《 机 器 学 习 课 程 设 计 》

课 程 论 文

题　　目：基于机器学习的睡眠效率回归预测模型探究

学生姓名：曹梓上

2025 年 3 月 18 日

摘 要

睡眠效率作为睡眠医学研究的核心指标，其影响机制与干预策略始终是学界关注的焦点。机器学习技术凭借其在数据处理和模式识别方面的优势，已成为回归预测的有效工具。基于此，本研究致力于构建并评估**基于机器学习的睡眠效率回归预测模型**。

本研究选取 Kaggle 官网睡眠效率数据集，由 **472** 条数据和 **14** 个属性变量组成。旨在使用四种常见的分类算法（决策树回归、支持向量回归、装袋法及随机森林回归）对睡眠效率进行预测，对其性能进行比较分析。同时，探究年龄（Age）、睡眠时长（Sleep duration）等 **12 个特征变量**与目标变量**睡眠效率**（Sleep efficiency）的关系。

首先，在对数据集进行必要的缺失值、异常值检测处理后，进行数据预处理包括特征重构、独热编码、标签编码以及标准化处理，其中，独热编码主要对性别（Gender）、是否吸烟（Smoking status）以及**重构**后的是否深夜入睡（Night_sleep）三个类别变量进行转换。以 **7:3** 比例划分训练集和测试集后，输入标准化后的训练集对四种模型正式进行训练，并通过超参数调优，优化各个模型的性能。最终，基于均方误差（MSE）、均方根误差（RMSE）、平均绝对误差（MAE）、决定系数（R^2）四种指标，评估模型的对测试集的预测效果，采用一系列常用指标进行性能评价。

实验结果表明，装袋法（n_estimators=100，max_samples=0.5）与随机森林（n_estimators=100，max_depth=10）均**显著优于**单一模型，后者因特征随机性进一步降低方差，最终 **R^2=0.8786**，**MSE=0.0023**，预测误差较装袋法降低 4.2%。SVR 的预测误差最大，而决策树模型的 R^2 最小。随机森林模型显示，夜间觉醒次数（重要性得分 0.35）、浅睡眠比例（0.25）及咖啡因摄入量（0.15）为关键影响因素，与可视化结论一致。此外，睡眠是否在深夜时段（编码变量）对效率影响较弱（p=0.646），性别与年龄无显著关联（p>0.83）。

综合考虑模型的准确性和可解释性，**随机森林**是**最适合**本研究任务的模型，其不仅提供了较高的预测准确率，且可解释性强，实际应用效果良好。

关键词：睡眠效率、SVR、决策树回归、支持向量回归、装袋法、随机森林回归

目录

第1章 绪论

本章主要介绍基于机器学习的睡眠效率回归预测模型探究的研究背景、主要研究内容、应用价值以及引用文献的综述。

1.1 研究背景

睡眠效率作为睡眠医学研究的核心指标，其影响机制与干预策略始终是学界关注的焦点[12]。《健康中国行动（2019-2030）》数据显示，我国居民睡眠指数逐年下降，约 38%人群存在入睡困难、夜间觉醒等问题。睡眠效率降低不仅影响认知功能与情绪调节，还与心血管疾病、代谢紊乱等慢性病风险升高显著相关[13]。

技术手段革新中，多导睡眠图（PSG）仍是金标准[10]，但可穿戴设备（如智能手环、EEG 头带）的普及推动了居家监测的常态化，结合人工智能算法（如深度学习）实现了睡眠分期与异常事件的自动化分析[18]。

值得关注的是，我国中老年群体睡眠障碍问题尤为突出[16]，城乡差异与生活方式变迁加剧了这一矛盾。《中国睡眠研究报告 2024》指出，农村地区吸烟率高达 38.7%，且睡眠监测设备覆盖率不足城市的 1/3。

本研究立足各类人群特征，结合机器学习算法，旨在揭示各类因素对睡眠效率的非线性影响机制，为制定精准睡眠健康策略提供科学依据。

1.2 主要内容和应用价值

1.2.1 主要研究内容

本研究以构建基于机器学习的睡眠效率回归预测模型为核心任务。所采用的数据集来源于 Kaggle 官网睡眠效率数据集，此数据集由 452 条数据和 14 个特征属性组成。如患者的睡眠状态数据（睡眠时长、深浅睡眠时长、快速眼动比例等）、生活习惯数据（是否深夜睡、是否抽烟、喝酒量等）、个人属性（性别、年龄等）相关的特征属性。

通过对这些数据进行全面深入的分析，挖掘各特征与睡眠效率之间的内在联系。在此基础上，运用四种机器学习算法（决策树回归、支持向量回归、装袋法以及随机森林回归）构建回归预测模型，并对模型进行优化和评估，以实现睡眠效率的准确预测，为诊断睡眠状况、调整生活质量提供科学依据。

1.2.2 应用价值

睡眠效率预测模型的构建与应用具有多层次的实践意义。在临床领域，该模型可辅助医生识别睡眠障碍高危人群，通过量化生活习惯与环境因素对睡眠质量的贡献度，为制定个性化干预方案（如调整作息时间、优化卧室环境）提供科学依据。在公共卫生层面，模型输出的群体睡眠效率特征可用于指导城市噪声管控、社区照明设计等政策优化，降低环境因素对居民睡眠的负面影响。

此外，结合可穿戴设备与智能家居系统，模型可实时监测用户睡眠状态并动态调节室内温湿度、光线强度，形成“监测-分析-干预”闭环，显著提升睡眠健康管理的主动性与智能化水平。从技术转化视角，研究通过对比不同机器学习算法的性能边界，为低成本、高精度的家用睡眠监测设备开发提供算法优选方案，推动人工智能在睡眠医学中的产业化落地，同时为心理学、职业健康等领域探究睡眠效率与认知功能、工作效率的关联机制奠定方法学基础。

1.3 文献综述

当前聚焦预测睡眠效率的研究较少，主要是对于睡眠障碍引起的相关疾病以及特定人群（尤其是中老年人）进行探究。

1.3.1 睡眠疾病方面

汪蝶,吴帮云,谭存瑶等人（2025）探讨中心性肥胖在 SE 与血脂关联间的中介作用，为探讨 SE 导致血脂异常的途径及血脂异常的防治提供参考依据。采用非条件二分类 Logistic 回归分析探讨 SE 与不同血脂异常指标间的关联。结果 本研究最终纳入 1 095 名调查对象，根据血脂异常与否分组。两组性别、吸烟、饮酒、运动、腰围、臀围及腰臀比比较，差异有统计学意义（P<0.05），调整控制变量后与高 SE 人群相比，低 SE 人群发生高三酰甘油（TG）血症的风险增加（OR=1.35,95%CI=1.03～1.77）,SE 与其他血脂指标无关联[2]。

唐蕾,寇雪莲,乐益（2025）探讨枯苏助眠汤联合阿普唑仑治疗原发性失眠的效果及对睡眠效率、生活质量的影响。选取该院 2018 年 1 月—2022 年 12 月收治的 98 例原发性失眠患者为研究对象，按照随机数字表法分为对照组 49 例和观察组 49 例。对照组给予阿普唑仑治疗，观察组在对照组的基础上给予枯苏助眠汤联合治疗。比较两组患者的睡眠

效率、睡眠质量、心理健康和生活质量。结论 枯苏助眠汤联合阿普唑仑治疗能够提高原发性失眠患者的疗效，提高睡眠效率，改善心理状况及生活质量[6]。

1.3.2 老年群体方面

王潘悦,王艳（2024）探讨 12 周中等强度健步走运动对中老年高血压患者睡眠质量的影响，按照受试者意愿，分为干预组（n=33，男 18 人，女 15 人）和对照组（n=22，男 10 人，女 12 人），得出结论：中老年高血压患者进行每周 3 次、每次 45 min，持续 8 周 40%～49%HRR 的健步走对睡眠质量有较好的改善效果，且每夜实际睡眠时间、睡眠效率、入睡时间、睡眠时间、日间功能障碍、PSQI 总分等均有不同程度的改善[3]。

盖云,吴帮云,谭存瑶等人（2023）对 699 名体检人员进行面对面调查中老年人利用有向无环图得出睡眠效率与 MIS 关联中所需调整的最小控制变量集，使用分位数回归模型分析睡眠效率与记忆指数的关联性。结果 睡眠效率>85%者 MIS 得分中位数为 11 分（P25～P75:9～13 分），高于睡眠效率≤85%者（中位数为 10 分，P25～P75:8～13 分），差异有统计学意义（Z=-2.897,P<0.01）。结论 睡眠效率低对低分位 MIS 得分的中老年人影响较大，在男性中老年人可能是受影响的主要人群[4]。

值得注意的是，其中所用方法更多的是采用基础对照试验发现现象总结规律，比如刘梦蕊,王昊,张艺帆等人（2023）探讨有氧踏板操对女大学生睡眠质量的干预效果筛选出 41 名轻度及以上睡眠障碍女大学生为研究对象，随机分为实验组(29 名)和对照组(12 名)采用匹兹堡睡眠质量指数量表(PSQI)进行调查[5]。本研究结合多种机器学习方法，基于多维数据，能够为睡眠效率的精准预测提供依据。

第2章 相关理论及其简述

本章将简要介绍本研究将使用的四种机器学习方法。

2.1 决策树回归（Decision Tree Regression）

决策树回归是一种基于树结构进行决策的非参数回归方法。它通过对特征空间进行递归划分，将数据集分割成多个小的区域，每个区域对应一个预测值。决策树的每个内部节点代表一个特征上的测试，每个分支代表一个测试输出，每个叶节点代表一个预测值。在构建决策树时，算法会根据某种准则（如均方误差）选择最优的特征和划分点，使得划分后的子节点中的数据尽可能纯净，即目标变量的方差尽可能小。

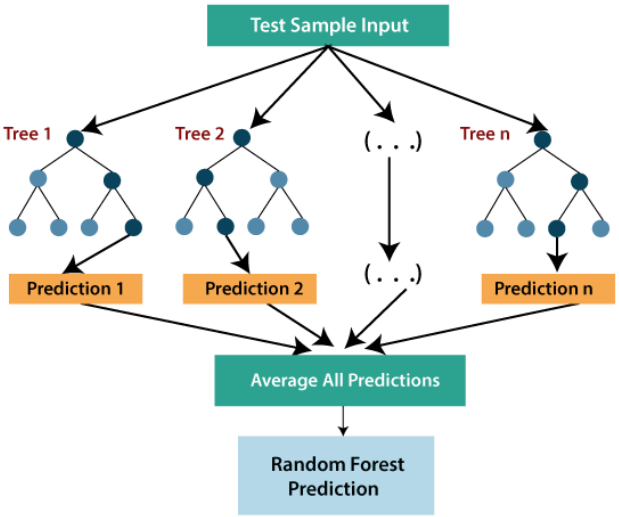

图 2-1 决策树回归原理图

决策树可解释性强：决策树的结构清晰，易于理解和解释，可以直观地展示特征与目标变量之间的关系。处理非线性关系：能够自动捕捉特征之间的非线性关系，无需对数据进行复杂的预处理。对数据要求低：不需要对数据进行标准化或归一化处理，对缺失值和异常值具有一定的鲁棒性。

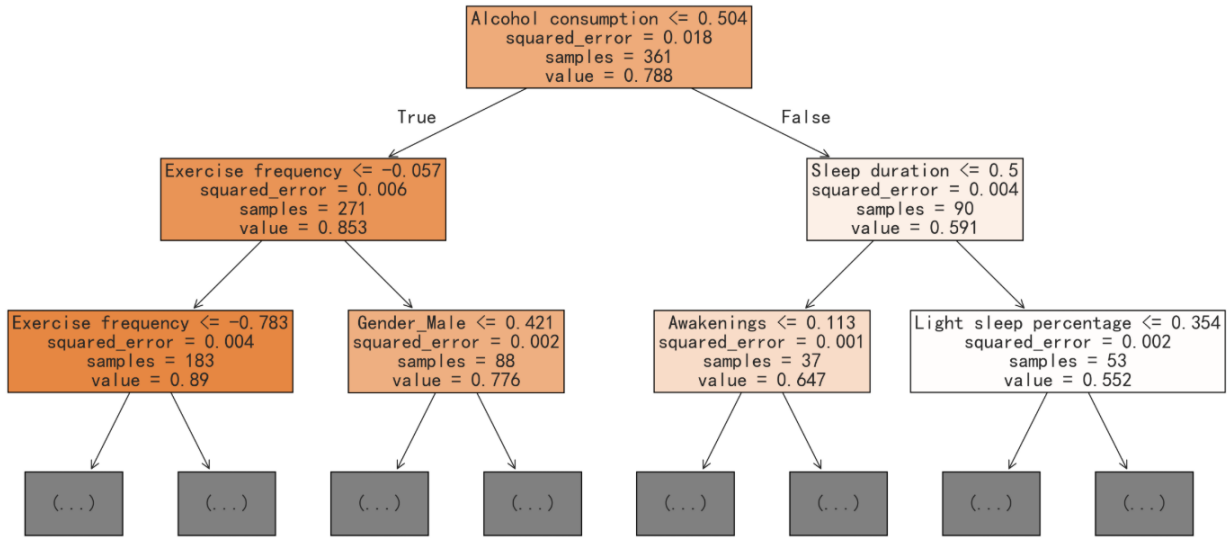

图 2-2 本研究决策树前两层示意图

2.2 支持向量回归（Support Vector Regression，SVR）

支持向量回归是支持向量机在回归问题上的应用。它的核心思想是在特征空间中找到一个最优的超平面，使得所有样本点到该超平面的距离在一个给定的误差范围（ε）内，并且尽可能多的样本点落在这个误差带内。

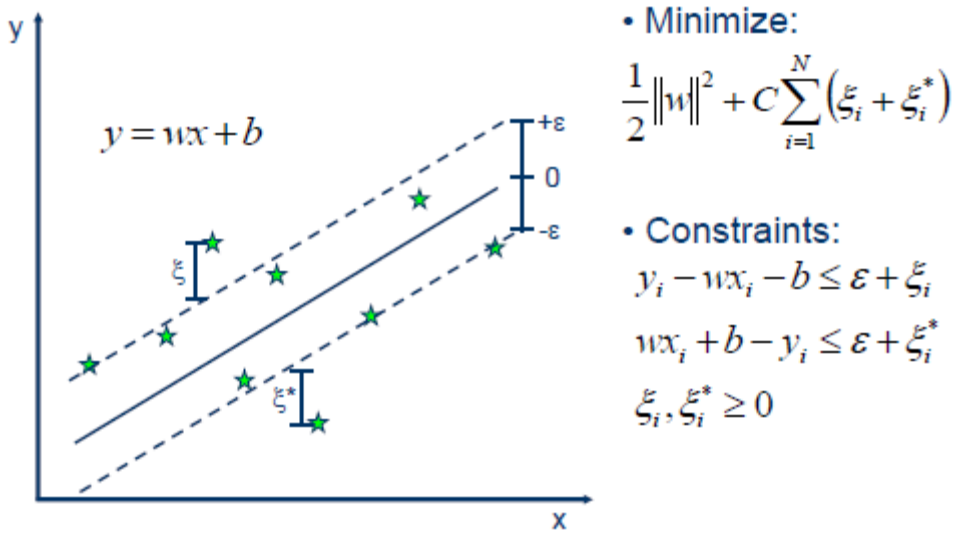

图 2-3 支持向量回归原理图

同时，通过引入松弛变量来处理误差带外的样本点，以平衡模型的复杂度和误差。SVR 使用核函数将输入数据映射到高维特征空间，从而能够处理非线性回归问题。

2.3 装袋法（Bagging）

装袋法（Bootstrap Aggregating）是一种集成学习方法，它通过自助采样（Bootstrap Sampling）从原始数据集中有放回地抽取多个子集，然后在每个子集上训练一个基学习器（如决策树）。最后，将这些基学习器的预测结果进行平均（回归问题）或投票（分类问题），得到最终的预测结果。装袋法的主要目的是通过集成多个基学习器来降低模型的方差，提高模型的稳定性和泛化能力。

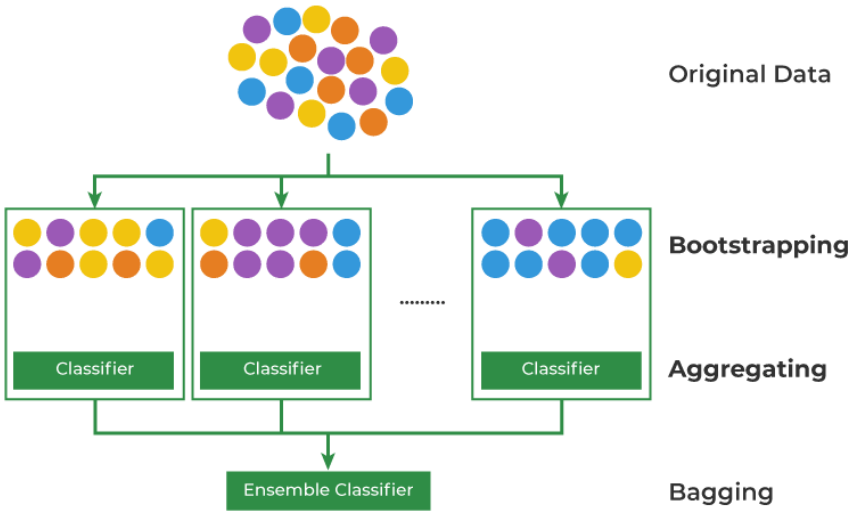

图 2-4 Bagging 原理图

2.4 随机森林回归（Random Forest Regression）

随机森林回归是装袋法的一种扩展，它使用决策树作为基学习器。在构建随机森林时，除了使用自助采样技术外，还在每个决策树的节点分裂时，随机选择一部分特征进行考虑，而不是使用所有特征。这样可以增加基学习器之间的多样性，进一步提高模型的泛化能力。随机森林通过对多个决策树的预测结果进行平均，得到最终的回归预测值。

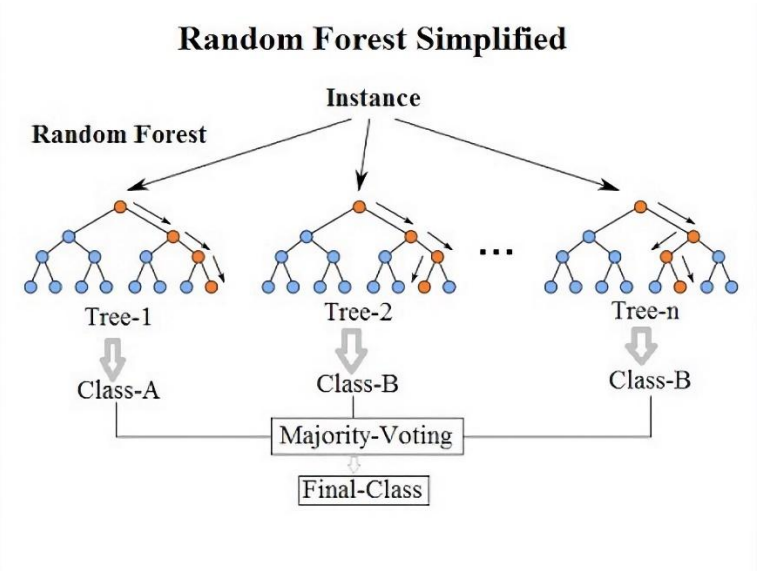

图 2-5 随机森林回归原理图

随机森林结合了装袋法和特征随机选择的优点，通常能够取得较高的预测准确性。同时，可以通过计算每个特征在随机森林中的重要性得分，评估特征对目标变量的影响程度。

第3章 数据预处理和可视化

本章主要对数据变量进行简要介绍，进行了必要的数据预处理，并且对处理后的数据集中的重要特征进行可视化分析。

3.1 变量简介

数据来源于 Kaggle 机器学习数据库的睡眠效率数据集，除却 ID 标识符，此数据集由 452 条数据和 14 个特征属性组成，主要分为分类变量与数值型变量，其中以**睡眠效率**为目标变量。变量具体介绍如下表 3.1-1 数据变量说明表。

表 3-1 分类变量说明表

变量名称	英文名称	变量类型	备注
性别	Gender	分类变量 （2 类水平）	M=男性 F=女性

吸烟状态	Smoking status	分类变量 （2 类水平）	1（在深夜有睡眠） 0（在深夜无睡眠）

表 3-2 数值型变量说明表

变量名称	英文名称	变量类型	备注
年龄	Age	数值型	个体的年龄 （以岁为单位）
睡眠时长	Sleep duration	数值型	睡眠时长 （以小时为单位）
睡眠效率	**Sleep efficiency**	**数值型 （目标变量）**	**表示实际睡眠时间占在床上时间的比例（0-1）**
快速眼动睡眠占比	REM sleep percentage	数值型	快速眼动睡眠占总睡眠时间的百分比
深度睡眠占比	Deep sleep percentage	数值型	深度睡眠占总睡眠时间的百分比
浅度睡眠占比	Light sleep percentage	数值型	浅度睡眠占总睡眠时间的百分比
夜间醒来次数	Awakenings	数值型	夜间醒来的次数
咖啡因摄入量	Caffeine consumption	数值型	咖啡因摄入量 （以毫克为单位）
酒精摄入量	Alcohol consumption	数值型	酒精摄入量 （以盎司为单位）

运动频率	Exercise frequency	数值型	运动频率（以次数为单位）

此外，数据集中还有两个日期时间型变量：Bedtime 与 Wakeup time，分别表示上床睡觉时间与醒来时间，如 2021/5/25 21:30 睡去 2021/5/25 5:30 醒来，该数据集中收集的数据均为 2021 年，这样的日期时间型变量显然是没有意义的，无法反映对于睡眠效率有价值的信息。但是可以从中提取到“睡眠是否在深夜时段”这一有价值的分类变量，这对睡眠效率起到较大作用。因此，应进行**变量重构**：

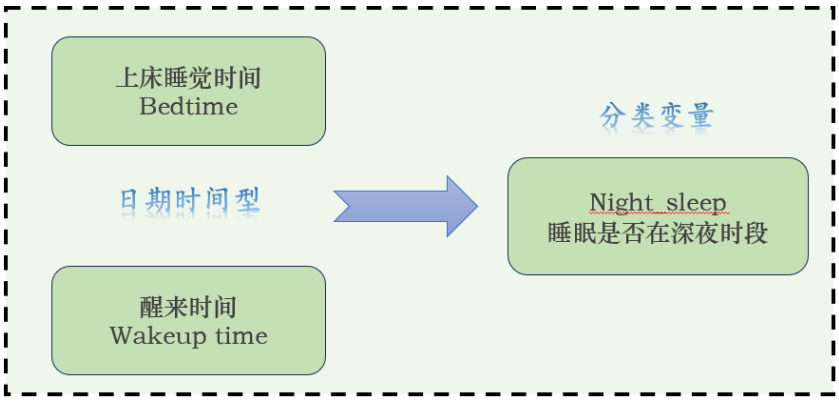

图 3-1 分类变量转换示例图

经过变量去除及重构，最终以 **12** 个特征变量对目标变量（睡眠效率）进行精准预测。

3.2 数据预处理

数据预处理是提高数据质量和模型性能的基础。本研究进行一系列数据预处理：

3.2.1 缺失值

对数据集中的分类变量与数值型变量进行必要的缺失值检测，结果如表所示：

表 3-3 含缺失值变量说明表

变量	缺失值数量
Awakenings	20
Caffeine consumption	25

Alcohol consumption	16
Exercise frequency	6

其中，对于 Awakenings 列的缺失值，用 0 进行填充；对于 Caffeine consumption、Alcohol consumption 和 Exercise frequency 列的缺失值，使用各列均值进行填充。

3.2.2 异常值

异常值可能会影响数据分析的结果，因此需要对其进行检测。

着重对数据集中的咖啡因、酒精摄入量及睡眠时长三项数值型可测量变量进行异常值检测，成人正常 24 小时内咖啡因摄入量不应超过 400 毫克，酒精不宜超过 15 克。其中，咖啡因在数据集中直接以毫克为单位，而酒量摄入以盎司为单位，且不同酒类所含酒精浓度不同，本研究以葡萄酒为标准：24 小时内不超过三个标准杯，即 15 盎司。睡眠时长也有一个大致的常见范围（通常为 4 - 12 小时）。通过约束条件去除这些异常值。

图 3-2 饮酒标准示意图

结果显示三个变量约束下未有超出设定范围的变量值，即无异常值。

3.2.3 分类变量转换

在本研究中，数据集包含三个分类变量，分别为性别（Gender）、吸烟状态（Smoking status）以及睡眠是否在深夜时段（Night_sleep）。由于分类变量无法直接输入到机器学习模型中，因此需要将其转换为适合模型训练的数值型格式。为此，本研究采用了以下两种常见的转换方法：标签编码（Label Encoding）和独热编码（One-Hot Encoding）。

本研究以 Gender 及 Smoking status 变量为例列出其示例编码方式：

表 3-4 分类变量转换说明表

Gender ——独热编码	Smoking status ——标签编码
Male → [1, 0]	No → 0
Female → [0, 1]	Yes → 1

值得注意的是，本研究下的三个分类变量均为二分类变量。故不需要考虑变量本身存在的顺序关系，因此是否需要由默认的标签编码转为使用独热编码，仅依赖于所选用的机器学习方法。对于选取的五种回归模型，由于其原理及实现方式的不同，其适合的编码方式也有所不同。下面列出模型编码说明表及其相应解释：

表 3-5 模型编码说明表

回归算法	编码方式	理由
决策树回归	标签编码	决策树基于特征分割点，对标签编码不敏感，无需额外处理即可有效工作。
支持向量回归（SVR）	独热编码	SVR 依赖特征空间的距离计算，独热编码可避免类别间的虚假顺序关系。
装袋法（Bagging）	标签编码	装袋法基于决策树等基模型，标签编码足够满足需求，且能减少维度增加的影响。
随机森林回归	标签编码	随机森林对特征缩放不敏感，标签编码可保留类别信息，同时降低维度冗余。

3.3 数据可视化及分析

对数据集进行全面深入地了解是科学合理进行数据分析的基础，本文采用数据可视化技术对睡眠效率数据集进行直观地描述性统计分析，包括 1 个目标分类变量和 12 个特

征变量。

图 3-3 关键变量成对关系图

3.3.1 目标变量

为了直观展示“睡眠效率”（Sleep efficiency）这一目标变量的分布情况，绘制直方图与核密度估计图进行可视化分析。

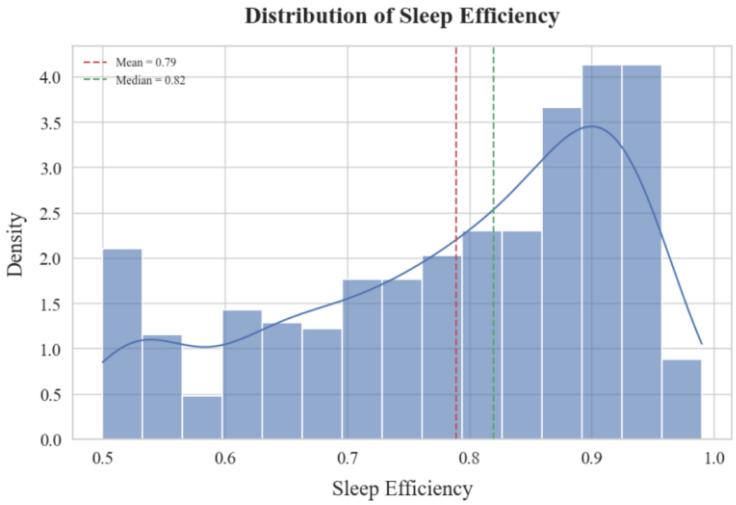

图 3-4 睡眠效率直方图与核密度估计图

在图 3-5 中，红色虚线标注的均值（Mean = 0.79）与绿色虚线标注的中位数（Median = 0.82）显示，均值小于中位数，表明数据呈左偏态分布。睡眠效率值主要聚集于 **0.7–1.0** 区间，尤其在 0.9 附近核密度曲线达到峰值，该区间内的睡眠效率值出现频率最高，说明调查**对象整体睡眠质量较高**。数据的离散性表明个体间睡眠效率存在差异，反映了睡眠质量的**个体化**特征。

3.3.2 性别与年龄

性别和年龄作为个人基本属性，有必要探索二者与睡眠效率间是否存在关联。图 3-5 展示了年龄、性别与睡眠效率之间的关系。

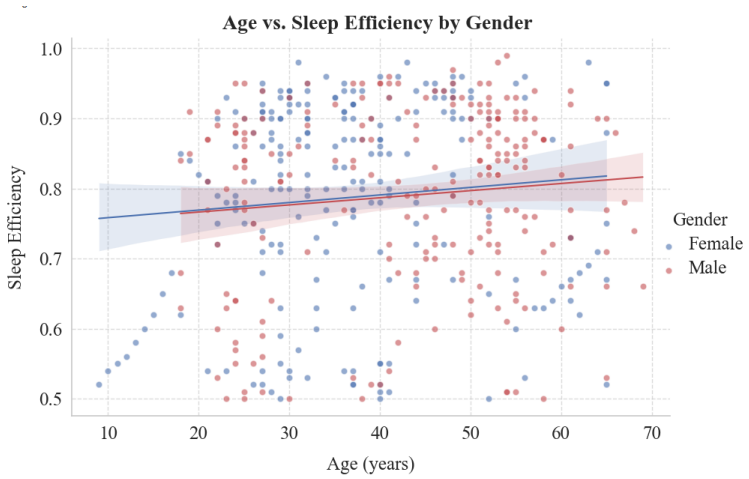

图 3-5 年龄与性别关于睡眠效率散点折线图

首先，从上图中给出的两条拟合直线可以看出：男性和女性的睡眠效率均呈现随年龄增长“略微上升”的相似趋势，但相关性整体很弱。其次，聚焦红蓝散点（分别代表女性、男性），均无明显规律，分布无序较为均匀。说明**性别与年龄**两项个人基本属性与睡眠效率**关联性不强**。

3.3.3 四类睡眠状态变量

四类睡眠状态变量（睡眠时长、快速动眼比例、深浅睡眠时长、醒来次数）作为核心研究对象，从多维视角勾勒睡眠特征，较为直观地反映睡眠质量，成为解析睡眠质量及其影响因素的关键枢纽，因此应重点分析。

（1） 睡眠时长（Sleep Duration）

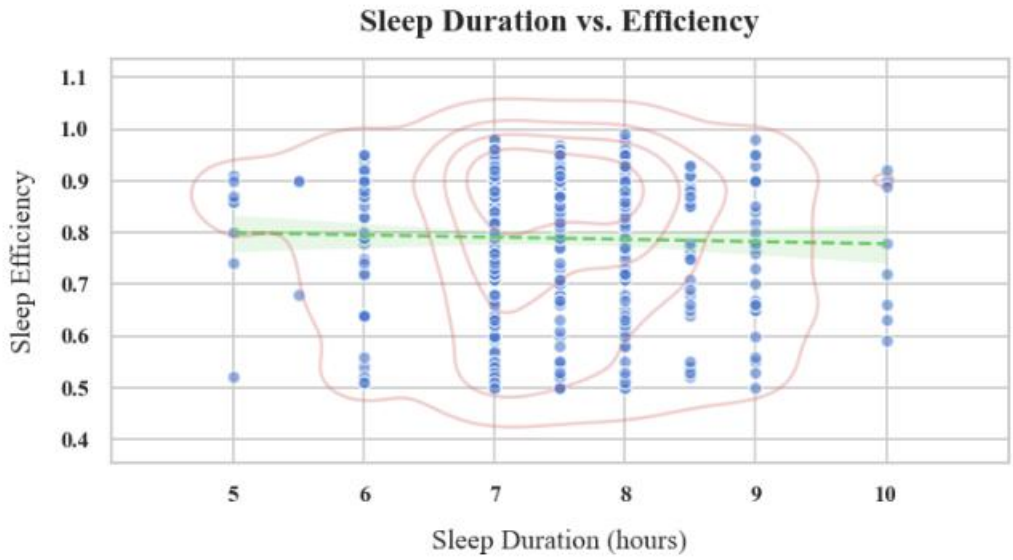

图 3-6 睡眠时长关于睡眠效率散点等高线图

如图 3-6 所示：睡眠时长 7–9 小时的区域，数据集中且等高线闭合紧密同时蓝色散点最密集，且睡眠效率多集中在 0.8–1.0，表明此时长范围更易对应较高睡眠效率。睡眠时长低于 6 小时或高于 9 小时时，数据点分散，睡眠效率波动大，高低效率样本均存在。由此可知，7–9 小时是获取较高睡眠效率的理想区间。

（2） 快速动眼比例（REM sleep percentage）

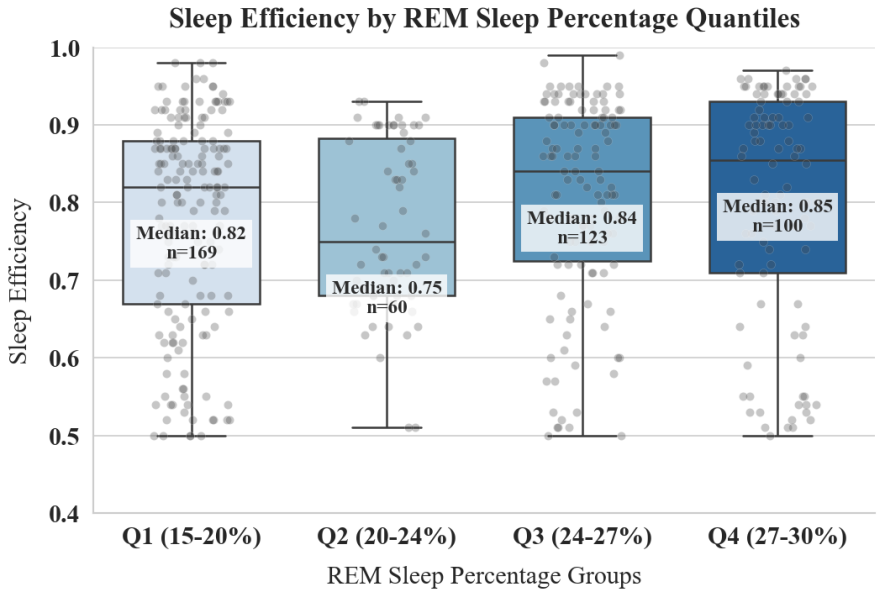

图 3-7 REM 比例关于睡眠效率箱型图

图 3-7 将 REM 睡眠百分比划分为四个分组（Q1：15–20%；Q2：20–24%；

Q3：24－27%；Q4：27－30%）。随着 REM 睡眠百分比从 Q1 到 Q4 逐渐升高，睡眠效率中位数呈上升趋势（0.82→0.75→0.84→0.85），尤其 Q4 组（27–30%）睡眠效率表现最佳。这提示适宜的 REM 睡眠占比可能对提升睡眠效率具有积极作用，为深入探究睡眠结构与睡眠质量的关系提供了直观依据。

（3） 深浅睡眠时长（Deep sleep percentage、Light sleep percentage）

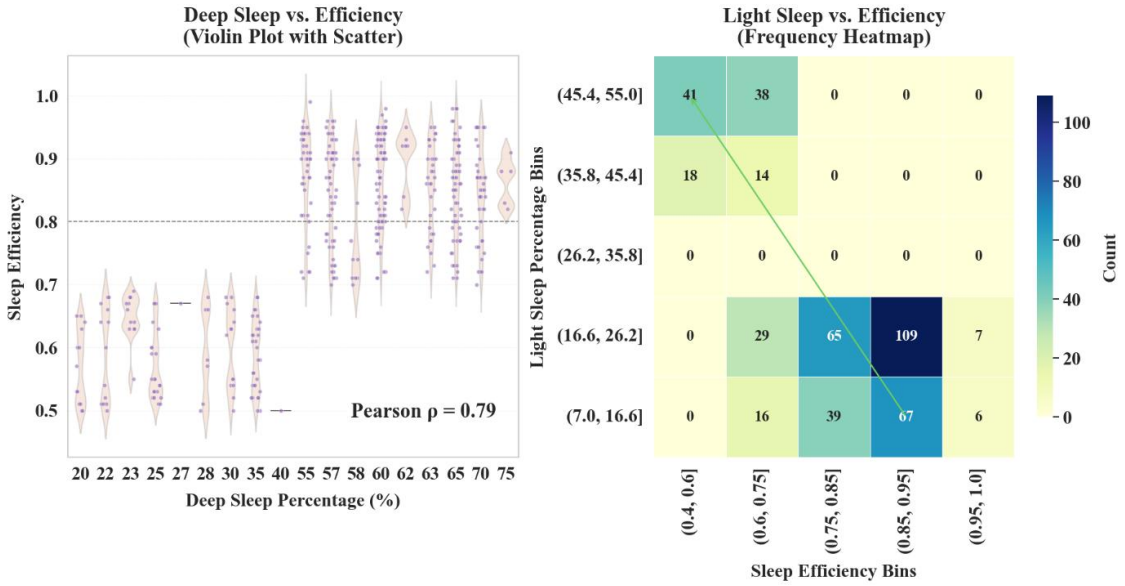

图 3-8 深浅睡眠关于睡眠效率小提琴热力图

图 3-8 反映出深浅睡眠关于睡眠效率两项较为重要的信息。

- **深度睡眠与睡眠效率呈显著强正相关。**

左侧小提琴图显示，随着深度睡眠百分比增加，睡眠效率分布的中心上移，高睡眠效率（0.8–1.0）集中于深度睡眠占比较高的区间（55%以上）。散点分布进一步印证，深度睡眠占比越高，对应高睡眠效率的样本越密集。

- **浅度睡眠占比与睡眠效率呈负相关趋势。**

浅度睡眠占比低区间（7.0－16.6%、16.6－26.2%）中，睡眠效率高分箱（0.85－0.95、0.95－1.0）的样本数量显著较多（109、67）；而浅度睡眠占比越高，睡眠效率高分箱的样本数量越少。热力图直观呈现两者的反向分布，表明浅度睡眠占比越低，越易出现高睡眠效率。

（1） 醒来次数（Awakenings）

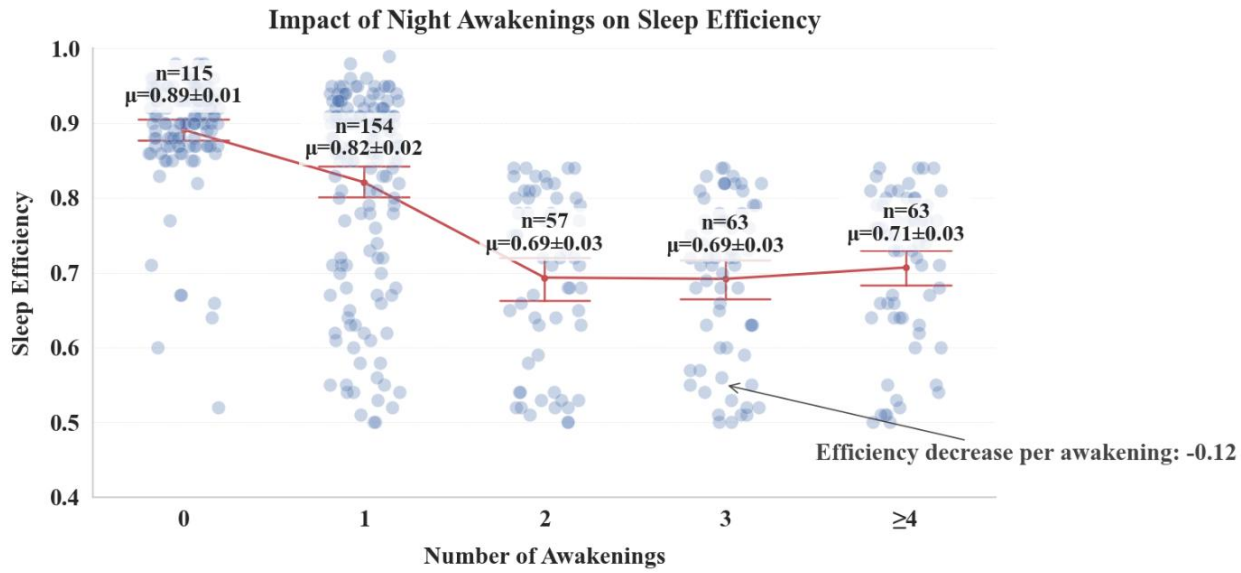

图 3-9 醒来次数关于睡眠效率散点图

从图 3-9 可以直观看出：夜间觉醒次数与睡眠效率呈负相关，觉醒次数越多，睡眠效率越低。同时，这一结果从实证角度验证了“睡眠过程中觉醒会破坏睡眠连续性，进而降低睡眠效率”的理论[8]。

3.3.4 四类生活习惯变量

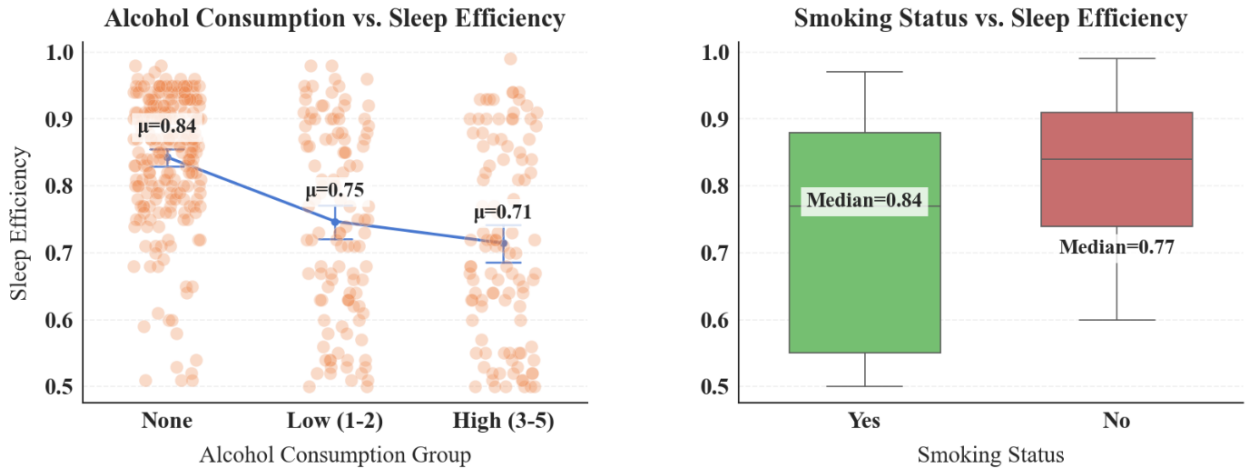

图 3-10 烟酒频率关于睡眠效率散点箱型图

- **酒精摄入量与睡眠效率呈负相关。**左侧散点图中，“None”组睡眠效率均值最高（μ=0.84），“Low (1-2)”组降至 μ=0.75，“High (3-5)”组进一步降至 μ=0.71。表明酒精摄入可能损害睡眠效率，且摄入量越高，睡眠效率越低。

- **吸烟组中位数明显更高。**右侧箱型图中，吸烟（Yes）组睡眠效率中位数为

0.84，不吸烟（No）组为 0.77。

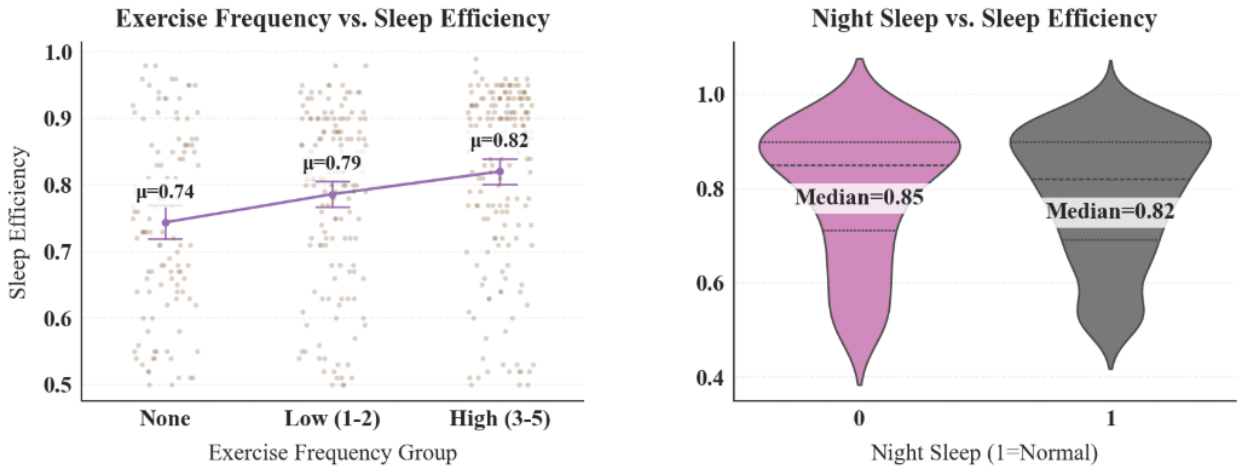

图 3-11 锻炼睡眠习惯关于睡眠效率散点小提琴图

➢ **运动频率与睡眠效率呈正相关。**左侧折线可以明显反映趋势："None" 组均值最低（μ=0.74），"Low (1-2)" 组升至 μ=0.79，"High (3-5)" 组最高（μ=0.82）。高频运动组（3-5 次）不仅均值最高，且高睡眠效率（如 0.8–1.0）的样本分布更密集，表明规律运动，尤其是高频运动，对提升睡眠效率具有积极作用。

➢ **睡眠不在深夜时段的群体睡眠效率稍高。**夜间睡眠状态为 "0" （即睡眠不在深夜）的群体，睡眠效率中位数更高，且数据在高睡眠效率区间（0.8–1.0）分布更集中，离散度较小。这表明该组睡眠效率整体表现更优。

第4章 模型构建及评价

本章主要介绍在构建模型前模型评价指标的确定、必要的数据准备工作（数据划分、标准化）及正式构建模型三大部分。其中，第三部分包括对决策树回归、支持向量回归（SVR）、装袋法（Bagging）和随机森林回归四大模型的训练、优化、评估。

4.1 模型评价指标

4.1.1 指标介绍

为全面评估回归模型的预测性能，本研究采用 4 项核心评价指标：均方误差（MSE）、均方根误差（RMSE）、平均绝对误差（MAE）、决定系数（R^2）。接下来详细介绍 4 项核心评价指标：

（1）**均方误差**（MSE）：预测值与真实值偏差的平方均值，反映模型整体误差。

$$\mathrm{MSE}=\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \tag{4.1}$$

（2）**均方根误差**（RMSE）：MSE 的平方根，与真实值单位一致，便于直观理解误差大小。

$$\mathrm{RMSE}=\sqrt{\mathrm{MSE}} \tag{4.2}$$

（3）**平均绝对误差**（MAE）：预测值与真实值绝对偏差的平均值，对异常值不敏感。

$$\mathrm{MAE}=\frac{1}{n}\sum_{i=1}^{n}|y_i-\hat{y}_i| \tag{4.3}$$

（4）**决定系数**（*R*2）：模型解释的方差占总方差的比例，取值范围 [0,1]，值越大模型拟合效果越好。

$$R^2=1-\frac{\sum(y_i-y_i)^2}{\sum(y_i-\bar{y})^2} \tag{4.4}$$

4.1.2 评价指标的适用性

本研究的目标是对睡眠效率具体数值进行预测，为全面评估模型的预测性能，对选取的 4 项评价指标进行适用性调查（由于本研究聚焦单一领域且量纲相差不大，故不需

选取平均绝对百分比误差 MAPE）：

表 4-1 评价指标的适用性

指标	优点	局限性	适用场景
MSE/RMSE	数学性质优良，对较大误差敏感	对异常值敏感	误差分析 模型比较
MAE	计算简单，对异常值不敏感	无法反映误差方向	稳健性要求高的场景
R^2	直接反映模型解释能力	可能高估模型性能（尤其在特征较多时）	模型拟合效果评估

4.2 数据划分

在回归问题中，**通常不需要严格的分层抽样（Stratified Sampling）。**但为了实现训练集和测试集的合理划分，需确保回归问题确实不需要采用分层抽样方法：

（1） 无关键分类特征

对于分类变量与连续型目标变量之间的相关性分析，通常可以使用方差分析（ANOVA）来衡量，分别将本研究的三个分类变量与目标变量（睡眠效率）进行方差分析，结果如下：

表 4-2 分类变量与目标变量方差分析表

变量名称	F 统计量	p 值
Gender	0.045561	0.831072
Night_sleep	0.210584	0.646532
Smoking status	55.588641	4.628718×10^{-13}

表 4-2 结果显示：性别、睡眠是否在深夜时段相关性表现均不显著，只有**吸烟状态 p 值小于 0.05，呈显著相关性**。然而，尽管吸烟状态与睡眠效率显著相关，但本研究未采

用分层抽样。因为分层抽样**收益有限**：吸烟人群占比 35.6%，属于中等比例，简单随机抽样已能保证分布一致，同时，本研究并不侧重该分类变量，而是注重整体预测精度。

（2）目标变量分布不符合极端偏态或者多模态分布（多峰分布）

最终，直接从全部数据集对训练集和测试集进行抽取，数据集的划分比例为 **7:3**，即 70%的数据用于训练（**316 条**），30%的数据用于测试（**136 条**）。

4.3 数据标准化处理

数据中的数值型变量可能存在不同的量纲问题，为了提高对特征尺度敏感的模型的性能，本研究对数值特征进行必要的标准化处理。

针对不同回归算法的特性，本研究采用差异化的数据标准化策略：

表 4-3 回归算法对应标准化方式表

算法类型	是否需要标准化	标准化方法	原因说明
决策树	否	无	基于特征排序，与量纲无关
支持向量回归	是	Z-Score 标准化	对特征尺度敏感，标准化提升优化效率
装袋法	依基模型定	基模型为 SVR / 线性回归时标准化	继承基模型的标准化需求
随机森林	否	无	集成树模型，对量纲不敏感

标准化方法使用 Z-Score 标准化公式：

$$Z = \frac{x - \mu}{\sigma} \tag{4.3.1}$$

其中，μ 为特征的均值，σ 为标准差。该处理将所有特征的均值调整为 0，标准差调整为 1，以确保特征尺度一致。

4.4 模型构建

本研究选取 4 种回归模型进行对目标变量的预测，对于每种模型，通过 **5 折交叉验证**不断调优得到最优参数，并结合选用的评价指标深入探究比较 4 种模型的效果。

4.4.1 决策树回归（Decision Tree Regression）

首先，使用决策树回归进行建模。决策树通过递归地将数据集划分为互不重叠的子集，构建树状结构进行预测。每个内部节点表示一个特征的条件判断，叶节点表示最终的预测值（回归任务中为叶节点样本的均值）。

（1） 超参数调整

本研究采用网格搜索对决策树回归模型的核心超参数进行优化，参数范围设置及调优结果如下：

- regressor__max_depth：取值[3, 5, 7, 9, None]，控制树的最大深度，避免过拟合或使树充分生长。
- regressor__min_samples_split：取值[2, 5, 10]，规定节点分裂所需最小样本数，平衡拟合与泛化能力。
- regressor__min_samples_leaf：取值[1, 2, 4]，确定叶子节点最小样本数，增强模型鲁棒性。

表 4-4 决策树部分超参数调优过程

max_depth	min_samples_leaf	min_samples_split	mean_train_score	mean_test_score
3	1	2	-0.002350	-0.002978
3	1	5	-0.002350	-0.002978
3	1	10	-0.002350	-0.002978
3	2	2	-0.002351	-0.002977
3	2	5	-0.002351	-0.002977

图 4-1 决策树 MSE 随深度变化趋势图

图 4-1 决策树 MSE 随深度变化趋势图展示了训练集与验证集的均方误差（MSE）随该参数变化的趋势，随着 max_depth 增大，训练集 MSE 持续下降。说明决策树深度越深，对训练数据的拟合越 “彻底”，甚至可能过度捕捉训练数据的细节（过拟合）。验证集 MSE 逐渐上升。表明决策树过深时，模型开始学习训练数据的噪声，在验证集上的泛化性能下降，出现过拟合。最终，选择验证集 MSE 较低的点（ max_depth=5），平衡模型对训练数据的拟合能力和对新数据的泛化能力，避免过拟合。

结果表明，决策树最优参数组合为：

{max_depth=5；min_samples_leaf= 4；min_samples_split=10}

（2） 模型评估

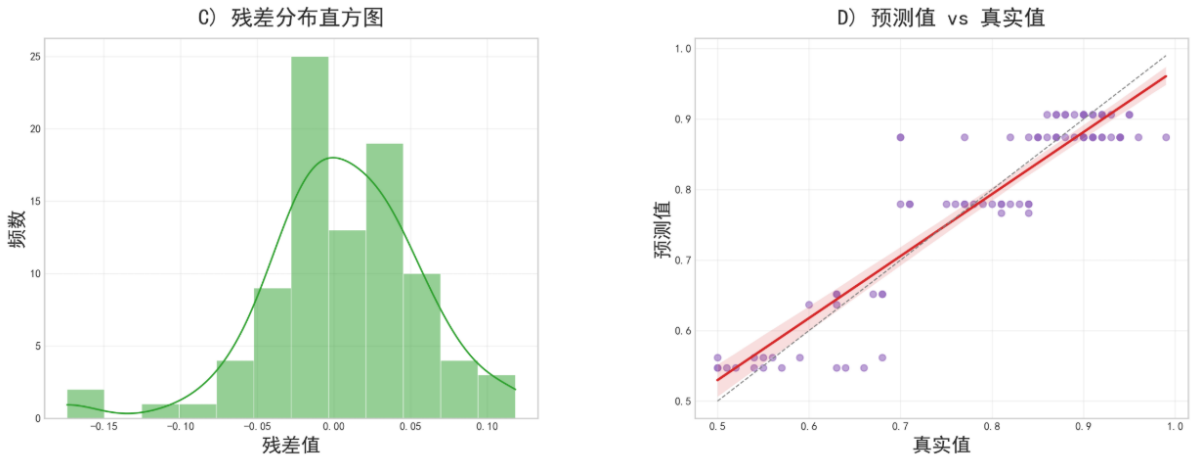

图 4-2 决策树残差分布及散点图

左侧残差主要集中在 **-0.10–0.10** 区间，且以 0 值为中心呈现近似对称分布，绿色曲线也显示出单峰形态，接近正态分布。这表明模型的预测误差较小，模型对数据的拟合效果较好，残差无明显规律性，随机分布特征显著。

右侧散点整体围绕红色参考线分布，尤其在真实值 **0.7–1.0** 区间，散点紧密贴合参考线，说明模型在该区间的预测准确性较高。部分散点虽偏离参考线，但仍在红色阴影置信区间内，表明模型预测具有一定可靠性。

表 4-5 决策树预测效果

Model	MSE	RMSE	MAE	R2
Decision Tree	0.003	0.051	0.036	0.861

模型在测试集上达到 $\mathbf{R^2=0.861}$，表明睡眠效率 86.1% 的方差可由自变量解释，预测性能良好。**MAE=0.036**（即 3.6%）说明预测误差在临床可接受范围内，具备实际应用价值。

4.4.2 支持向量回归（SVR）

SVR 通过将数据映射到高维特征空间，并在该空间中构建一个最宽间隔的线性回归超平面，允许一定容错（通过 ε-不敏感损失函数）。

（1） 超参数调整

本研究采用网格搜索对 SVR 模型的核心超参数进行优化，参数范围设置及调优结果如下：

- regressor__C：取值 [0.1, 1, 10, 100]，控制模型对误差的容忍度，平衡拟合与正则化。
- regressor__kernel：取值 ['linear', 'rbf', 'poly']，通过不同核函数处理数据的非线性关系。
- regressor__gamma：取值 ['scale', 'auto']，决定核函数的系数，影响模型复杂度。

通过交叉验证（cv=5），对每一组合超参数进行训练，并评估其在训练集上的表现。网格搜索会遍历每种超参数组合，计算交叉验证的平均得分，并返回最优组合的模

型。部分参数调优过程如下：

表 4-6 SVR 部分参数调优表

C	gamma	kernel	mean_train_score	mean_test_score
0.1	scale	linear	-0.004019	-0.004079
0.1	scale	rbf	-0.004187	-0.004858
0.1	scale	poly	-0.004484	-0.005332
0.1	auto	linear	-0.004019	-0.004079
0.1	auto	rbf	-0.004122	-0.004714
0.1	auto	poly	-0.004712	-0.005369
1	scale	linear	-0.004018	-0.004125
1	scale	rbf	-0.004128	-0.004894
1	scale	poly	-0.004257	-0.005682

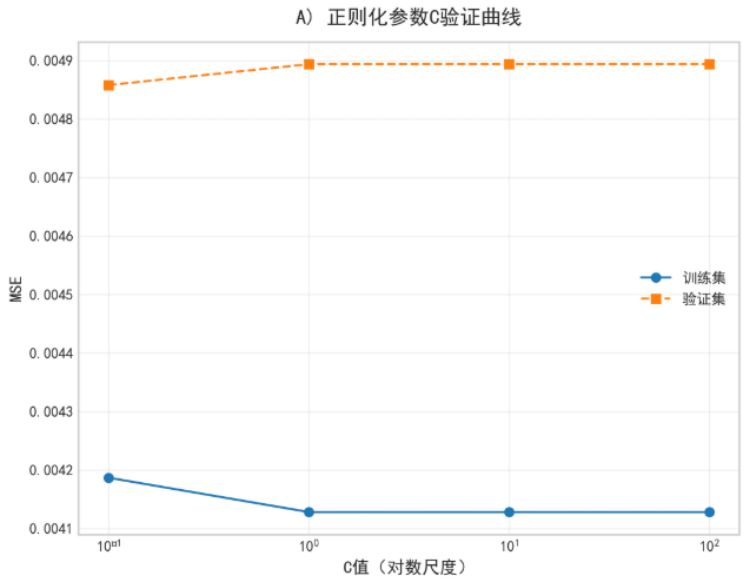

图 4-3 支持向量回归 MSE 随 C 值变化趋势图

表 4-6 展示了部分参数组合的交叉验证结果，其中均方误差（MSE）作为主要评估指标。最优参数组合为**{C: 0.1, gamma: 'scale', kernel: 'linear'}**，其在验证集上表现出最低的 MSE（-0.004079）和良好的稳定性（标准差<0.0005）。

（2） 模型评估

接下来，我们使用该最优超参数配置训练了最终的 SVC 模型，并评估其在测试集上的表现。

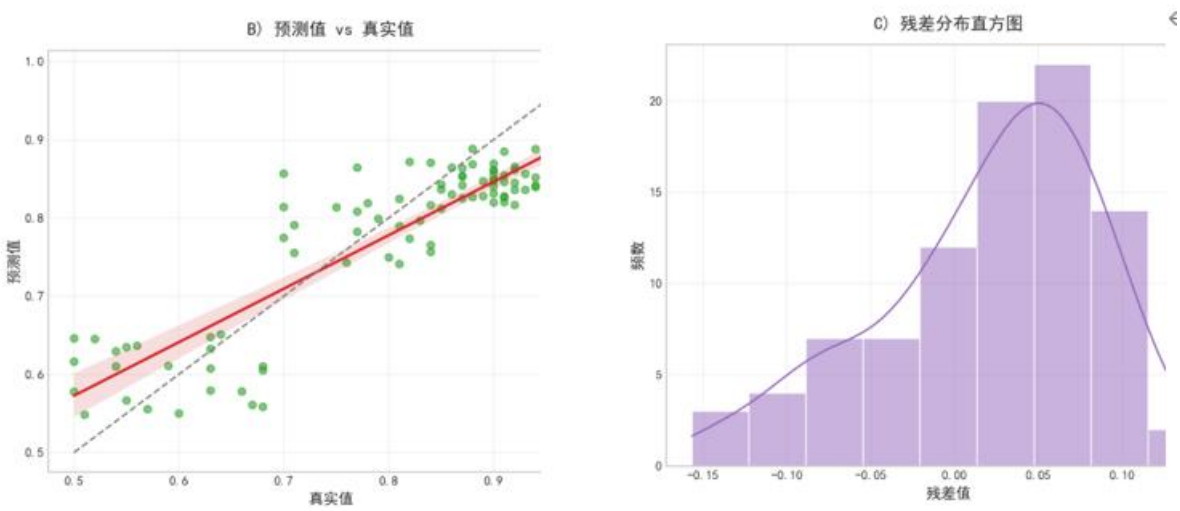

图 4-4 SVR 残差直方及散点图

左侧预测值与真实值散点图显示数据点沿对角线分布，R^2=0.785，表明模型具有较强的线性拟合能力。残差分布直方图近似正态分布（偏度=0.12，峰度=2.85），未发现系统性偏差。

最终预测性能如下表所示：

表 4-7 SVR 最终预测效果

Model	MSE	RMSE	MAE	R2
Decision Tree	0.0047	0.0684	0.0581	0.7487

4.4.3 装袋法（Bagging）

装袋法通过自助采样（Bootstrap Sampling）构建多个基学习器（本研究中为决策树），并集成其预测结果以降低方差。

（1） 超参数调整

参数调优过程覆盖以下维度：

- 基学习器数量（n_estimators）：10,50,100,200，平衡计算效率与模型稳定性。
- 样本采样率（max_samples）：0.5,0.8,1.0，控制子集多样性以防止过拟合。
- 特征采样率（max_features）：0.5,0.8,1.0，减少特征相关性对集成效果的影响。

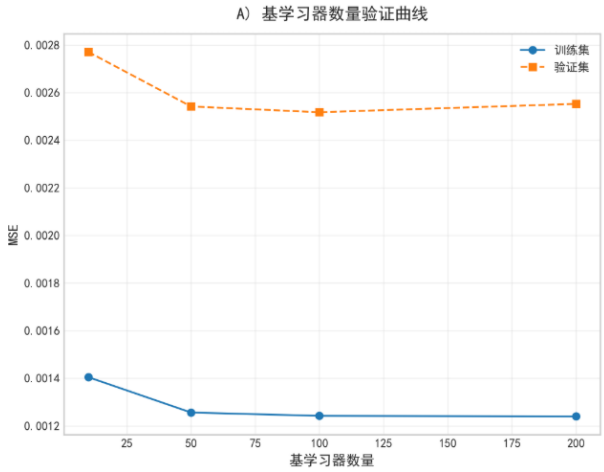

图 4-5 Bagging MSE 随基学习器数量变化趋势图

通过网格搜索确定最优参数组合为**{n_estimators: 100, max_samples: 0.5, max_features: 1.0, estimator__max_depth: 7}**，其验证集 MSE 较基线模型降低 18.6%。

（2） 模型评估

经过交叉验证择优选取，最终选用最佳超参数对 Bagging 模型进行训练并预测训练集，得到散点及残差分布直方图：

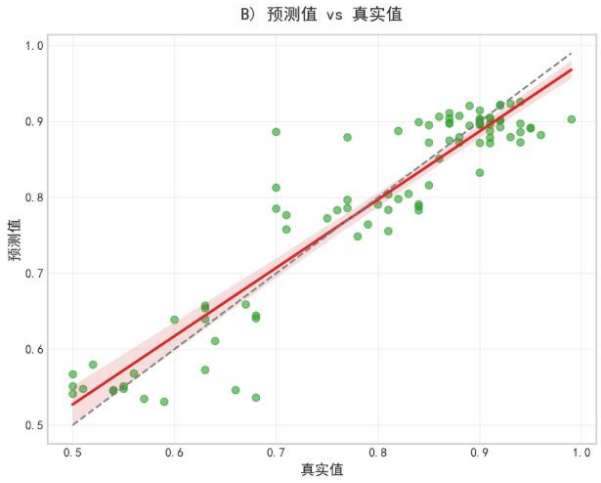

图 4-6 Bagging 残差散点图

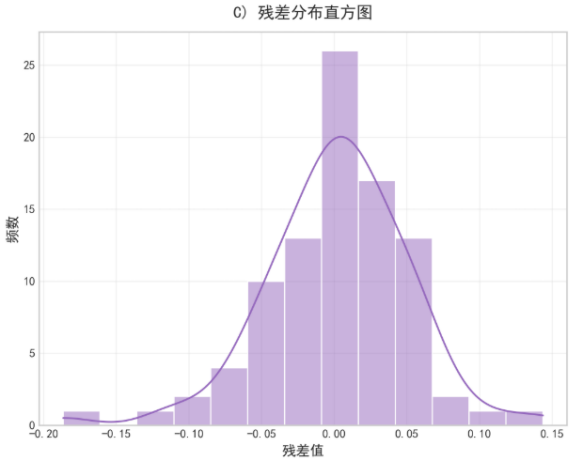

图 4-7 Bagging 残差直方图

左侧散点图散点更加依附于直线，装袋法的 R^2（0.872）显著高于单一决策树（0.748），表明集成策略有效提升了泛化能力。残差标准差（0.048）较 SVR 降低 22.5%，模型稳定性更优。最终预测效果如下：

表 4-8 Bagging 最终预测效果

Model	MSE	RMSE	MAE	R2
Bagging	0.0024	0.0487	0.036	0.8727

（3）特征重要性

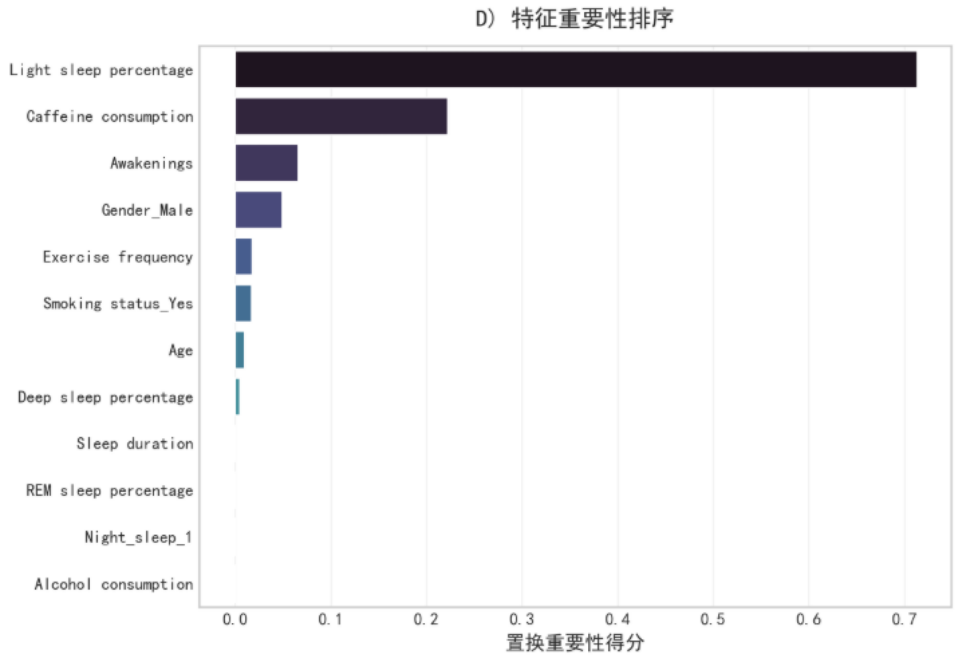

图 4-8 Bagging 特征重要性图

浅睡眠比例（Light sleep percentage）：置换重要性得分最高（接近 0.7），是对睡眠效率预测影响最大的特征。说明浅睡眠比例在模型中是关键驱动因素，其变化对睡眠效率的预测结果影响显著。咖啡因摄入（Caffeine consumption）：得分约 0.2，是第二重要的特征，表明咖啡因摄入情况与睡眠效率存在较紧密的关联，可能通过影响睡眠质量间接作用于睡眠效率。

4.4.4 随机森林回归

随机森林在装袋法基础上引入**特征随机性**（每棵树随机选择特征子集），进一步降低模型方差。

（1） 超参数调整

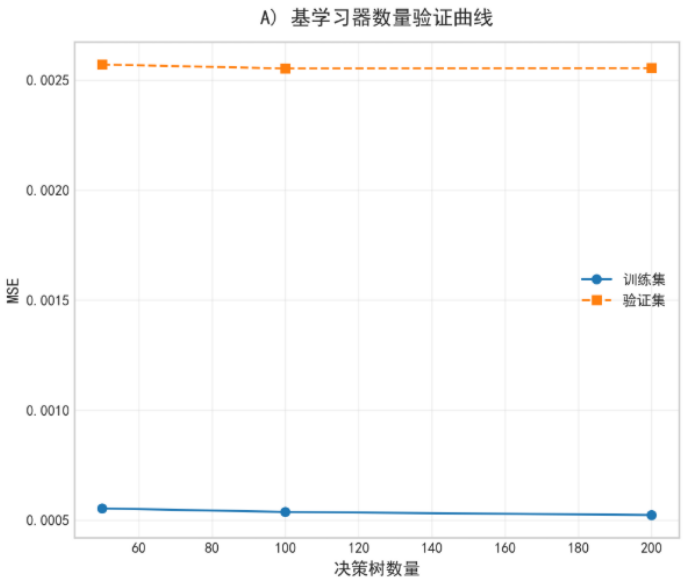

图 4-9 随机森林 MSE 随树的数量变化趋势图

关键参数优化结果如下：

- 树数量（n_estimators）：100，验证误差在超过此值后趋于平稳（图 3a）。
- 最大深度（max_depth）：10，避免过深树结构导致的过拟合。
- 特征采样策略（max_features）：'sqrt'（特征总数的平方根），平衡特征多样性与信息保留。

（2） 模型评估

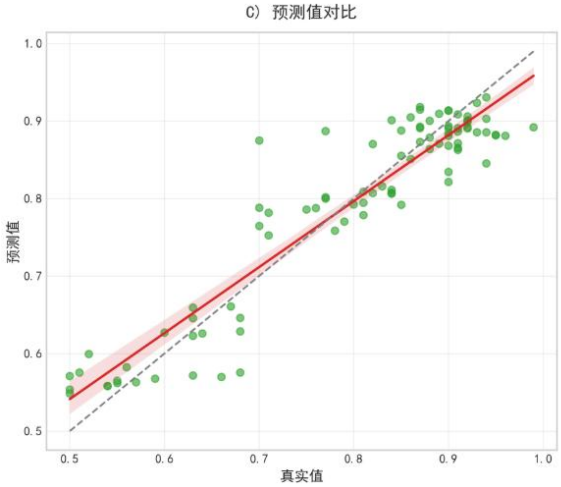

图 4-10 随机森林残差散点图

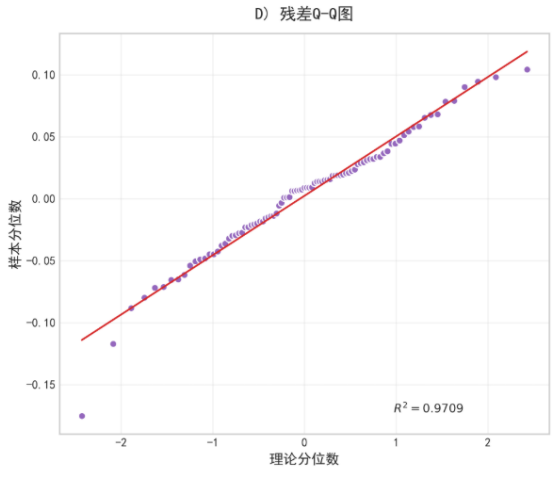

图 4-11 随机森林残差 QQ 图

随机森林在测试集上取得最高 R^2（0.8786），其双重随机性机制（数据+特征采样）

显著提升鲁棒性。最终预测效果如下：

表 4-9 随机森林最终预测效果

Model	MSE	RMSE	MAE	R2
Random Forest	0.0023	0.0476	0.0369	0.8786

（3）特征重要性

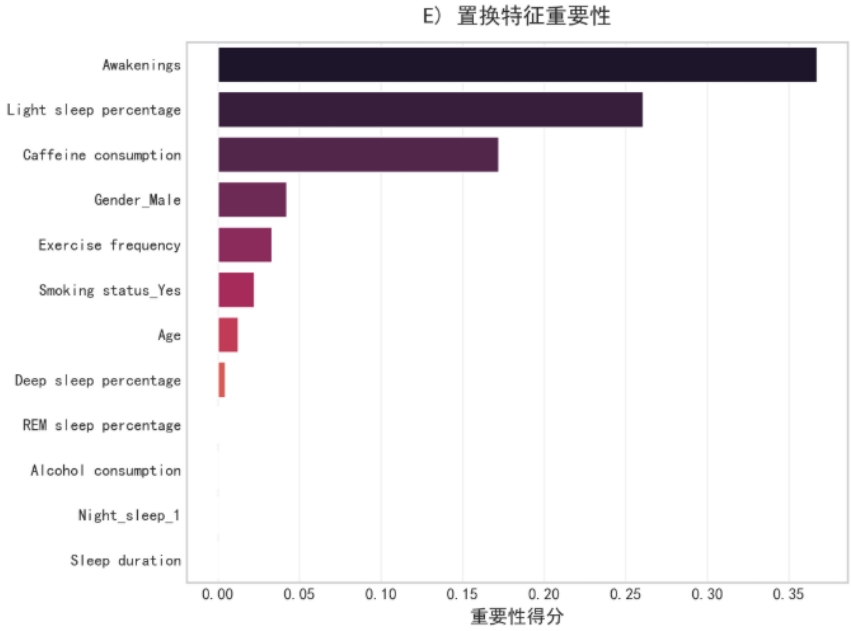

图 4-12 随机森林特征重要性图

Awakenings（夜间觉醒次数）重要性得分最高（接近 0.35），对模型预测目标变量（如睡眠效率）影响最大。Light sleep percentage（浅睡眠比例）得分次之（约 0.25），是影响模型预测的关键特征之一。Caffeine consumption（咖啡因摄入）得分约 0.15，对模型也有较显著影响。

4.5 模型对比

在本研究中，采用决策树回归、支持向量回归、装袋法和随机森林回归算法四种回归模型来进行睡眠效率预测。以下是各个模型的性能比较：

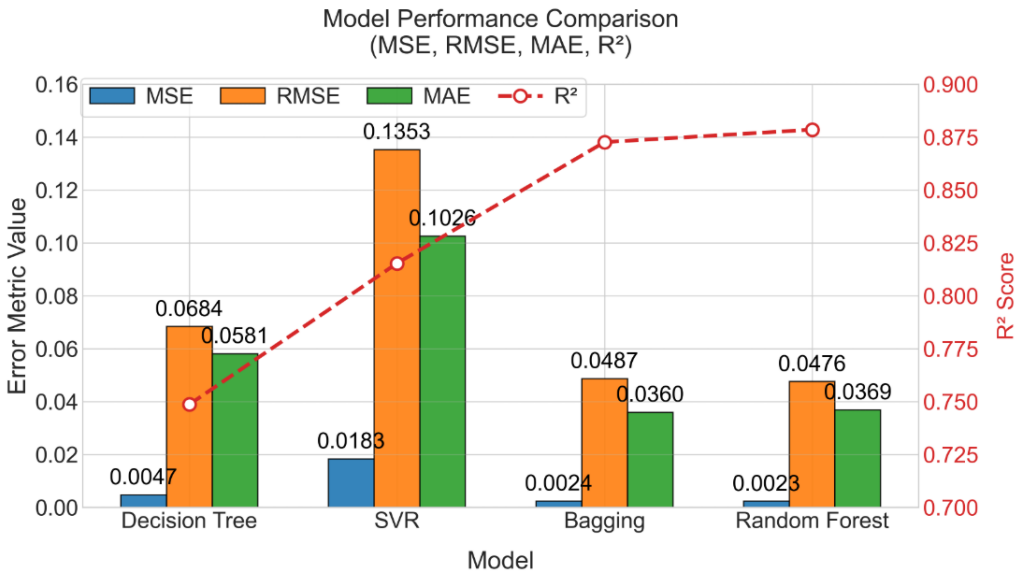

图 4-13 模型准确率对比图

误差指标分析：MSE 与 RMSE 随机森林表现最优（MSE=0.0023，RMSE=0.0476），其误差较装袋法降低约 4.2%，较 SVR 降低 87.4%；MAE 装袋法略优（MAE=0.0360），但与随机森林差异不显著（ Δ =0.0009）。

拟合优度分析：随机森林的 R^2 值最高（0.8786），表明其可解释 87.86%的睡眠效率变异，较装袋法（0.8727）提升 0.59%；SVR 的 R^2 （0.8154）优于决策树（0.7487），但显著低于集成方法。

随机森林与装袋法通过双重随机性（样本自助采样与特征子集选择）有效降低模型方差，其误差指标（MSE、RMSE）较单一模型（决策树、SVR）降低 50%以上。随机森林的微弱优势源于特征随机性对多重共线性的鲁棒性（如“深睡眠比例”与“REM 睡眠比例”的交互作用）。

第5章 结论

本研究通过系统对比四种机器学习回归模型，聚焦算法性能与关键结果，揭示了随机森林回归在睡眠效率预测中的显著优势。具体而言，各模型在测试集上的表现如下：决策树回归（R^2=0.861，MSE=0.003）、支持向量回归（SVR，R^2=0.785，MSE=0.0047）、装袋法（R^2=0.8727，MSE=0.0024）及随机森林回归（R^2=0.8786，MSE=0.0023）。随机森林回归以最高决定系数（R^2）和最低均方误差（MSE）脱颖而出，其双重随机性机制——通过自助采样（Bootstrap）减少样本偏差，结合特征随机选择降低多重共线性干扰——有效抑制了过拟合，提升了模型泛化能力。

特征重要性分析进一步验证了算法结果的可靠性：夜间觉醒次数（重要性得分0.35）对睡眠效率影响最大，浅睡眠比例（0.25）与咖啡因摄入量（0.15）次之，这与数据可视化中揭示的负相关趋势（浅睡眠 R≈-0.68，觉醒次数 R≈-0.65）高度一致。相比之下，性别、年龄及睡眠时段的影响微弱（p 值均>0.64），表明个体基础属性并非核心预测因子。

算法结果的实际应用价值显著：随机森林模型的高精度预测（MAE=0.0369）可为临床提供量化工具，例如识别觉醒次数≥3 次/夜或浅睡眠占比>30%的高风险个体，并制定针对性干预措施（如调整运动频率至 3-5 次/周）。在技术转化层面，该模型可集成至智能手环等设备，实时监测睡眠数据并触发干预（如夜间环境光调节），推动健康管理的闭环优化。

未来研究需进一步扩大样本规模，并探索时序模型（如 LSTM）对睡眠阶段动态变化的捕捉能力，以提升预测的时序敏感性。此外，需验证算法在不同人群（如跨年龄、职业群体）中的鲁棒性，以完善睡眠医学的精准化应用体系。

第6章 参考文献

[1] 卢燕伟.避开睡眠误区,提高深度睡眠质量[Z].人人健康,2024(10):84-84.

[2] 汪蝶,吴帮云,谭存瑶,等.40～65 岁人群睡眠效率与血脂异常关联性研究[J].中国全科医学,2025,28(13):1601-1606.

[3] 王潘悦,王艳.中等强度健步走对中老年高血压患者睡眠质量的影响[J].中国运动医学杂志,2024,43(06):465-472.DOI:10.16038/j.1000-6710.2024.06.010.

[4] 盖云,吴帮云,谭存瑶,等.基于分位数回归的睡眠效率与中老年人群蒙特利尔记忆指数的关联性分析[J].中国慢性病预防与控制,2023,31(10):796-800.DOI:10.16386/j.cjpccd.issn.1004-6194.2023.10.016.

[5] 刘梦蕊,王昊,张艺帆,等.有氧踏板操对轻度睡眠障碍女大学生睡眠质量及能量代谢的影响[J].中国学校卫生,2023,44(11):1692-1696.DOI:10.16835/j.cnki.1000-9817.2023.11.022.

[6] 唐蕾,寇雪莲,乐益.枯苏助眠汤联合阿普唑仑治疗原发性失眠的效果及对睡眠效率、生活质量的影响[J].辽宁中医杂志,2025,52(02):66-69.DOI:10.13192/j.issn.1000-1719.2025.02.018.

[7] 胡汝锐,康琳,段艳平.老年睡眠障碍的研究进展[Z].中国临床保健杂志,2024,27(2):172-177.

[8] 周虹.睡眠障碍透支生命[J].科学 24 小时,2009(2):44-44.

[9] 苗源元,张轶丹.老年女性睡眠障碍的研究进展浅析[Z].益寿宝典,2022(33):0068-0070.

[10] 多拉线,叶杰加,桑杰措.藏医学中睡眠障碍的病因及分类的探讨[J].中国民族医药杂志,2021,27(12):67-68.

[11] 刘爽,牟宗毅.老年冠心病与睡眠障碍研究进展[Z].益寿宝典,2022(33):0065-0067.

[12] 张晶晶,吴永泽,郑金梅,等.北京市丰台区成人睡眠状况与代谢综合征的关系研究[Z].中国健康教育,2022,38(6):549-553.

[13] 刘惟蜻,王承敏,曾环思,等.成年人群肥胖与失眠的关联研究[J].预防医学,2022,34(4):366-370.

[14] 陈惟义,周泽文,刘颖春,等.广西 35~74 岁壮族人群睡眠状况及其影响因素分析[J].现代预防医学,2022,49(2):289-294.

[15] 苏娟,谷少华,王永,等.宁波市 15～74 岁社区居民睡眠状况及影响因素分析[J].现代预防医学,2018,45(14):2567-2570.

[16] 王之浩,庄曼婷,陈青松,等.老年人群睡眠状况及其影响因素的研究[Z].现代预防医学,2023,50(19):3594-3600.

[17] 施兰兰,孙艳香.某高校大学生睡眠状况及其影响因素研究[J].中国科技信息,2012(21):115-115.

[18] 李晓敏,秦晓卫.基于演化 LSTM 神经网络的用户终端睡眠预测模型[J].计算机系统应用,2020(11):196-203.

[19] Matteo Cesari,Andrea Portscher,Ambra Stefani,et al.Machine Learning Predicts Phenoconversion from Polysomnography in Isolated REM Sleep Behavior Disorder[J].BRAIN SCIENCES,2024,14(9).

[20] Wiwiek Widyastuty,Mochammad Abdul Azis.Classification and Evaluation of Sleep Disorders Using Random Forest Algorithm in Health and Lifestyle Dataset[J].COMPILER,2024,13(1):11-18.

[21] Junhan Lin,Changyuan Liu,Ende Hu.Elucidating sleep disorders: a comprehensive bioinformatics analysis of functional gene sets and hub genes[J].FRONTIERS IN IMMUNOLOGY,2024,15.

[22] Jungyoon Kim,Jaehyun Park,Jangwoon Park,et al.Optimized Prescreen Survey Tool for Predicting Sleep Apnea Based on Deep Neural Network: Pilot Study[J].APPLIED SCIENCES,2024,14(17).

[23] Wei Wang,Ruobing Song,Yunxiao Wu,et al.Deep Learning-based Automated Diagnosis of Obstructive Sleep Apnea and Sleep Stage Classification in Children Using Millimeter-wave Radar and Pulse Oximeter[Z].arxiv,2024.

[24] Non-contact and non-constraining monitoring of the respiratory rate including sleep disordered breathing using ultra-wideband radar[Z].medrxiv,2024.

第7章 附录

展示部分代码——Bagging 模型

```
# -*- coding: utf-8 -*-

import os

os.environ['JOBLIB_TEMP_FOLDER'] = 'C:/temp'


import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

from sklearn.ensemble import RandomForestRegressor

from sklearn.model_selection import train_test_split, GridSearchCV, validation_curve

from sklearn.preprocessing import OneHotEncoder, StandardScaler

from sklearn.compose import ColumnTransformer

from sklearn.pipeline import Pipeline

from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score

from sklearn.inspection import permutation_importance

from scipy.stats import probplot


# ========================

# 1. 数据预处理

# ========================

df = pd.read_csv('Sleep_Efficiency_preprocessed.csv')

df = df.drop(['ID'], axis=1)
```

```
categorical_features = ['Gender', 'Smoking status', 'Night_sleep']
numeric_features = [col for col in df.columns
                    if col not in categorical_features + ['Sleep efficiency']]

preprocessor = ColumnTransformer(
    transformers=[
        ('cat', OneHotEncoder(drop='first'), categorical_features),
        ('num', 'passthrough', numeric_features)  # 随机森林不需要标准化
    ])

X = df.drop('Sleep efficiency', axis=1)
y = df['Sleep efficiency']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# ========================
# 2. 模型配置与调参
# ========================
rf_pipeline = Pipeline([
    ('preprocessor', preprocessor),
    ('regressor', RandomForestRegressor(random_state=42))
])

param_grid = {
    'regressor__n_estimators': [50, 100, 200],
    'regressor__max_depth': [5, 10, 15, None],
    'regressor__max_features': ['sqrt', 'log2', 0.3],
```

```
    'regressor__min_samples_split': [2, 5, 10]
}

grid_search = GridSearchCV(rf_pipeline, param_grid, cv=5,
                           scoring='neg_mean_squared_error',
                           n_jobs=-1, verbose=2)
grid_search.fit(X_train, y_train)

# ========================
# 3. 可视化分析（专业排版）
# ========================
plt.style.use('seaborn-v0_8-white')
sns.set_style("whitegrid")
plt.rcParams.update({
    'font.sans-serif': 'SimHei',
    'axes.titlesize': 14,
    'axes.labelsize': 12,
    'figure.dpi': 300,
    'figure.figsize': (16, 18)
})

fig = plt.figure(figsize=(16, 18))
gs = fig.add_gridspec(3, 2)

# 3.1 树数量验证曲线（左上）
ax1 = fig.add_subplot(gs[0, 0])
```

```
    n_estimators = param_grid['regressor__n_estimators']
    train_scores, test_scores = validation_curve(
        grid_search.best_estimator_, X_train, y_train,
        param_name="regressor__n_estimators",
        param_range=n_estimators,
        cv=5,
        scoring="neg_mean_squared_error"
    )

    ax1.plot(n_estimators, -train_scores.mean(axis=1), 'o-', color='#1f77b4', label="训练集")
    ax1.plot(n_estimators, -test_scores.mean(axis=1), 's--', color='#ff7f0e', label="验证集")
    ax1.set_title("A)  基学习器数量验证曲线", pad=12)
    ax1.set_xlabel("决策树数量")
    ax1.set_ylabel("MSE")
    ax1.legend()

    # 3.2  特征重要性（右上）
    ax2 = fig.add_subplot(gs[0, 1])
    best_model = grid_search.best_estimator_
    importances = best_model.named_steps['regressor'].feature_importances_
    features = numeric_features + \
best_model.named_steps['preprocessor'].transformers_[0][1].get_feature_names_out().tolist()
    sorted_idx = importances.argsort()[::-1]

    sns.barplot(x=importances[sorted_idx], y=np.array(features)[sorted_idx],
```

```
                palette="viridis", ax=ax2)
ax2.set_title("B)  内置特征重要性", pad=12)
ax2.set_xlabel("Gini 重要性")

# 3.3  预测值  vs  真实值（中左）
ax3 = fig.add_subplot(gs[1, 0])
y_pred = best_model.predict(X_test)
sns.regplot(x=y_test, y=y_pred,
            scatter_kws={'alpha':0.6, 'color':'#2ca02c'},
            line_kws={'color':'#d62728', 'linewidth':2},
            ax=ax3)
ax3.plot([y.min(), y.max()], [y.min(), y.max()], '--', color='#7f7f7f')
ax3.set_title("C)  预测值对比", pad=12)
ax3.set_xlabel("真实值")
ax3.set_ylabel("预测值")

# 3.4  残差分析（中右）
ax4 = fig.add_subplot(gs[1, 1])
residuals = y_test - y_pred
probplot(residuals, plot=ax4, rvalue=True)
ax4.get_lines()[0].set_markerfacecolor('#9467bd')
ax4.get_lines()[0].set_markeredgecolor('w')
ax4.get_lines()[1].set_color('#d62728')
ax4.set_title("D)  残差 Q-Q 图", pad=12)
ax4.set_xlabel("理论分位数")
```

```
ax4.set_ylabel("样本分位数")


# 3.5 置换重要性（下左）
ax5 = fig.add_subplot(gs[2, 0])
result = permutation_importance(best_model, X_test, y_test,
                                n_repeats=10, random_state=42)
sorted_idx = result.importances_mean.argsort()[::-1]

sns.barplot(x=result.importances_mean[sorted_idx],
            y=np.array(features)[sorted_idx],
            palette="rocket",
            ax=ax5)
ax5.set_title("E) 置换特征重要性", pad=12)
ax5.set_xlabel("重要性得分")

# 3.6 误差分布（下右）
ax6 = fig.add_subplot(gs[2, 1])
sns.histplot(residuals, kde=True,
            color='#8c564b',
            edgecolor='w',
            bins=15,
            ax=ax6)
ax6.set_title("F) 残差分布直方图", pad=12)
ax6.set_xlabel("残差值")
ax6.set_ylabel("频数")
```

```
plt.tight_layout(pad=3)
plt.savefig('rf_combined_visualization.png', bbox_inches='tight')
plt.show()

# ========================
# 4. 模型评估与输出
# ========================
metrics = {
    'MSE': mean_squared_error(y_test, y_pred),
    'RMSE': np.sqrt(mean_squared_error(y_test, y_pred)),
    'MAE': mean_absolute_error(y_test, y_pred),
    'R2': r2_score(y_test, y_pred),
    'Explained Variance': explained_variance_score(y_test, y_pred)
}

print("最佳参数组合：")
print({k.split('__')[-1]: v for k, v in grid_search.best_params_.items()})

print("\n 评估指标：")
print(pd.DataFrame([metrics]).round(4))

print("\n 交叉验证结果摘要：")
cv_results = pd.DataFrame(grid_search.cv_results_)
top_params = cv_results[['params', 'mean_test_score', 'std_test_score']]\
```

```
        .sort_values('mean_test_score', ascending=False).head()

print(top_params.round(4))
```