈机制分析、公共卫生系统运行机理探究；特征：系统复杂、	1. 结构合理性检验：反馈回路有效性、变量因果关系一致性；2. 灵敏度检验：核心变量扰动（±5%~±10%）

常见题型	高频算法	基本原理	适合场景	模型检验
		模拟系统随时间的演化规律，揭示系统内在运行机理	变量间存在因果反馈、需长期动态模拟	对系统输出的影响；3. 拟合度检验：模型输出与实际数据的误差（RMSE、R^2）
	贝叶斯网络（BN）	基于概率图模型的不确定性推理方法，用有向无环图表示变量间的依赖关系，通过先验概率和贝叶斯公式计算后验概率，量化变量间的因果关联，适配信息不完全、存在不确定性的场景	疾病传播因果分析、灾害风险因素关联探究、不确定环境下的机理推理（如污染物扩散成因分析）；特征：信息不完全、变量间存在概率依赖、需因果推理	1. 网络结构检验：BIC 评分（验证结构合理性）、PC 算法结构学习有效性；2. 参数检验：先验分布与后验分布一致性；3. 推理有效性检验：后验概率置信区间（95%）
	元胞自动机（CA）	以网格上的“元胞”为基本单位，每个元胞具有离散状态，通过预设的局部演化规则（依赖自身及邻域元胞状态），模拟宏观系统的时空变化，擅长刻画扩散、增长、演化类机理	城市用地扩张模拟、疫情空间扩散机理、生态种群分布演化、火灾/洪水蔓延过程模拟；特征：时空演化特性、局部规则决定宏观行为、扩散/增长类机理分析	1. 演化规则有效性检验：模拟结果与实际时空分布的相似度；2. 网格敏感性检验：不同网格尺寸对模拟结果的影响；3. 稳定性检验：长期演化后系统是否趋于合理稳态
决策类	贝叶斯决策理论	结合先验信息（历史经验/专家判断）和样本数据，通过贝叶斯公式更新后验概率，计算各决策方案的期望损失，选择期望损失最小的方案，兼顾不确定性和数据支撑，适配风险型决策	不确定环境下的投资决策、灾害应急风险决策、信息不完全的方案选择（如应急物资调度决策）；特征：信息不完全、存在风险概率、需量化决策损失	1. 先验分布合理性检验：对比不同先验分布（均匀/正态）的后验结果；2. 风险损失检验：期望损失计算的准确性、损失函数合理性；3. 决策稳健性检验：不同概率场景下最优方案一致性
	多准则决策（MCDM）	整合多个相互冲突的决策准则（如成本、效率、风险），通过权重确定、方案排序、折中优化等步骤，将多目标决策问题转化为可量化分析的问题，常融合 AHP、TOPSIS 等方法使用	多目标方案优选、区域发展规划决策、兼顾多方利益的复杂决策（如公共政策制定）；特征：多冲突准则、需平衡多方需求、决策目标多元化	1. 准则一致性检验：冲突准则的折中合理性；2. 权重有效性检验：不同赋权方法对决策结果的影响；3. 敏感性检验：准则权重扰动后的方案排序变化
	动态规划（DP）	将多阶段决策问题分解为若干个相互关联的子决策问题，利用“最优子	多阶段资源分配决策、生产调度的动态优化、路径规划的分段决策	1. 最优子结构检验：各阶段最优解是否构成全局最优解；2. 无

常见题型	高频算法	基本原理	适合场景	模型检验
		结构”和“无后效性”特性，逐阶段求解子问题最优解，最终整合得到全局最优决策，适配时序性、阶段性决策场景	（如多时段物流调度）；特征：决策分阶段进行、各阶段相互关联、需全局最优解	后效性检验：后续阶段决策不受前期状态影响；3. 边界条件检验：初始与终止状态设定的合理性
分类聚类类	K-Means 聚类	无监督聚类算法，先人为设定聚类数 K，随机初始化 K 个质心，通过迭代计算每个样本到各质心的距离，将样本归为距离最近的质心簇，更新质心位置直至收敛，实现数据的自动分群	用户画像分群、数据异常检测（噪声簇识别）、区域特征聚类、无标签数据的快速分群；特征：大样本、数据特征清晰、簇内密度均匀、适用于球形簇	1. K 值确定检验：肘部法则（Elbow Method）+轮廓系数（Silhouette Score）；2. 聚类有效性检验：Calinski-Harabasz 指数（越大越好）；3. 稳定性检验：不同初始质心下聚类结果一致性
	系统聚类（层次聚类）	无需指定簇数，从单个样本为一簇开始，逐步合并（凝聚式）或拆分（分裂式）相似度高的簇，形成聚类树（树状图），可灵活确定最优分类数	小样本、未知分类数的聚类赛题（如稀有样本特征分群）；特征：样本量少（n < 100）、需分析分类层次关系、适用于任意簇型	1. 距离度量检验：对比不同距离（欧式/曼哈顿）、连接方式（Ward/平均链接）的聚类效果；2. 簇数合理性检验：树状图切割点有效性；3. 聚类纯度检验：簇内样本相似度、簇间样本差异性
	DBSCAN 聚类	基于密度的无监督聚类算法，通过“核心点-边界点-噪声点”的定义，将密度相连的样本划分为一簇，自动识别噪声点，无需预设簇数	含噪声、非球形簇的聚类赛题（如不规则分布的地理数据分群）；特征：数据分布不规则、存在噪声点、无需指定簇数	1. 参数合理性检验：Eps 和 MinPts 参数通过 K-距离图确定；2. 聚类有效性检验：轮廓系数、噪声点比例（控制在合理范围）；3. 密度适配性检验：不同密度区域聚类效果一致性
	SVM（支持向量机）	监督分类算法，通过核函数将低维非线性数据映射到高维线性空间，寻找最优分类超平面，最大化两类样本的间隔（Margin），仅依赖边界样本（支持向量）分类，泛化能力强	高维数据分类（如图像、文本分类）、样本类别不平衡的分类问题、小样本下的精准分类（如稀有疾病诊断数据分类）；特征：小样本、高维数据、非线性分类、类别不平衡	1. 核函数选择检验：对比线性/高斯（RBF）核函数的分类准确率；2. 参数优化检验：通过交叉验证确定 C（惩罚系数）、γ（核函数参数）；3. 分类有效性检验：混淆矩阵、AUC 值、F1 分数

常见题型	高频算法	基本原理	适合场景	模型检验
	决策树	监督分类算法，基于信息增益、基尼系数等指标选择最优特征，按特征阈值逐级划分样本，构建树状决策规则（根节点-内部节点-叶节点），可解释性极强，无需数据标准化处理	样本类别识别、风险等级分类、简单规则导向的分类问题（如客户信用评级、故障类型判断）；特征：需明确决策规则、可解释性要求高、数据类型混合（定量+定性）	1. 过拟合检验：剪枝（预剪枝/后剪枝）前后准确率对比；2. 特征重要性检验：决策树特征重要性排序合理性；3. 分类效果检验：混淆矩阵、准确率、召回率
	随机森林	基于决策树的集成学习算法，通过 Bootstrap 抽样构建多个决策树，采用投票机制确定最终分类结果，降低过拟合风险，提升分类准确率和鲁棒性	复杂数据分类、高维数据降维分类、含噪声数据的分类（如环境监测数据异常分类）；特征：大样本、含噪声数据、需高准确率分类、抗过拟合需求	1. 集成效果检验：对比单棵决策树与随机森林的准确率、泛化能力；2. 参数优化检验：决策树数量、特征采样比例对结果的影响；3. 稳定性检验：不同样本子集下分类结果一致性