# 《 机 器 学 习 D 》

# 课 程 论 文

题目：基于机器学习的睡眠障碍分类预测模型研究

学生姓名：曹梓上

2024 年 12 月 10 日

# 摘 要

随着生活节奏的加快和心理压力的增加，睡眠障碍的患病率呈逐年上升趋势。有效的早期诊断和治疗对于改善患者的睡眠质量至关重要。机器学习技术凭借其在数据处理和模式识别方面的优势，已成为预测和分类睡眠障碍的有效工具。基于此，本研究致力于构建并评估基于机器学习的睡眠障碍分类预测模型。

本研究选取 Kaggle 官网睡眠健康和生活方式数据集，由 **374** 条数据和 **12** 个属性变量组成。旨在使用五种常见的分类算法（支持向量机 SVC、多层感知机 MLP、决策树、朴素贝叶斯和 K 近邻 KNN）对睡眠障碍进行分类，对其性能进行比较分析。同时，探究年龄（Age）、体重类别（BMI Category）、压力水平（Stress Level）等 11 个特征变量与目标变量睡眠障碍类别（Sleep Disorder）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独热编码、标签编码、SMOTE 过采样以及标准化处理，其中，独热编码主要对类别变量中的性别（Gender）及职业（Occupation）进行转换以消除类别之间的顺序关系。以 **7:3** 比例同分布划分训练集和测试集后，对训练集进行 SMOTH 过采样，使其由 **261** 份增至 **471** 份以解决类别不平衡问题。之后，输入标准化后的训练集对五种模型进行训练，并通过超参数调优，优化各个模型的性能。最终，基于宏平均及加权平均两种评价规则，评估模型的对测试集的分类效果，采用一系列常用指标进行性能评价。

实验结果表明，除朴素贝叶斯模型外其余四种模型均表现良好，准确率均在 **90%左右**，其中**决策树模型**在所有五种模型中表现最佳，准确率达到了 **90%**。朴素贝叶斯模型由于其假设特征之间独立，未能充分捕捉数据中的相关性，准确率为 78%，表现最差。各特征中，**BMI 类别**（BMI Category）对睡眠障碍影响最大。

综合考虑模型的准确性和计算效率，决策树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，且在宏平均和加权平均下表现接近（加权精确率为 0.91、加权召回率为 0.90、加权 F1 分数为 0.90），有效地消除了数据不平衡带来的影响，实际应用效果良好。

**关键词**：睡眠障碍、机器学习、支持向量机、多层感知机、决策树、朴素贝叶斯、K 近邻

# Abstract

The prevalence of sleep disorders is on the rise year after year as the pace of life accelerates and psychological pressure increases. Effective early diagnosis and treatment are crucial to improve the sleep quality of patients. Machine learning technology, with its advantages in data processing and pattern recognition, has become an effective tool for predicting and classifying sleep disorders. Based on this, this study is devoted to constructing and evaluating a predictive model for classifying sleep disorders based on machine learning.

The sleep health and lifestyle dataset from the official Kaggle website was selected for this study, consisting of **374** data and **12** attribute variables. It aims to classify sleep disorders using five common classification algorithms (Support Vector Machine SVC, Multi-Layer Perceptron MLP, Decision Tree, Plain Bayes, and K Nearest Neighbor KNN), and to compare and analyze their performance. Meanwhile, the relationship between 11 feature variables such as Age, BMI Category, and Stress Level and the target variable Sleep Disorder category (Sleep Disorder) is explored.

First, after the necessary missing values and outliers were detected in the dataset, data preprocessing was carried out, including unique heat coding, label coding, SMOTE oversampling, and standardization, in which the unique heat coding mainly converted the gender and occupation in the category variables to eliminate the sequential relationship between the categories. After dividing the training set and test set with the same distribution in the ratio of **7:3**, the training set was subjected to SMOTH oversampling to increase it from **261** to **471** to solve the category imbalance problem. After that, the input standardized training set is used to train the five models, and the performance of each model is optimized by hyperparameter tuning. Finally, the models' classification effects on the test set are assessed based on two evaluation rules, macro average and weighted average, and a series of commonly used metrics are used for performance evaluation.

The experimental results show that all four models except the plain Bayesian model perform well with an accuracy of **around 90%,** with **the decision tree model** performing the best among all five models with an accuracy of **90%**. The plain Bayesian model, which fails to adequately capture the correlations in the data due to its assumption of independence between features, performed the worst with an accuracy of 78%. Among the features, **BMI Category** had the greatest impact on sleep disorders the most.

Considering the accuracy and computational efficiency of the model, the decision tree is the most suitable model for the task of this study, which not only provides a high prediction accuracy, but also performs close to the macro-averaging and weighted averaging (weighted precision of 0.91, weighted recall of 0.90, and weighted F1 scores of 0.90), which effectively removes the effects of data imbalance, and is effective for practical application.

**Keywords**: sleep disorders, machine learning, support vector machine(SVC), multilayer perceptron(MLP), decision tree, naive bayes, k-nearest neighbor(KNN).

# 目录

# 第1章 绪论

本章主要介绍睡眠障碍研究的现状和前景、主要研究内容和应用价值以及引用文献的综述。

## 1.1 研究现状

近年来的研究表明，睡眠障碍在中国已成为一个普遍的问题。根据多项调查显示，中国成年人中约有 30%到 50%存在不同程度的睡眠障碍，其中失眠是最常见的类型[1]。睡眠障碍的发生率随年龄增长而上升，中老年群体尤其需要关注[2]。此外，城市化进程、生活节奏加快以及工作压力增加被认为是睡眠障碍流行率增长的重要原因[3]。

睡眠障碍的病因复杂多样，包括心理社会因素（如压力、焦虑）[15]、生理因素（如神经化学物质的改变）、生活方式（如不规律的作息时间、饮食习惯）和环境因素（如噪音、光污染）等[5]。遗传因素在某些类型的睡眠障碍中也起到一定作用，如家族性失眠，这表明基因在睡眠障碍的形成中可能扮演重要角色[9]。

在诊断与评估方面，多导睡眠监测（PSG）虽为“金标准”[7]，但其有设备昂贵、操作复杂等局限。展望未来，多学科交叉融合助力睡眠障碍研究。机器学习有望实现早期自动诊断、精准分型及风险预测[20]，为个性化治疗提供依据。新治疗靶点和药物研发将持续推进，同时，睡眠健康教育和预防干预也将成为重点[9]，以降低发病率，提升公众睡眠质量和整体健康水平。

## 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障碍分类预测模型为核心任务。所采用的数据集来源于 Kaggle 官网睡眠健康和生活方式数据集，此数据集由 400 条数据和 13 个特征属性组成。如患者的基本生理指标（BMI 类别、心率、血压等）、生活习惯数据（睡眠质量、睡眠持续时间等）、心理状态评估（压力水平等）以及睡眠相关的特征属性。

通过对这些数据进行全面深入的分析，挖掘各特征与睡眠障碍类型及程度之间的内在联系。在此基础上，运用多种先进的机器学习算法（如决策树、支持向量机、神经网络等）构建分类预测模型，并对模型进行优化和评估，以实现对不同类型睡眠障碍的准确识别和预测，为临床诊断和个性化治疗提供科学依据。

### 1.2.2 应用价值

本研究成果在医学和公共卫生领域意义重大。于临床诊断方面，辅助医生快速精准判断睡眠障碍类型，减少误诊漏诊。在治疗上，能据此制定个性化方案，涵盖药物、心理、生活方式干预，提升疗效。同时，有助于识别高危人群与风险因素，推动疾病预防和公众良好睡眠习惯养成，降低发病率。还可为睡眠医学研究提供数据与理论依据，助力深入探究发病机制、病理生理过程，开发新治疗方法。

## 1.3 文献综述

在睡眠障碍逐渐被重视的大背景下，研究睡眠障碍的早期诊断和有效分类成为医学领域的重要课题。

传统技术方面，苏娟、谷少华（2018）等人分层多阶段抽样方法，对宁波市社区居民的睡眠状况进行调查，并应用 Logistic 回归模型分析居民睡眠质量的影响因素，多因素 Logistic 回归分析显示，女性（OR=1.88，95%CI：1.532.30，P〈0.001）、离婚或丧偶（OR=1.70，95%CI：1.092.64，P=0.019）、近期饮酒者（OR=1.32，95%CI：1.06~1.66，P=0.014）以及年龄在 35 岁以上的居民睡眠质量较差[12]。同样，袁帆、丁彩翠（2018）等人 694 名职业人群作为研究对象，分析了他们的睡眠状况及其影响因素。多因素 logistic 回归分析：我国职业人群存在睡眠不足与睡眠过多并存的现象，且睡眠时间分布受地区、性别、年龄、文化程度、婚姻状况、经济状况及职业的影响[13]。

模型应用方面，王之浩、庄曼婷（2023）等人为描述老年人群的睡眠现况及影响，调查广东省中山市两镇参与 2020 年度老年人体检的 9,020 名≥60 岁老年人，进行多因素线性回归和 logistic 回归分析，得出可靠结论[14]。Wiwiek Widyastuty、Mochammad Abdul Azis（2024）旨在利用机器学习（ML）的能力，特别是随机森林算法，来检测和分析收集的数据集中指示睡眠障碍的模式。表明随机森林算法在睡眠障碍检测中可以达到高水平的准确性。该模型展示了 97.33%的测试准确率，96%的精确度，100%的召回率。此外，它还达到了 98%的 F1 分数和 0.945 的 Kappa 分数，验证了该算法在产生精确分类方面的可靠性[18]。Jungyoon Kim 、Jaehyun Park（2024）等人阻塞性睡眠呼吸暂停（OSA）共有 66 名参与者，包括 45 名男性和 21 名女性，应用了多种建模技术，包括深度神经网络结合主成分分析（DNN-PCA）、随机森林（RF）、自适应增强分类器（ABC）、决策树分类器（DTC）、K 近邻分类器（KNC）和支持向量机分类器

（SVMC）结果显示，DNN-PCA、RF、ABC、DTC、KNC 和 SVMC 模型的 AUROC 值分别为 0.95、0.62、0.53、0.53、0.51 和 0.78，表明不同模型的适用性及可行性[20]。

综合创新方面，Wei Wang、Ruobing Song（2023）等人评估毫米波雷达装置和多导睡眠图(PSG)在诊断儿童睡眠阻塞性睡眠呼吸暂停(OSA)和睡眠分期方面的一致性。研究对象为 281 名年龄在 1 至 18 岁之间的儿童，这些儿童于 2023 年 9 月至 11 月期间在北京儿童医院睡眠中心接受了睡眠监测。所有参与的儿童同时接受了 PSG 和 QSA600 的睡眠监测。QSA600 的记录通过深度学习模型自动分析，而 PSG 数据则手动评分。使用深度学习模型对 QSA600 录音进行自动分析，同时对 PSG 数据进行人工评分。通过交叉验证评估的深度学习模型在诊断 OAHI>1、OAHI>5 和 OAHI>10 的儿童时表现出良好的敏感性（分别为 81.8%、84.3%和 89.7%）和特异性（分别为 90.5%、95.3%和 97.1%）。受试者工作特征曲线下面积分别为 0.923、0.955 和 0.988[21]。

# 第2章 相关理论及其简述

本章将简要介绍本文将使用的五种机器学习方法。

## 2.1 支持向量机（Support Vector Machine, SVM）

支持向量机是一种监督学习模型，常用于分类和回归问题。SVM 的目标是找到一个超平面，能够将不同类别的样本有效地分开且使得该超平面与样本点的距离最大化。最大化间隔能够提高模型的泛化能力，从而在未见过的测试数据上有更好的表现。

支持向量分类器（SVC）是 SVM 的一个实现，专门用于处理分类问题。SVC 的目标是通过最大化间隔来找到一个最优的超平面，从而有效地区分不同类别的样本。与 SVM 理论中所讨论的优化目标一致，SVC 通过求解最优化问题，找到一个最优的超平面。

在现实应用中，数据通常是线性不可分的，或者数据中存在一些噪声。为了使模型能够适应这些情况，SVM 引入了软间隔（Soft Margin），即允许某些样本点被错误分类，但对错误分类的样本给予惩罚。优化目标为：

$$\text{Minimize} \quad \frac{1}{2}\| w \|^2 + C\sum_{i=1}^{n}\xi_i \tag{2.1.1}$$

$$\text{Subject to} \quad y_i(\omega^T x_i + b) \geq 1 - \xi_i, \forall i \tag{2.1.2}$$

- $\xi_i$ 是松弛变量，用来允许样本点被错误分类。

●C 是惩罚参数，控制间隔的宽度与分类错误的权衡。较大的 C 值意味着对错误分类的惩罚较大，模型更倾向于找到一个更严格的决策边界。

软间隔使得 SVM 能够更好地处理具有噪声或不完全分离的数据。

## 2.2 多层感知器（Multilayer Perceptron, MLP）

多层感知器（MLP）是一种前馈神经网络，由输入层、多个隐藏层和输出层组成。每层的节点（神经元）通过加权连接与前一层的节点相连。MLP 通过非线性激活函数（如 ReLU、Sigmoid 等）对数据进行变换，能够有效地拟合复杂的非线性函数。训练过程通常使用反向传播算法，通过梯度下降方法优化网络的权重参数。

MLP 常用于分类和回归任务，尤其适用于处理非线性关系强的数据。

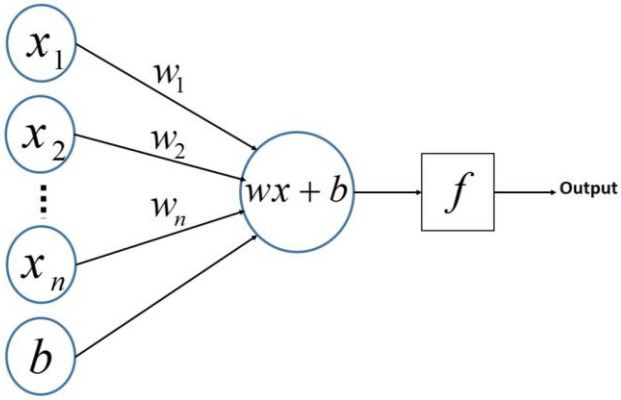



**图 2.2-1 MLP 神经网络简示图**

MLP 通常包括三个基本层：

- 输入层（Input Layer）：接收外部输入数据。输入层的节点数等于特征的维度。
- 隐藏层（Hidden Layer）：一或多个隐藏层是 MLP 模型的重要组成部分。隐藏层的节点数（即神经元个数）是超参数，通过调节它们可以增加网络的复杂度和学习能力。隐藏层中的每个节点通过激活函数（如 ReLU、sigmoid、tanh 等）对加权输入进行非线性变换。
- 输出层（Output Layer）：输出层的节点数根据任务的需求而定，分类任务中输出层的节点数等于类别数，回归任务中输出层通常只有一个节点。

## 2.3 K 近邻（K-Nearest Neighbors, KNN）

K 近邻算法是一种基于实例的学习方法，主要用于分类和回归问题。KNN 的基本思

想是，给定一个待分类的样本，找到训练集中与其最相似的 K 个邻居，根据这些邻居的类别（多数表决原则）来决定该样本的类别。KNN 是一个懒惰学习算法，即它在训练阶段不学习任何显式的模型，而是在预测阶段通过计算样本间的距离来进行分类。

其工作过程包括以下步骤：

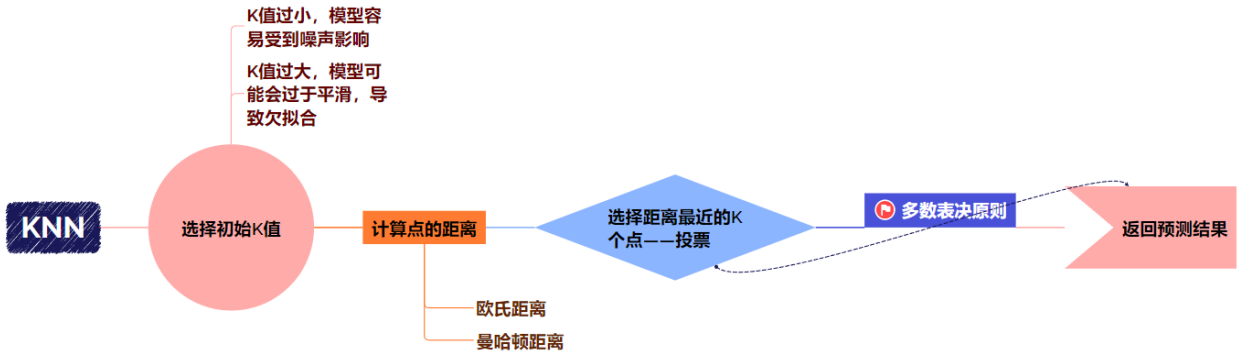



**图 2.3-1 KNN 模型流程图**

KNN 算法依赖于距离度量来计算样本之间的相似度。常见的距离度量方法包括：

欧氏距离（Euclidean Distance）： 欧氏距离是 KNN 中最常用的距离度量方法。它是两点之间直线距离的度量，计算公式为：

$$d(x,y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2} \tag{2.1.3}$$

其中 $x_i$ 和 $y_i$ 是样本点 x 与 y 在第 i 维上的特征值，n 是特征的维度。

曼哈顿距离（Manhattan Distance）： 曼哈顿距离也被称为“城市街区距离”，它是所有维度上差值的绝对值之和，计算公式为：

$$d(x,y)=\sum_{i=1}^{n}|x_i-y_i| \tag{2.1.4}$$

曼哈顿距离在处理某些类型的高维数据时，可能比欧氏距离更有效。

## 2.4 决策树（Decision Tree）

决策树是一种常用于分类与回归分析的机器学习模型。其核心思想是通过递归地将数据集划分成不同的子集，从而形成一颗树状结构。每一个内部节点代表一个特征的判断条件，每一个叶子节点表示最终的分类标签或回归值。决策树模型通过一系列条件判断将数据从根节点沿树枝流向叶节点，以达到预测的目的。

决策树的构建基于某种度量标准来选择最佳特征进行分裂，常用的分裂标准包括**信息增益**（Information Gain）、**基尼指数**（Gini Impurity）和**均方误差**（Mean Squared Error，MSE）。信息增益常用于 ID3 和 C4.5 算法中，而 CART 算法则采用基尼指数作为分裂标准。决策树通过选择能够最有效区分样本的特征进行分裂，直到达到预设的停止条件，如树的深度达到最大值，或节点的纯度满足要求。

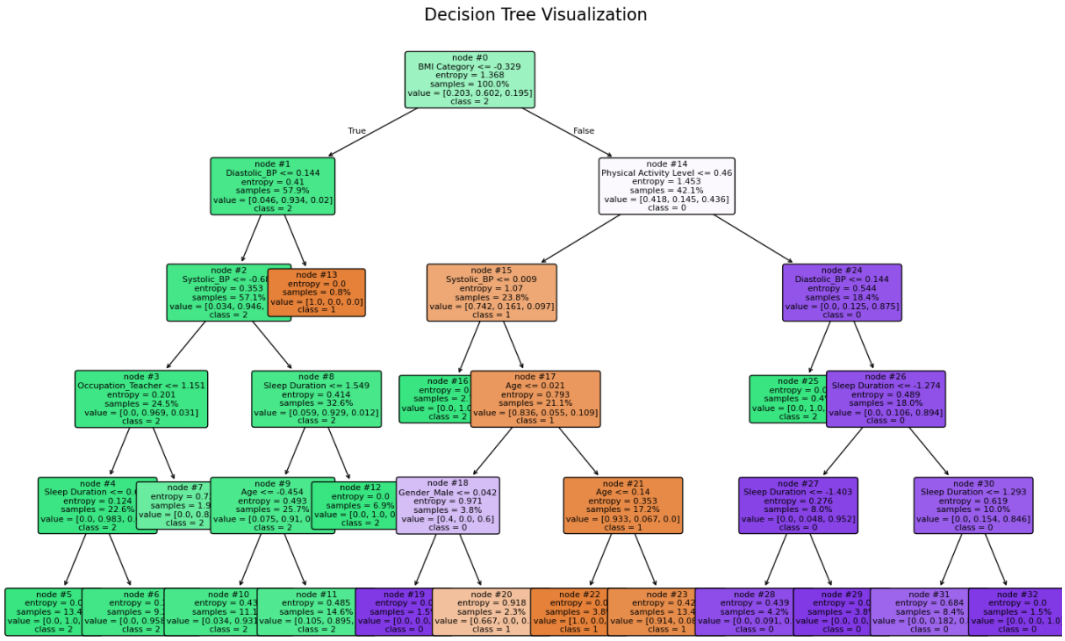



**图 2.4-1 本研究中的决策树**

为了解决过拟合问题，通常会采用**剪枝**技术。剪枝可以分为**预剪枝**和**后剪枝**，前者在树构建过程中限制树的生长，后者则在树构建完成后去除多余的分支。

## 2.5 朴素贝叶斯（Naive Bayes）

朴素贝叶斯是一类基于贝叶斯定理的概率分类方法。其核心假设是特征之间条件独立，即给定类别的情况下，各特征之间互不依赖。虽然这种假设在现实中往往不成立，但朴素贝叶斯在许多实际应用中表现良好，特别是在文本分类和其他高维数据处理中。

贝叶斯定理的形式为：

$$P(C \mid X) = \frac{P(X \mid C)P(C)}{P(X)} \tag{2.1.5}$$

其中，P(C|X)是给定输入特征 X 下属于类别 C 的后验概率，)P(X|C)是类别 C 下特征 X 的似然，P(C)是类别的先验概率，P(X)是特征 X 的边际概率。

本研究所用高斯朴素贝叶斯（Gaussian Naive Bayes）是朴素贝叶斯分类器的一种常见变体：假设特征是连续的，并且符合高斯分布。适用于数值型数据。

# 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，处理了数据的缺失值，并且对其进行可视化分析直观描述数据集。

## 3.1 变量简介

数据来源于 Kaggle 机器学习数据库的睡眠健康和生活方式数据集，除却 ID 标识符，此数据集由 374 条数据和 12 个特征属性组成，主要分为分类变量与数值型变量。变量具体介绍如下表 3.1-1 数据变量说明表。

**表 3.1-1 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<br>F=女性 |
| 职业 | Occupation | 分类变量（11 类水平） | 个人的职业或专业（如护士、律师等） |
| BMI 类别 | BMI Category | 分类变量（4 类水平） | 个人的 BMI 类别（偏瘦、正常、超重、肥胖） |
| 睡眠障碍 | Sleep Disorder | 分类变量（3 类水平） | 无、失眠、睡眠呼吸暂停 |

**表 3.1-2 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 | 人的年龄（以岁为单位） |
| 睡眠持续时间 | Sleep Duration | 数值型 | 每天的睡眠小时数 |
| 睡眠质量 | Sleep Quality | 数值型 | 对睡眠质量的主观评分，范围从 1 到 10 |
| 身体活动水平 | Physical Activity Level | 数值型 | 每天进行身体活动的分钟数 |
| 压力水平 | Stress Level | 数值型 | 对压力水平的主观评分，范围从 1 到 10 |
| 血压 | Blood Pressure | 数值型 | 收缩压/舒张压的测量值 |
| 心率 | Heart Rate | 数值型 | 静息心率，以每分钟心跳次数为单位 |
| 每日步数 | Daily Steps | 数值型 | 每天走的步数 |

## 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：

**（1）缺失值**

本数据集较为完整，对分类变量与数值型变量进行必要的缺失值检测，结果显示均无缺失值。

**（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。着重对数据集中的血压、休眠时长两项数值型可测量变量进行异常值检测，正常（非病患）人体一般血压极限：高压[90,180];低压[60,120],休眠时长正常范围[5,10]。常用的方法包括箱线图分析和 Z-score 法，通过这些方法识别出数据中与大多数数据点差距较大的值。

结果显示，二者均无异常值。对于剩余变量，检测其是否在值域或类别内，结果显示并未有超出设定范围的变量值，即无异常值。

**（3）分类变量转换**

在本研究中，数据集包含四个分类变量，其中包括目标变量睡眠障碍类别（Sleep Disorder）和三个特征变量（BMI Category、Gender、Occupation）。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）。对不同变量适宜选取不同转换方式，本研究以 Gender 及 BMI Category 为例列出其对应：

**表 3.2-1 分类变量转换说明表**

| Gender ⟶ 独热编码 | BMI Category ⟶ 标签编码 |
|---|---|
| **Male → [1, 0]**<br>**Female→ [0, 1]** | **Normal → 0**<br>**Overweight → 1**<br>**Obese → 2** |

本研究中的 BMI Category 体重类别特征变量存在顺序关系，因此固定选取标签编码格式。而余下的 Gender 与 Occupation 两种特征变量是否需要由默认的标签编码转为使用独热编码，对于选取的五种分类算法，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 3.2-2 模型编码说明表**

| 模型 | 适合的编码方式 | 解释 |
|---|---|---|
| SVC | 独热编码 | 对数值关系敏感，独热编码确保类别之间没有顺序关系。 |
| MLP | 独热编码 | 对数值关系敏感，独热编码能避免误解类别之间的顺序关系。 |
| 决策树 | 标签编码 | 不依赖于数值关系，标签编码直接将类别转化为整数即可。 |
| KNN | 独热编码 | 距离度量敏感于类别的数值化，独热编码避免了类别之间的顺序关系问题。 |

| | | |
|---|---|---|
| 朴素贝叶斯 | 标签编码 | 适合使用标签编码，计算条件概率时不会误解类别之间的关系。 |

## 3.3 数据可视化及分析

对数据集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健康和生活方式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 11 个特征变量。

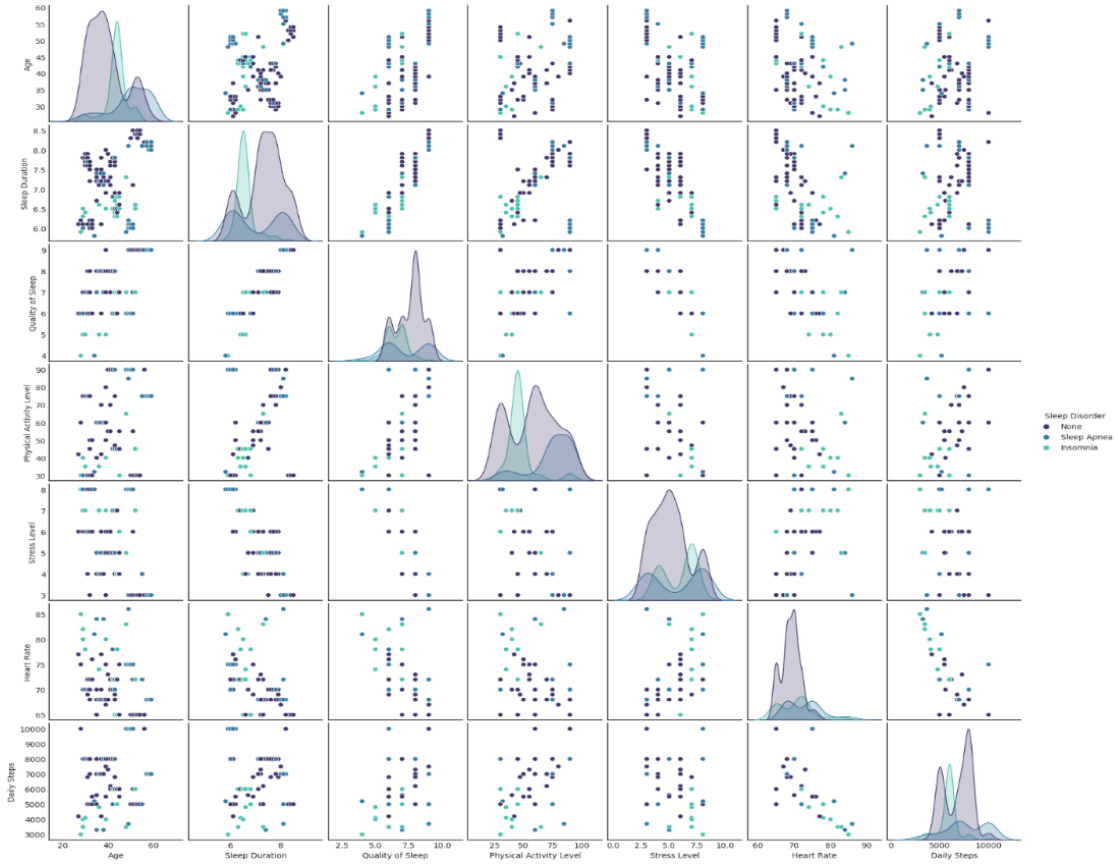



**图 3.3-1 成对变量关系图**

### 3.3.1 目标变量

为了直观展示“睡眠障碍”（Sleep Disorder）这一目标变量的分布情况，我们采用了柱状图进行可视化分析。

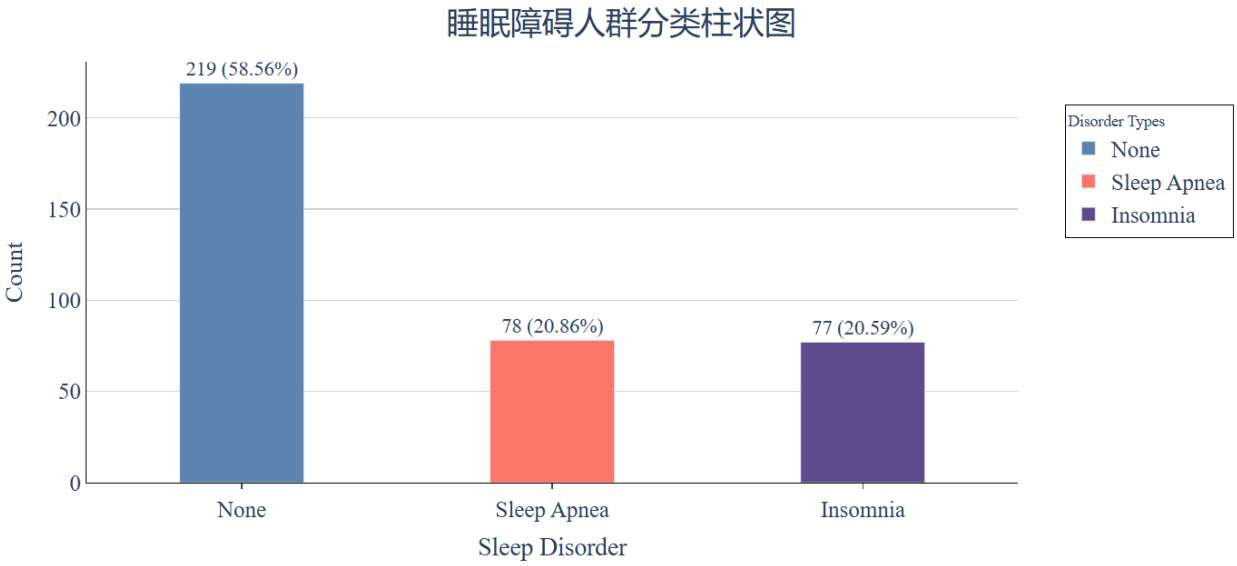



**图 3.3-2 睡眠障碍人群分类柱状图**

从图 3.3-2 可以看出，约 58.56%的受访者没有睡眠障碍，而约 41.44%的受访者存在睡眠障碍。这一结果与实际调查中睡眠障碍的普遍程度一致，其中睡眠呼吸暂停和失眠群体的比例接近，反映了这两种障碍在调查样本中的较高发病率。

### 3.3.2 性别与年龄

性别可能与睡眠障碍的发生存在一定关系。图 3.3-3 展示了性别与睡眠障碍类型之间的关系。

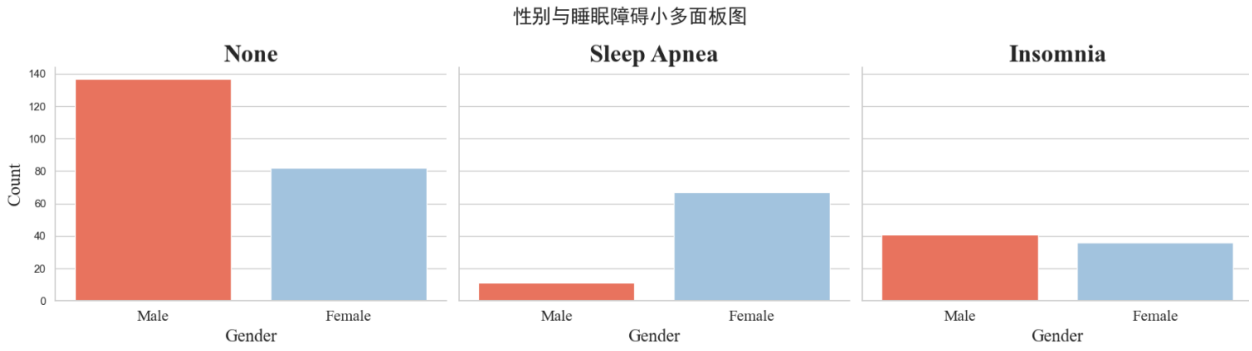



**图 3.3-3 性别与睡眠障碍小多面板图**

可以看出，男性在无睡眠障碍群体中的比例较高，而女性在睡眠呼吸暂停群体中的比例较高。值得注意的是，现有研究表明，男性与女性患睡眠呼吸障碍的比例约为 2:1，

这可能是由于样本量较小而产生的偏差。此外，失眠群体的性别比例较为均衡，表明该障碍对男女的影响较为普遍。

对于睡眠障碍病症，相比于性别，年龄可能更具有研究价值与分析可行性。

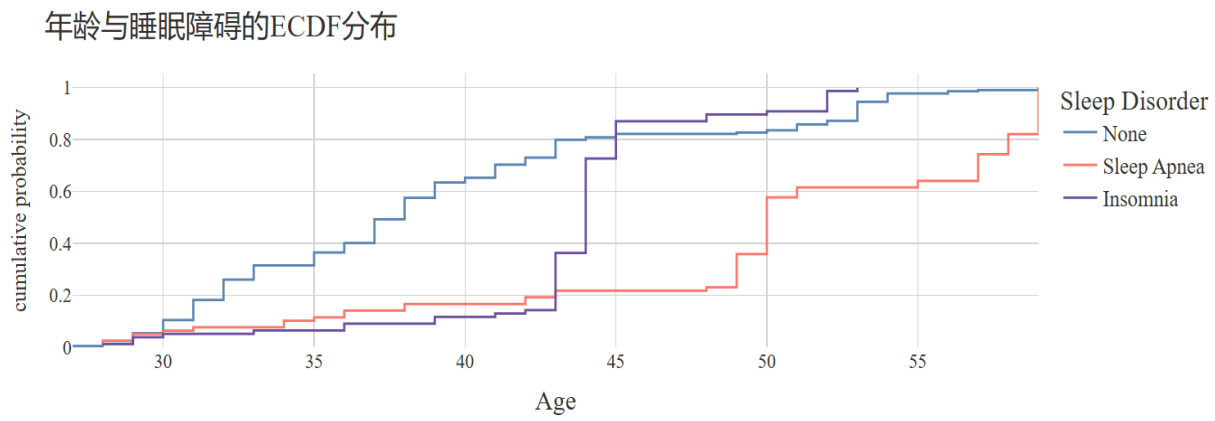



**图 3.3-4 睡眠障碍类别经验累积分布函数（ECDF）图**

上图表现出无障碍、睡眠呼吸暂停、失眠三类表现随年龄的累计变化，从中可以直观地观察到三类表现的阶段性概率增长变化及对比。

其中，为更细致地调查各年龄段患病概率及最大概率患病年龄，画出年龄段中三类表现概率分布图像：

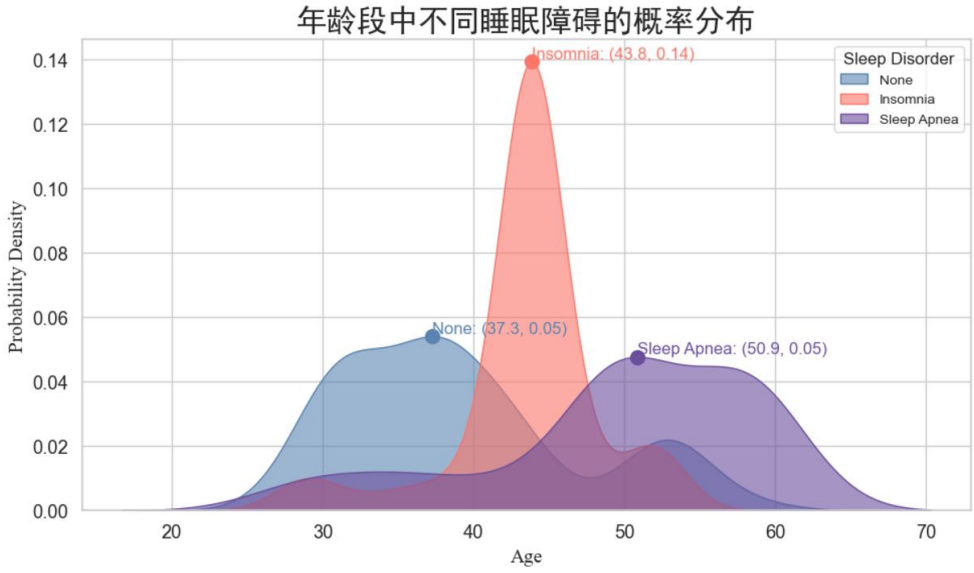



**图 3.3-5 各年龄段不同睡眠障碍的概率分布情况**

结果显示，43 岁左右的群体失眠的概率最高，而 51 岁左右群体患睡眠呼吸暂停的概率最大。

### 3.3.3 其它变量

除了性别和年龄，其他变量也与睡眠障碍的发生具有重要关系。例如，职业、身体活动水平、压力水平和体重等因素都可能对睡眠健康产生影响。


**图 3.3-6 睡眠质量与睡眠障碍太阳辐射图**

从图 3.3-6 可以看出，无睡眠障碍群体的睡眠质量普遍较高，而失眠群体的睡眠质量较差。睡眠呼吸暂停群体的睡眠质量则呈现较大差异，这可能与部分患者对呼吸暂停的感知较弱或无感知有关。

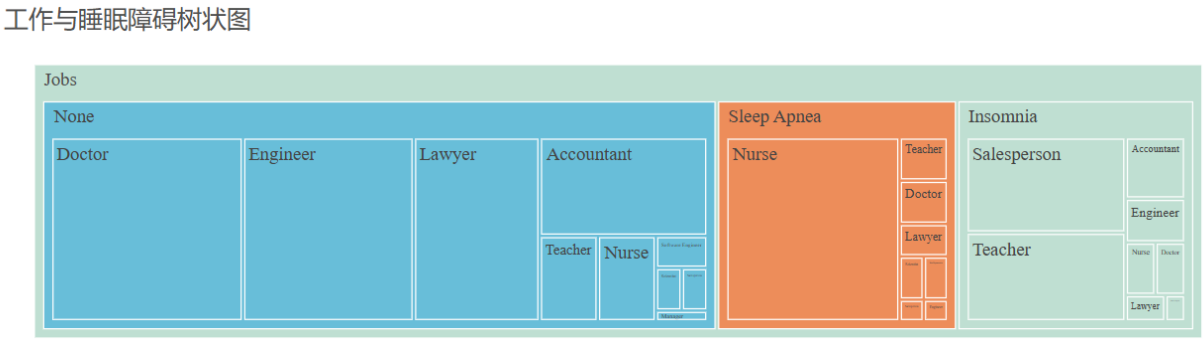



**图 3.3-7 工作与睡眠障碍树状图**

图 3.3-7 展示了不同职业群体中睡眠障碍的分布情况。调查结果表明，失眠群体主要集中在销售员和教师等高压力职业群体中，而睡眠呼吸暂停群体则主要集中在护士职业。这可能与职业性质和工作环境的差异密切相关。

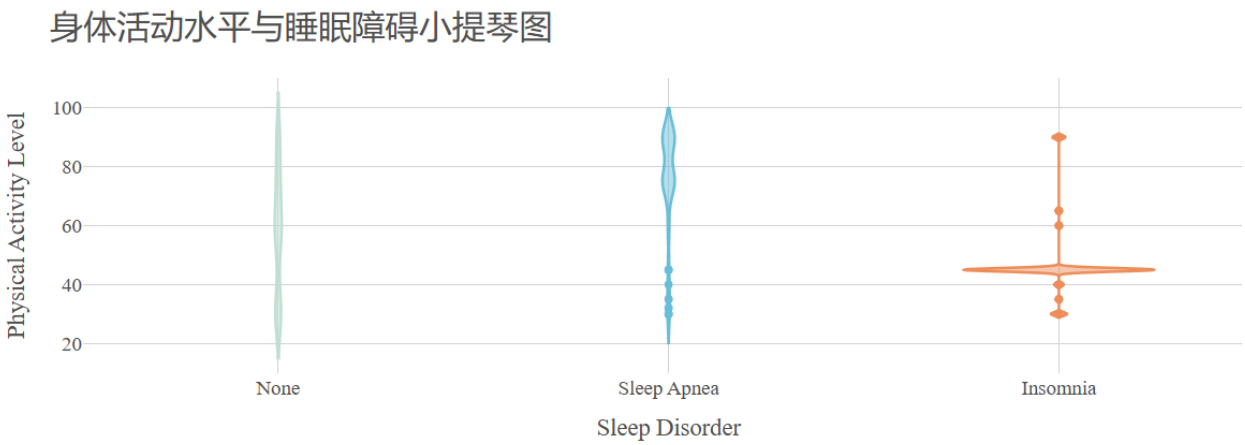



**图 3.3-8 身体活动水平与睡眠障碍小提琴图**

由上图可以看到，较为突出的是，大部分失眠群体的身体活动水平维持在较低水平；其余两类群体的身体活动水平较为平均，并未表现出明显差异。

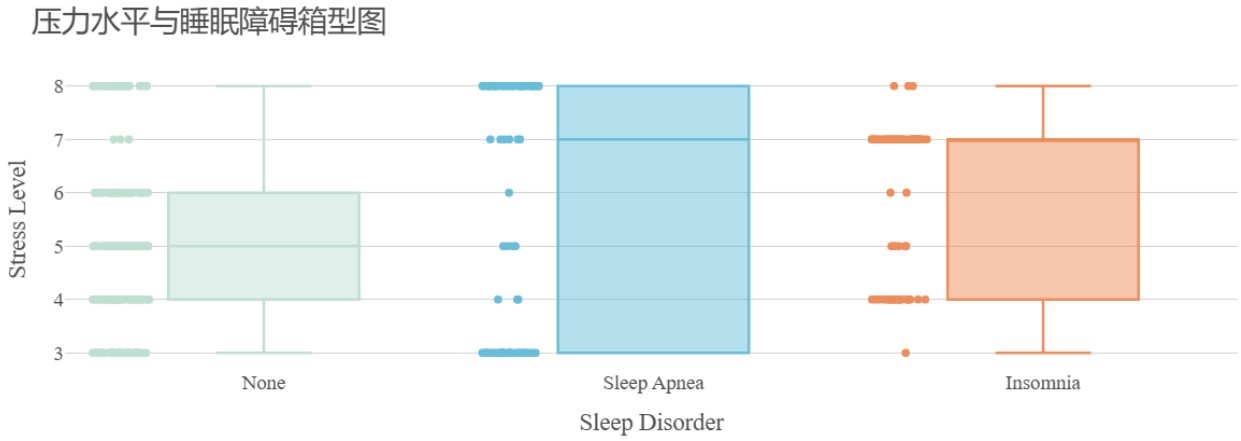



**图 3.3-9 压力水平与睡眠障碍箱型图**

由上图可以看到，无睡眠障碍群体的压力水平显著低于存在睡眠障碍群体，中位数为 5；呼吸暂停群体与失眠群体中位数一致，均为 7；其中，睡眠呼吸暂停群体压力水平差异较大——压力水平为 3 和 8 的人数均较多。

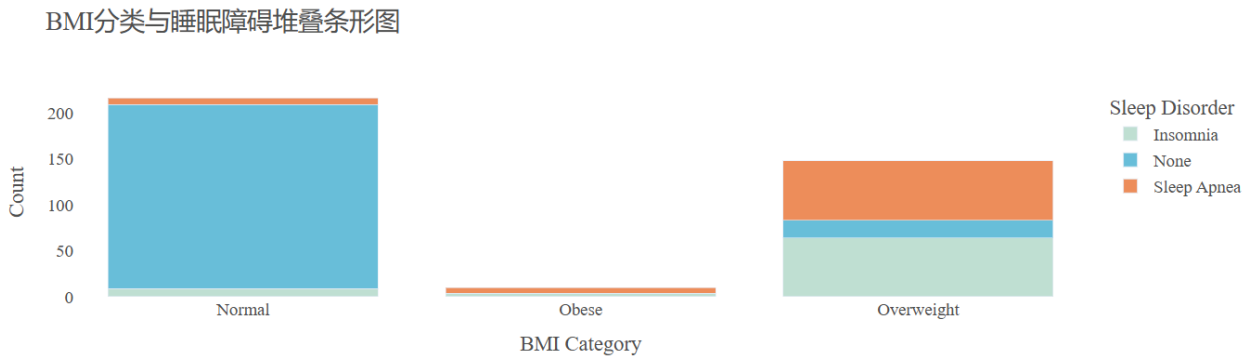



**图 3.3-10　BMI 与睡眠障碍堆叠条形图**

由上图可以看到，正常体重群体大多数无睡眠障碍；超重群体患病人数较多，其中，无失眠与睡眠呼吸暂停的人数相当；肥胖群体调查人数较少，但大多数均患有睡眠呼吸暂停病症，这与其发病机理相吻合。

# 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、SMOTH 过采样、标准化）及正式构建模型三大部分。其中，第三部分包括对支持向量机（SVC）、多层感知机（MLP）、决策树、K 近邻（KNN）及朴素贝叶斯五大模型的训练、优化、评估。

## 4.1　模型评价指标

### 4.1.1 指标介绍

为全面评估模型的分类性能，本研究采用常用分类评价指标：准确率（Accuracy）、精确率（Precision）、召回率（Recall）、F1-score、ROC 曲线下的面积（AUC-ROC）以及混淆矩阵（Confusion Matrix）。接下来以混淆矩阵为例详细介绍 5 项分类评价指标：

**表 4.1-1 二分类混淆矩阵**

| 实际值 | 预测值 | |
|---|---|---|
| | 正类 | 负类 |
| 正类 | TP | FN |
| 负类 | FP | TN |

其中，TP（True Positive）表示正类被正确分类的样本数，TN（True Negative）表示负类被正确分类的样本数，FP（False Positive）和 FN（False Negative）分别表示误判为正类和误判为负类的样本数。

（1）准确率（Accuracy）：评价模型整体性能的重要指标。定义为正确分类样本数量占总样本数量的比例，计算公式为：

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.1.1}$$

（2）精确率（Precision）：衡量模型预测为正类的样本中实际为正类的比例。计算公式为：

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4.1.2}$$

（3）召回率（Recall）：也称灵敏度，衡量实际为正类的样本中被正确分类的比例。计算公式为：

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4.1.3}$$

（4）F1 分数（F1-Score）：综合评估模型性能的有效指标。是精确率和召回率的调和平均数，用于平衡两者之间的权衡关系。计算公式为：

$$F1 - Score = \frac{2*\text{Precision}*\text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.1.4}$$

（5）ROC 曲线和 AUC：

ROC 曲线（Receiver Operating Characteristic Curve）反映模型在不同分类阈值下的表现。曲线横轴为假正率（False Positive Rate, FPR），纵轴为真正率（True Positive Rate, TPR），定义如下：

$$FPR = \frac{FP}{FP + TN}, TPR = \frac{TP}{TP + FN} \tag{4.1.5}$$

ROC 曲线下的面积（Area Under the Curve, AUC）是衡量模型区分能力的关键指标，其取值范围为[0,1]。AUC 越接近 1，说明模型区分正负类的能力越强。当 AUC 接近 0.5 时，模型等同于随机猜测。

（6）Kappa 系数

Kappa 系数（Cohen's Kappa）衡量模型分类结果与随机猜测之间的一致性，定义为：

$$Kappa = \frac{p_o - p_e}{1 - p_e} \tag{4.1.6}$$

其中，$p_o$ 为实际分类准确率，$p_e$ 为随机猜测的期望准确率。Kappa 系数在评估类别分布不均的情况下尤为有用。

### 4.1.2 评价指标的适用性

本研究的目标是对睡眠障碍的类别进行分类，无睡眠障碍、睡眠呼吸暂停、失眠患者分别为 219、78、77 个样本。为全面评估模型在处理类别不平衡问题中的性能，本研究在采用 SMOTE 技术对训练集进行过采样后，选择宏平均（Macro-average）和加权平均（Weighted-average）两种方法对测试集进行评价。

**（1）宏平均（Macro-average）**

宏平均通过对每个类别的分类指标（如精确率、召回率和 F1-score）独立计算后，取这些指标的算术平均值，其计算公式为：

$$Macro - Precision = \frac{1}{C}\sum_{i=1}^{C} Precision \tag{4.2.1}$$

$$Macro - Recall = \frac{1}{C}\sum_{i=1}^{C} Recall_{\ i} \tag{4.2.2}$$

$$Macor - F1 = \frac{1}{C}\sum_{i=1}^{C} F1_i \tag{4.2.3}$$

其中，C 为类别总数，$Precision_i$、$Recall_i$ 和 $F1_i$ 分别是第 i 类的精确率、召回率和 F1-score。

宏平均对每个类别的评价指标（如精确率、召回率和 F1-score）独立计算后取算术平均，忽略类别分布的影响，能够反映模型对所有类别的分类性能是否公平。在本研究中，宏平均用于验证 SMOTE 技术是否有效改善了少数类别的分类能力。

**（2）加权平均（Weighted-average）**

加权平均根据每个类别的样本比例，对分类指标进行加权计算，其公式为：

$$Weighted-Precision=\sum_{i=1}^{C}(\frac{N_i}{N}\cdot Precision_i) \tag{4.2.4}$$

$$Weighted-Recall=\sum_{i=1}^{C}(\frac{N_i}{N}\cdot Recall_i) \tag{4.2.5}$$

$$Weighted-F1=\sum_{i=1}^{C}(\frac{N_i}{N}\cdot F1_i) \tag{4.2.6}$$

其中，$N_i$为第 i 类的样本数量，N 为总样本数量。

加权平均根据类别的样本比例对分类指标进行加权计算，更贴近测试集中类别的真实分布。加权平均能够反映模型在实际应用场景中的整体分类性能，是衡量模型全局表现的重要指标。

## 4.2 数据划分

为了实现训练集和测试集的合理划分，本研究采用了分层抽样（Stratified Sampling）方法。分层抽样能够确保在数据划分过程中，每个类别的样本比例在训练集和测试集中保持一致，从而减少类别不平衡对模型训练和评估的影响。数据集的划分比例为 7:3，即 70%的数据用于训练，30%的数据用于测试。其中，70%的训练集需进一步进行 SMOTH 过采样，以便在模型训练过程中最大程度上减轻类别不平衡的影响。

| | None | Sleep Apnea | Insomnia | 样本合计 |
|---|---|---|---|---|
| 训练集 | 153 | 54 | 54 | 261 |
| 测试集 | 66 | 24 | 23 | 113 |
| 样本合计 | 219 | 78 | 77 | 374 |

## 4.3 SMOTH 数据过采样

由于训练集存在类别不平衡问题，类别 1 的样本数量远大于类别 2 和类别 3，可能导致模型在训练过程中偏向于多数类别，从而影响少数类别的分类效果。本研究在训练集上应用了 SMOTE（Synthetic Minority Oversampling Technique）过采样技术。

SMOTE 通过生成合成样本，扩充少数类别的样本量，从而平衡训练集中的类别分布。该方法通过对少数类别样本的邻近样本进行插值生成新样本，有效提高了模型对少数类别的学习能力，本研究中过采样后的训练集由 261 份增至 471 份。

**表 4.3-1 SMOTE 采样后部分训练数据集**

| 变量名称 | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| Age | 30 | 58 | 33 | 41 | 29 | 43 |
| Sleep Duration | 7.6 | 8.0 | 6.0 | 7.1 | 6.5 | 6.5 |
| Quality of Sleep | 7 | 9 | 6 | 7 | 5 | 6 |
| Physical Activity Level | 75 | 75 | 30 | 55 | 40 | 45 |
| Stress Level | 6 | 3 | 8 | 6 | 7 | 7 |
| BMI Category | 0 | 2 | 0 | 2 | 0 | 2 |
| Heart Rate | 70 | 68 | 72 | 72 | 80 | 72 |
| Daily Steps | 8000 | 7000 | 5000 | 6000 | 4000 | 6000 |
| Systolic_BP | 120 | 140 | 125 | 125 | 132 | 130 |
| Diastolic_BP | 80 | 95 | 80 | 82 | 87 | 85 |
| Gender_Male | True | False | True | True | False | True |
| Occupation_Doctor | True | False | True | False | False | False |
| Occupation_Engineer | False | False | False | False | False | False |
| Occupation_Lawyer | False | False | False | True | False | False |
| Occupation_Manager | False | False | False | False | False | False |
| Occupation_Nurse | False | True | False | False | True | False |
| Occupation_Sales Representative | False | False | False | False | False | False |
| Occupation_Salesperson | False | False | False | False | False | True |
| Occupation_Scientist | False | False | False | False | False | False |
| Occupation_Software Engineer | False | False | False | False | False | False |
| Occupation_Teacher | False | False | False | False | False | False |

| | | | | | | |
|---|---|---|---|---|---|---|
| Sleep Disorder | 1 | 2 | 1 | 1 | 0 | 0 |

## 4.4 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（本研究中为支持向量机（SVC）、多层感知机（MLP）和 K 近邻三种算法）的性能，本研究对数值特征进行了标准化处理。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，$\mu$ 为特征的均值，$\sigma$ 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。

## 4.5 模型构建

### 4.5.1 支持向量机（SVC）

首先，使用支持向量分类（SVC）进行建模。SVC 是一个非常强大的分类模型，适用于高维度数据，尤其是对于分类边界不明确的情况表现良好。

**（1） 预处理选取**

- 特征变量编码：性别（Gender）和职业（Occupation）被转换为独热编码，以确保这些分类特征不被误处理为数值型。BMI 类别（BMI Category）使用标签编码。
- 数据标准化：所有特征变量（包括独热编码后的变量）都经过标准化处理，以确保每个特征在相同的尺度下进行训练，从而避免某些特征因数值范围过大而对模型产生较大影响。

**（2） 超参数调整**

该 SVC 模型采用径向基函数核（RBF kernel），该核函数能够有效地处理非线性可分的数据。为了得到最优的 SVC 模型，本研究通过最大化训练集交叉验证得分进行超参数搜索，选择最佳的惩罚参数 C 和 gamma 值。超参数搜索空间包括 C 值的范围：[0.1, 1.0, 10.0]及 gamma 的选择：['scale', 'auto']。

**表 4.5-1 SVC 超参数调优过程**

| C | gamma | 交叉验证中平均测试得分 |
|---|---|---|
| 0.1 | scale | 0.902396 |
| 0.1 | auto | 0.902396 |
| 1.0 | scale | 0.938477 |
| 1.0 | auto | 0.938477 |
| 10.0 | scale | 0.936372 |
| 10.0 | auto | 0.936372 |

结果表明，C=1.0 和 gamma='scale'在交叉验证过程中提供了最佳的性能表现。

### （3） 模型评估

根据（2）的结果，最终选用最佳超参数(C=1.0 和 gamma='scale')训练了最终的 SVC 模型并对测试集进行预测，得到混淆矩阵及 ROC-AUC 曲线：

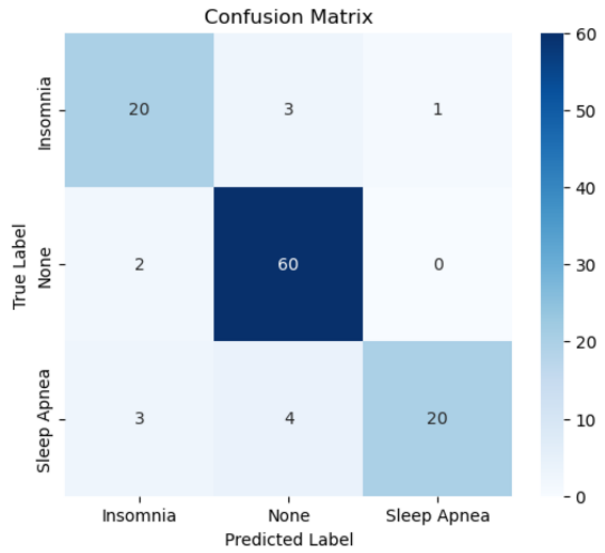



**图 4.5-1 SVC 混淆矩阵**

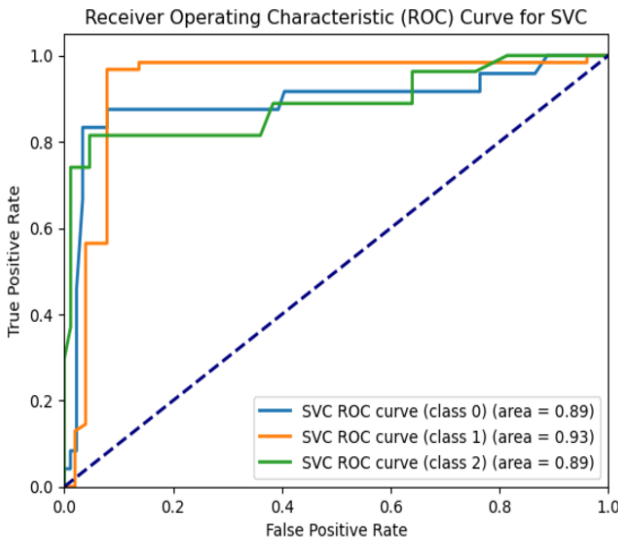



**图 4.5-2 SVC 的 ROC-AUC 曲线**

由混淆矩阵得出该模型准确率为 0.88，表示模型整体上在所有类别上的预测准确性较高，能正确分类 88%的样本。同时，可以观察到，AUC 曲线并不平滑，这是由于选取数据集较小导致的。AUC 值越高，表示模型的分类性能越好。该模型预测三种类别（失眠、无障碍、睡眠呼吸暂停）的 AUC 值分别为 0.89、0.93、0.89，说明该 SVC 模型在测

试集上有较高的分类性能。

左侧混淆矩阵进一步得出分类评价指标：

**表 4.5-2 SVC 分类评价指标表**

| 类别 | 精确率 | 召回率 | F1 分数 | 样本数 |
|---|---|---|---|---|
| 失眠 | 0.80 | 0.83 | 0.82 | 24 |
| 无睡眠障碍 | 0.90 | 0.97 | 0.93 | 62 |
| 睡眠呼吸暂停 | 0.95 | 0.74 | 0.83 | 27 |

该 SVC 模型对失眠类别的识别效果较好，但仍存在一定的错误分类，尤其是在少数类别的召回上（如将部分失眠样本预测为无睡眠障碍）；对无睡眠障碍类别的识别表现非常出色，具有很高的精确率和召回率；对睡眠呼吸暂停类别，精确率较高（95%），表明大部分预测为睡眠呼吸暂停的样本是真正的睡眠呼吸暂停。然而，由于召回率只有 74%，这意味着类别不平衡的影响并未完全消除，仍有部分睡眠呼吸暂停的样本被误分类为其他类别（如无睡眠障碍或失眠）。

为进一步探究模型在实际应用中的表现，本研究将宏平均与加权平均两种指标计算方式进行对比：

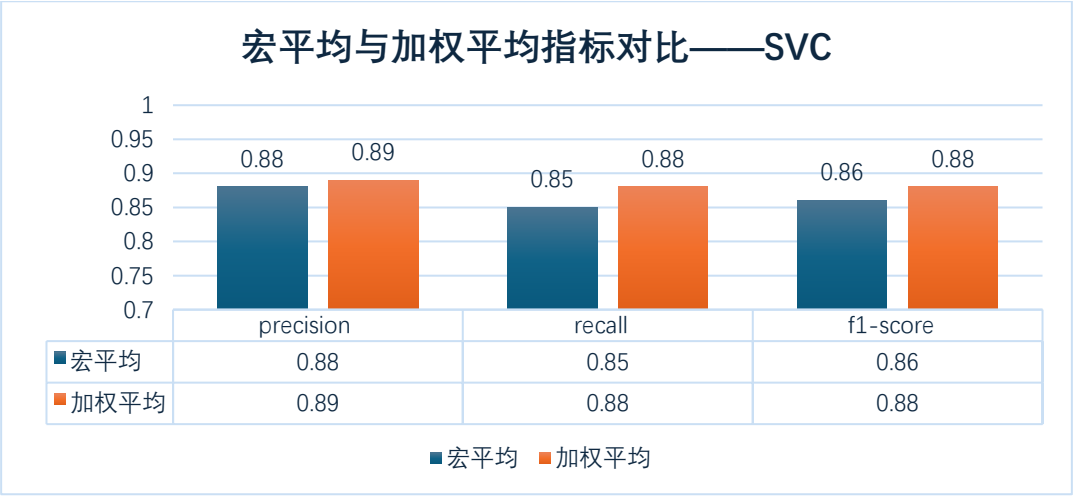



**图 4.5-3 宏平均与加权平均指标对比图——SVC**

结果显示：二者所得精确率、召回率与 F1 分数相差很小，说明类别不平衡问题在很大程度上解决，并不会影响 SVC 模型关于睡眠障碍类别预测的实际应用。

### 4.5.2 多层感知机（MLP）

MLP 模型是一种前馈神经网络，具有多个隐藏层，可以用于捕捉输入特征之间的复杂非线性关系。本研究建立该神经网络模型探究其表现，同时与其余四种机器学习方法进行比较。

**（1） 预处理选取**

- 特征变量编码：与支持向量机（SVC）模型相同，性别（Gender）和职业变量经过独热编码转换，以避免它们被误处理为数值型特征。BMI 类别使用标签编码，转换为数值类型。
- 数据标准化：所有特征变量（包括独热编码后的变量）都经过标准化处理。

**（2） 超参数调整**

为获得该 MLP 模型最佳表现，本研究使用了网格搜索（Grid Search）方法来调整 MLP 模型的超参数，以下是调优的超参数空间：

- 隐藏层大小（hidden_layer_sizes）: [(50, ), (100, ), (100, 50)]
- 激活函数（activation）: ['relu', 'tanh']
- 优化算法（solver）: ['adam', 'sgd']

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。参数调优过程如下：

**表 4.5-3 MLP 参数调优过程表**

| | 隐藏层大小 | 激活函数 | 优化器 | 交叉验证中平均测试得分 |
|---|---|---|---|---|
| 0 | (50,) | relu | adam | 0.940627 |
| 1 | (50,) | relu | sgd | 0.917312 |
| 2 | (100,) | relu | adam | 0.944860 |
| 3 | (100,) | relu | sgd | 0.885487 |

| | | | | |
|---|---|---|---|---|
| 4 | (100, 50) | relu | adam | 0.942732 |
| 5 | (100, 50) | relu | sgd | 0.904591 |
| 6 | (50,) | tanh | adam | 0.932161 |
| 7 | (50,) | tanh | sgd | 0.906697 |
| 8 | (100,) | tanh | adam | 0.932184 |
| 9 | (100,) | tanh | sgd | 0.898253 |
| 10 | (100, 50) | tanh | adam | 0.938567 |
| 11 | (100, 50) | tanh | sgd | 0.917312 |

最终选定的超参数组合为：隐藏层大小（hidden_layer_sizes）: (100,)；激活函数（activation）: 'relu'；优化算法（solver）: 'adam'。这一组合在交叉验证的过程中提供了最佳的性能表现，模型能够在训练集上以较高的准确率和稳定性进行训练。

**（3） 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 MLP 模型，并评估其在测试集上的表现。

首先可以观察模型训练过程中损失函数曲线趋势变化：

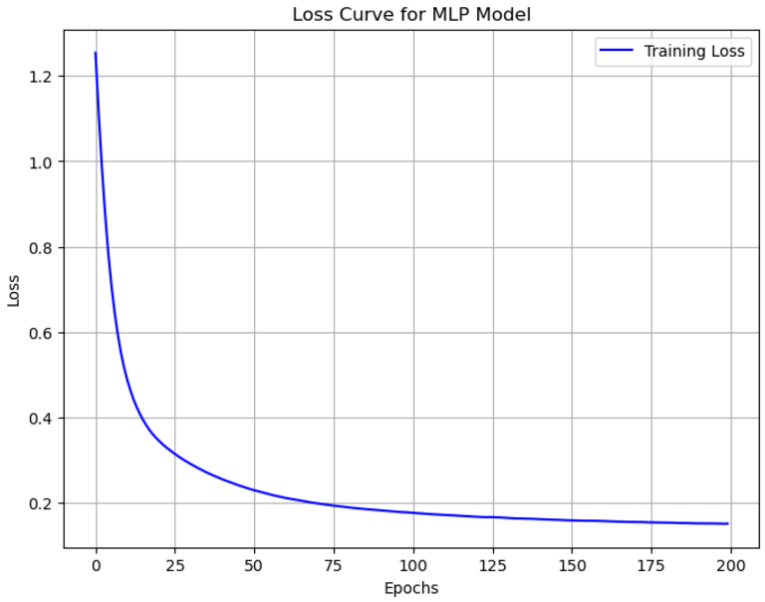



**图 4.5-4 MLP 损失函数曲线**

损失曲线平稳下降并趋于平稳，说明 MLP 模型正在有效学习并接近最优解。

训练完成进行预测，得到混淆矩阵及 ROC-AUC 曲线：

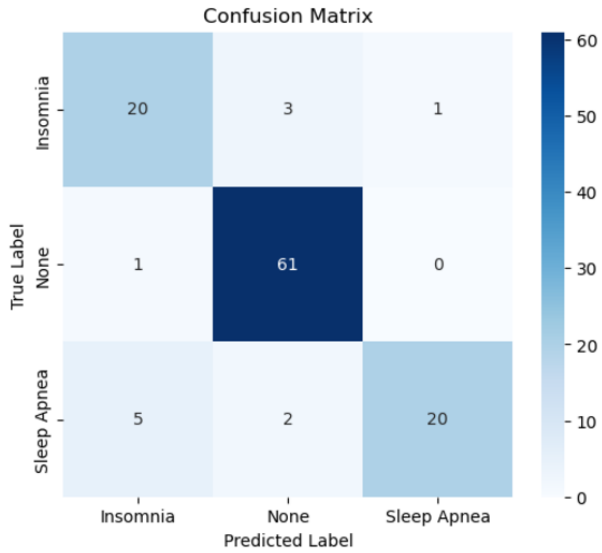



图 4.5-5 MLP 混淆矩阵

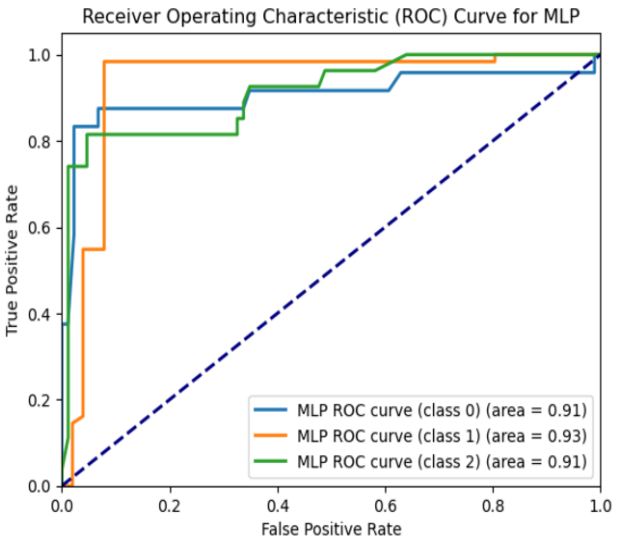



图 4.5-6 MLP 的 ROC-AUC 曲线

由混淆矩阵得出该模型准确率为 0.89，表示模型整体上在所有类别上的预测准确性较高，能正确分类 89%的样本。该模型预测三种类别（失眠、无障碍、睡眠呼吸暂停）的 AUC 值分别为 0.91、0.93、0.91，说明该 SVC 模型在测试集上有较高的分类性能。

左侧混淆矩阵进一步得出分类评价指标：

表 4.5-4 MLP 分类评价指标表

| 类别 | 精确率 | 召回率 | F1 分数 | 样本数 |
|---|---|---|---|---|
| 失眠 | 0.77 | 0.83 | 0.80 | 24 |
| 无睡眠障碍 | 0.92 | 0.98 | 0.95 | 62 |
| 睡眠呼吸暂停 | 0.95 | 0.74 | 0.83 | 27 |

各类别表现同 SVC 模型并无较大差异，值得注意的是，MLP 模型在无睡眠障碍类别的表现更为优秀（F1 分数提升 2%），而在失眠类别中表现稍有下降（F1 分数下降 2%）。这很可能与神经网络模型更强的拟合能力有关，MLP 能够通过层次结构和非线性激活函数捕捉到更多的特征信息。然而，可能过于关注较大样本类别，导致对小类别的泛化能力较差。

为进一步探究模型在实际应用中的表现，本研究将宏平均与加权平均两种指标计算方

式进行对比：

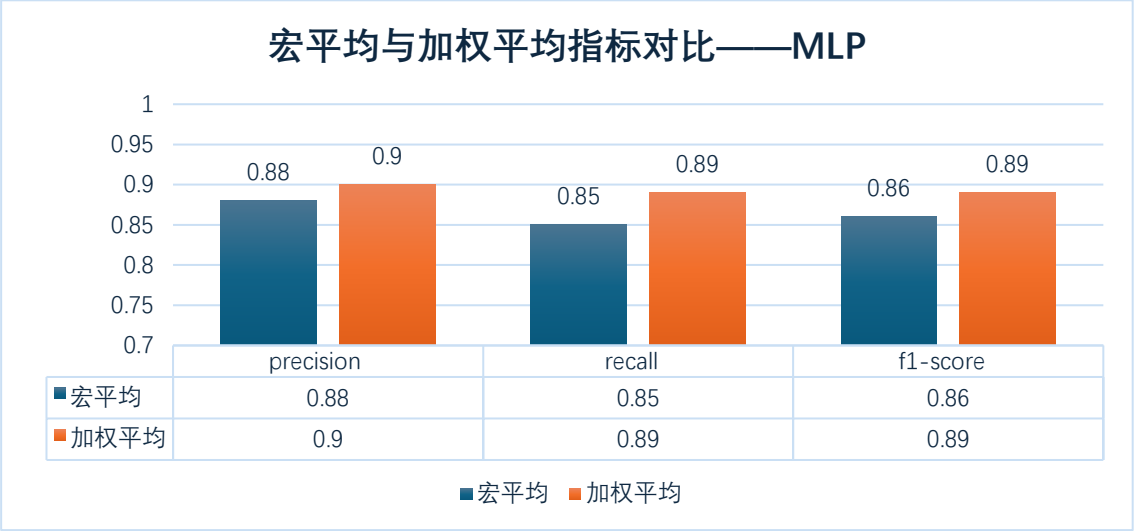



**图 4.5-7 宏平均与加权平均指标对比图——MLP**

结果显示：二者所得精确率、召回率与 F1 分数均较高（90%左右），说明 MLP 模型关于睡眠障碍类别预测的实际应用较为理想。

### 4.5.3 K 近邻（KNN）

K 近邻（K-Nearest Neighbors, KNN）是一种基于实例的监督学习算法，用于分类和回归任务。在分类问题中，KNN 通过测量样本间的距离来决定每个测试样本的分类。

**（1） 预处理选取**

➢ 特征变量编码：性别（Gender）和职业（Occupation）被转换为独热编码，BMI 类别（BMI Category）使用标签编码。

➢ 数据标准化：KNN 模型对特征的尺度敏感，因此在训练前，所有的特征都进行了标准化处理（Z-score 标准化）。标准化确保每个特征的均值为 0，标准差为 1，避免特征间尺度差异影响模型的性能。

**（2） 超参数调整**

KNN 算法的超参数包括：

- n_neighbors：K 值，即决定分类时考虑的邻居数目。
- weights：邻居样本的权重设置，选择"uniform"（所有邻居权重相同）或"distance"

（根据与当前样本的距离给邻居样本加权）。

- metric：距离度量方法，可以选择"euclidean"（欧氏距离）或"manhattan"（曼哈顿距离）。

**表 4.5-5 KNN 超参数调优流程表**

| | K 值 | 邻居样本权重设置 | 距离度量方法 | 交叉验证中平均测试得分 |
|---|---|---|---|---|
| 0 | 3 | uniform | euclidean | 0.902352 |
| 1 | 3 | distance | euclidean | 0.893863 |
| 2 | 5 | uniform | euclidean | 0.921501 |
| 3 | 5 | distance | euclidean | 0.917268 |
| 4 | 7 | uniform | euclidean | 0.910862 |
| 5 | 7 | distance | euclidean | 0.917268 |
| 6 | 9 | uniform | euclidean | 0.904524 |
| 7 | 9 | distance | euclidean | 0.913012 |
| 8 | 3 | uniform | manhattan | 0.910840 |
| 9 | 3 | distance | manhattan | 0.898096 |
| 10 | 5 | uniform | manhattan | 0.923606 |
| 11 | 5 | distance | manhattan | 0.915118 |
| 12 | 7 | uniform | manhattan | 0.917223 |
| 13 | 7 | distance | manhattan | 0.915118 |
| 14 | 9 | uniform | manhattan | 0.915118 |

| | | | | |
|---|---|---|---|---|
| 15 | 9 | distance | manhattan | 0.917245 |

结果表明，[5,’uniform’,’manhattan’]组合在交叉验证过程中表现出最佳的性能。

**（3） 模型评估**

经过交叉验证择优选取，最终选用最佳超参数([5,’uniform’,’manhattan’])对 KNN 模型进行训练并预测训练集，得到混淆矩阵及 ROC-AUC 曲线：

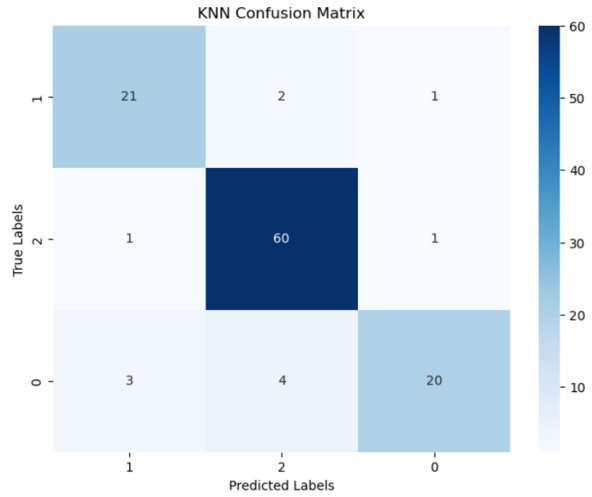



**图 4.5-8 KNN 混淆矩阵**

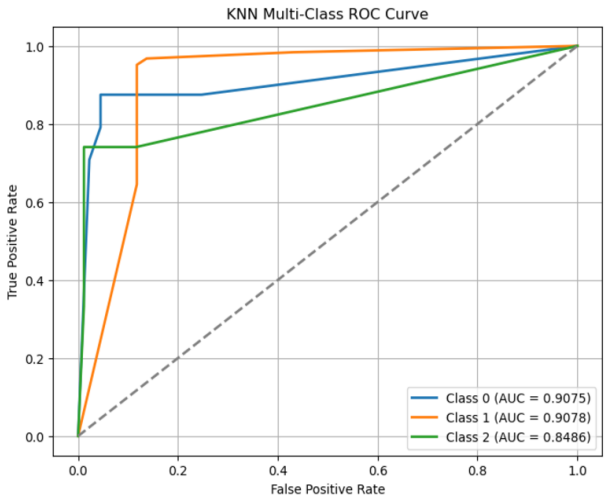



**图 4.5-9 KNN 的 ROC-AUC 图**

由混淆矩阵得出该模型准确率为 0.89，表示模型整体上在所有类别上的预测准确性较高，能正确分类 89%的样本。该模型预测三种类别（失眠、无障碍、睡眠呼吸暂停）的 AUC 值分别约为 0.91、0.91、0.85，说明 KNN 模型在测试集上有较高的分类性能。

左侧混淆矩阵进一步得出分类评价指标：

**表 4.5-6 KNN 分类评价指标**

| 类别 | 精确率 | 召回率 | F1 分数 | 样本数 |
|---|---|---|---|---|
| 失眠 | 0.84 | 0.88 | 0.86 | 24 |
| 无睡眠障碍 | 0.91 | 0.97 | 0.94 | 62 |
| 睡眠呼吸暂停 | 0.91 | 0.74 | 0.82 | 27 |

值得关注的是，KNN 模型对失眠这一少量类别的预测表现有较为明显提升。其余表现无较大差异，同样较为优秀。

为进一步探究模型在实际应用中的表现，本研究将宏平均与加权平均两种指标计算方式进行对比：

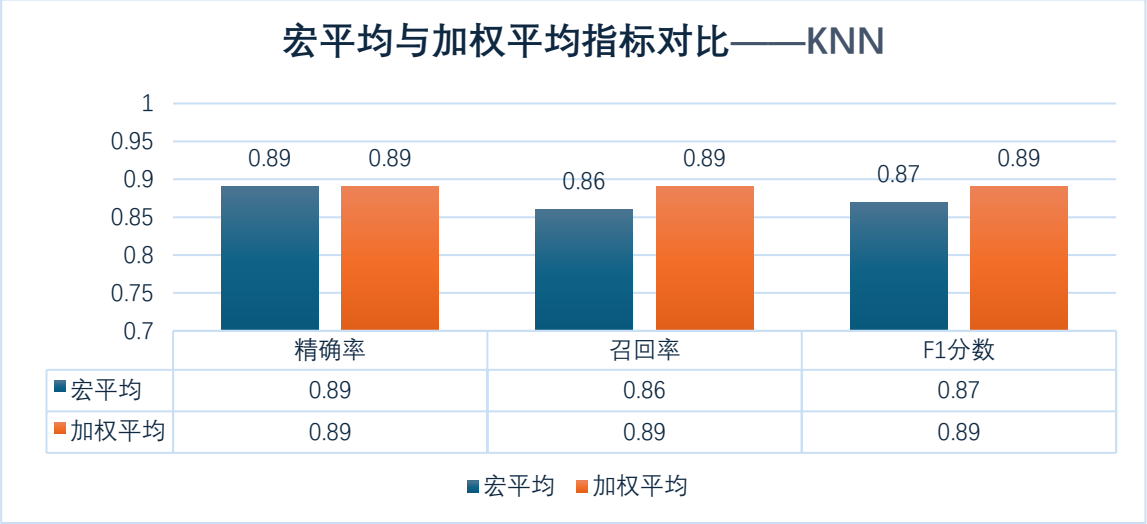



**图 4.5-10 宏平均与加权平均指标对比图——KNN**

结果显示：KNN 模型的宏平均和加权平均都较为接近，表明模型在各类别的表现较为均衡，并且整体分类效果较好。

### 4.5.4 决策树（Decision Tree）

决策树是常用的分类模型，具有良好的解释性和较强的适应性。通过递归地分裂数据集，决策树能够有效地处理特征与目标变量之间的非线性关系。

**（1） 预处理选取**

➢ 特征变量编码：为保证后续五种模型对比的一致严谨性，使用同一划分后训练测试集，故 Gender、Occupation 依旧选取独热编码。

➢ 数据标准化：本研究并未对决策树进行标准化处理，因为决策树模型对特征的尺度不敏感。标准化在树模型中不是必要步骤。

**（2） 超参数调整**

为了优化决策树模型，我们通过网格搜索（寻找交叉验证平均分数最大值）进行超参数调优。调优过程包括以下参数：

- max_depth：树的最大深度，控制模型的复杂度。

- min_samples_split：一个节点再分裂所需的最小样本数，防止过拟合。
- min_samples_leaf：叶节点的最小样本数，控制模型的复杂度。
- criterion：决策树分裂时的标准，这里选择 gini 和 entropy。

超参数搜索空间为 max_depth：5、10、20；min_samples_split：5、10、20；min_samples_leaf：1、2、4。criterion：['gini', 'entropy']。通过最大化交叉验证平均分数实现超参数搜索，选择最佳的组合。由于数据量较大，这里仅展示最大深度及最小样本分割数为例：

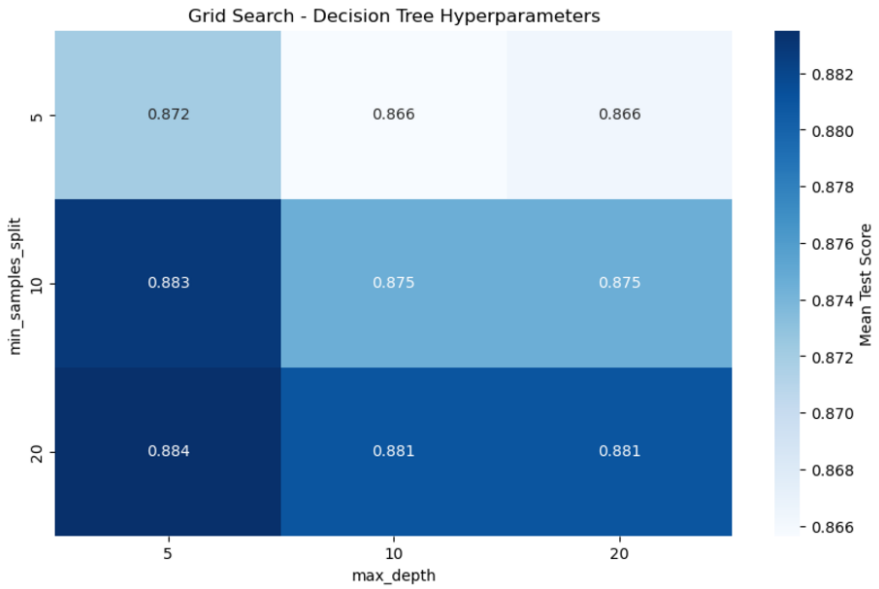
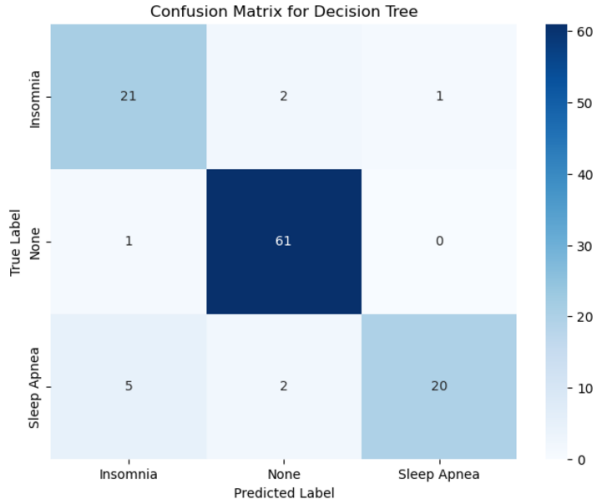



**图 4.5-11 决策树超参数调优图**

结果表明，最优组合为：最大深度 5、最小样本分割数 10、叶节点的最小样本数 1、决策树分裂时的标准 entropy。

### （3） 模型评估

选用最佳超参数训练决策树模型并对测试集进行预测，得到混淆矩阵及 ROC-AUC 曲线：



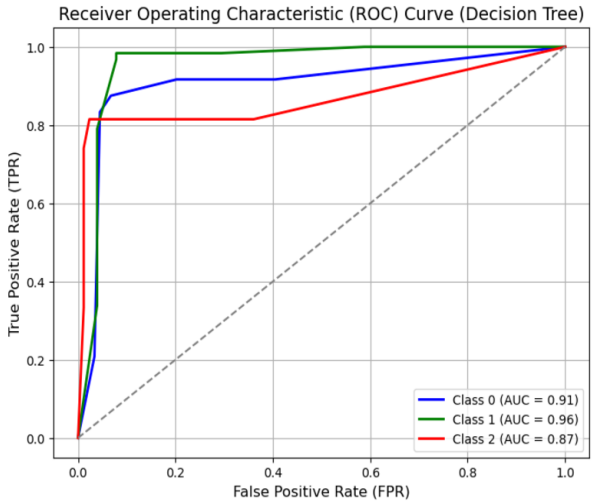




**图 4.5-12 决策树混淆矩阵**　　**图 4.5-13 决策树 ROC-AUC 曲线**

由混淆矩阵得出该模型准确率为 0.90，表示模型整体上在所有类别上的预测准确性较高，能正确分类 90%的样本。同时，该模型预测三种类别（失眠、无障碍、睡眠呼吸暂停）的 AUC 值分别为 0.91、0.96、0.87，说明该 SVC 模型在测试集上有较高的分类性能。

左侧混淆矩阵进一步得出分类评价指标：

**表 4.5-7 决策树分类评价指标表**

| 类别 | 精确率 | 召回率 | F1 分数 | 样本数 |
|---|---|---|---|---|
| 失眠 | 0.78 | 0.88 | 0.82 | 24 |
| 无睡眠障碍 | 0.94 | 0.98 | 0.96 | 62 |
| 睡眠呼吸暂停 | 0.95 | 0.74 | 0.83 | 27 |

该决策树模型在本研究中的表现较为优秀，能够有效地对不同睡眠障碍类别进行分类，特别是在无睡眠障碍类别上表现优异。尽管在睡眠呼吸暂停类别的召回率不够突出，但整体分类效果依然达到较高水平。

为进一步探究模型在实际应用中的表现，本研究将宏平均与加权平均两种指标计算方式进行对比：

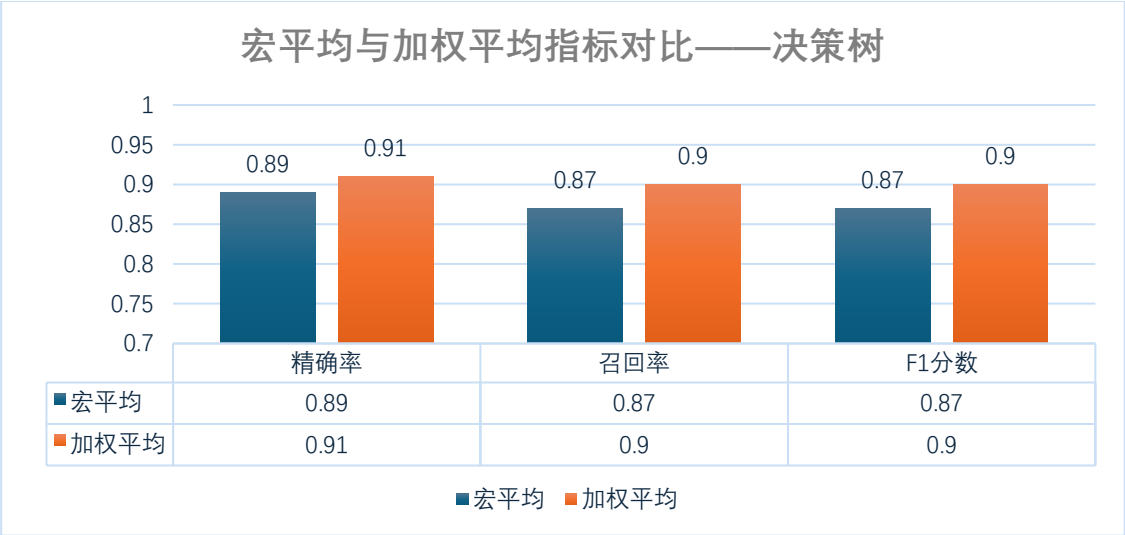



**图 4.5-14 宏平均与加权平均指标对比图——决策树**

最后，通过决策树模型绘制了特征重要性条形图：

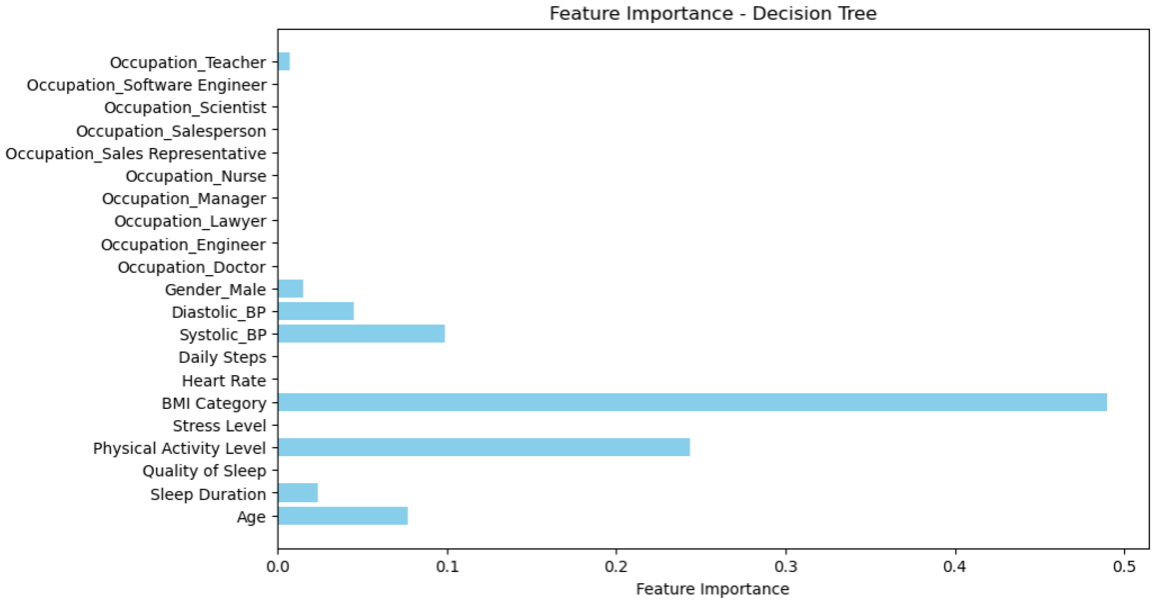



**图 4.5-15 决策树特征重要性条形图**

由图可知，较为重要的几个特征依次为：体重类别（BMI Category）、身体活动水平（Physical Activity Level）、收缩压（Systolic）、年龄（Age）等，其中，BMI 类别重要性尤甚，可见，控制体重是避免患睡眠障碍的关键。

由于本研究中的决策树模型在参数调优过程中已经间接控制了树的深度及叶节点大小，且模型本身并不复杂，于训练集及测试集均表现良好，因此不再进行剪枝。

### 4.5.5 高斯朴素贝叶斯

朴素贝叶斯分类器是一种基于贝叶斯定理的概率分类方法。它假设各个特征是条件独立的，因此适用于特征之间相关性较低的情况。本研究中，选择了高斯朴素贝叶斯（GaussianNB）模型。该模型假设特征遵循高斯（正态）分布，适用于数值型数据。

#### （1） 相关性分析

对于朴素贝叶斯模型是否能够较好地被应用，变量间的相关性是极为重要的因素。本研究首先分析了数据集中各个特征之间的相关性，以评估它们是否存在强烈的线性关系。高相关性的特征可能导致多重共线性问题，从而影响模型的稳定性和解释性。

下面对经过预处理（独热编码、SMOTH 过采样及标准化）后的训练数据集进行相关性分析，计算特征之间的皮尔逊相关系数：

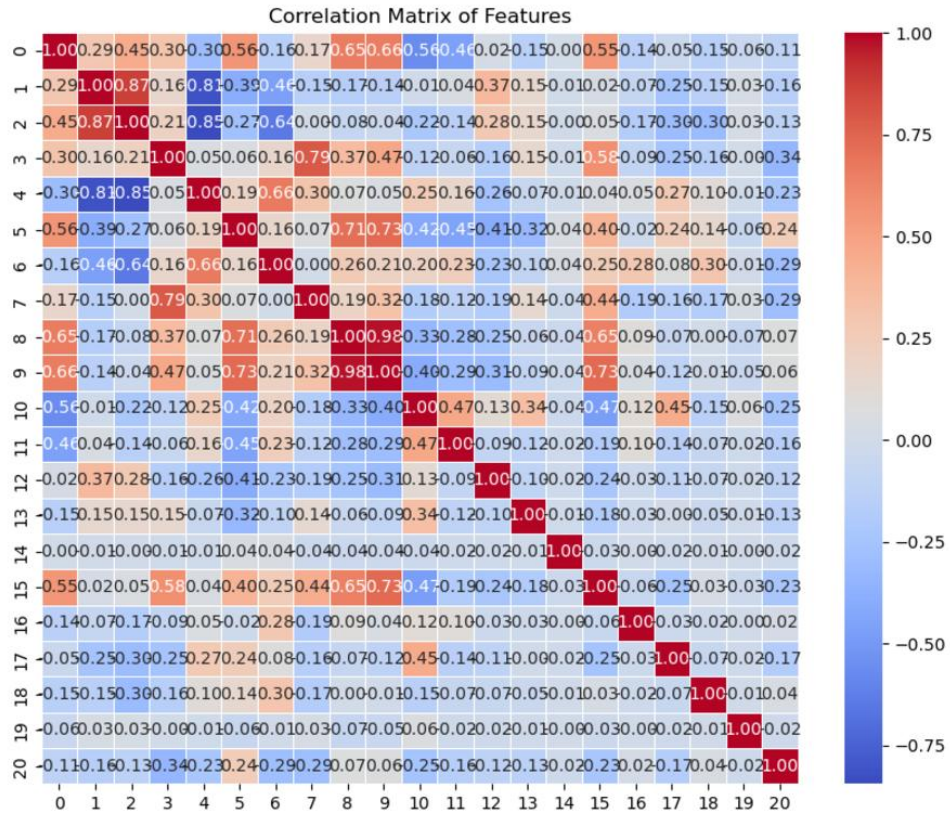



**图 4.5-16 训练集皮尔逊相关系数热力图**

相关性热力图显示，在经过预处理（包括独热编码、SMOTE 过采样及标准化）后的训练数据集中特征之间的相关系数大多数较低。只有少数特征对之间的相关性较高，因此本研究认为，基于这些特征构建朴素贝叶斯模型是可行的。

**（2） 预处理选取**

➢ 特征变量编码：性别（Gender）和职业（Occupation）被转换为独热编码，BMI 类别（BMI Category）使用标签编码。

➢ 数据标准化：在高斯朴素贝叶斯模型中，特征的标准化尤其重要，因为该模型假设特征符合正态分布，标准化能够更好地符合这一假设。

**（3） 超参数调整**

朴素贝叶斯算法的调优主要集中在平滑参数(var_smoothing)上。该参数用于防止在计算条件概率时出现零概率问题。默认值为 1e-9，经过调参后，我们确定最优的平滑参数为 1e-5，模型的整体准确率也得到大幅提升，由原来的 0.47 提升到为 0.78。、

此结果表明，平滑参数对模型性能的影响非常大。适当的平滑可以使模型更好地适应数据，避免过度拟合或欠拟合，从而提高预测准确性。

**（4） 模型评估**

经过调参后的朴素贝叶斯模型，以 1e-5 的平滑参数对朴素贝叶斯模型进行训练并预测训练集，得到混淆矩阵及 ROC-AUC 曲线：

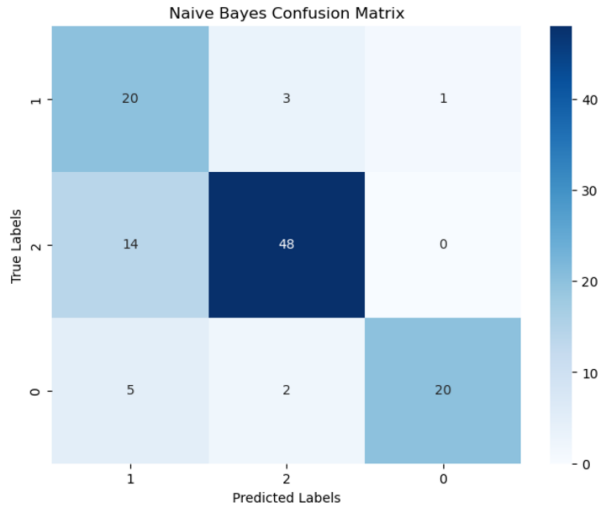



图 4.5-17 朴素贝叶斯混淆矩阵

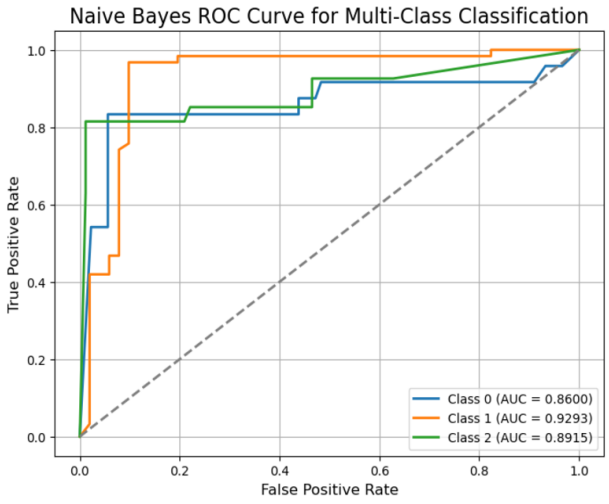



图 4.5-18 朴素贝叶斯 ROC-AUC 曲线

由混淆矩阵得出该模型准确率为 0.78，可以看到朴素贝叶斯模型在本研究中表现较为一般，相较于其他四类模型准确率较低。其中，较多的睡眠呼吸暂停类别被误判为无睡眠障碍类别，说明睡眠呼吸暂停和无睡眠障碍在一些特征上可能具有较高的相似性，

同时凸显出类别不平衡问题。该模型预测三种类别（失眠、无障碍、睡眠呼吸暂停）的 AUC 值分别约为 0.86、0.93、0.89，表明模型在测试集上仍具有较高的分类性能。

## 4.6 模型对比

在本研究中，采用了支持向量机（SVC）、多层感知机（MLP）、决策树（Decision Tree）、朴素贝叶斯（Naive Bayes）和 K 近邻（KNN）五种分类模型来进行睡眠障碍预测。以下是各个模型的性能比较。

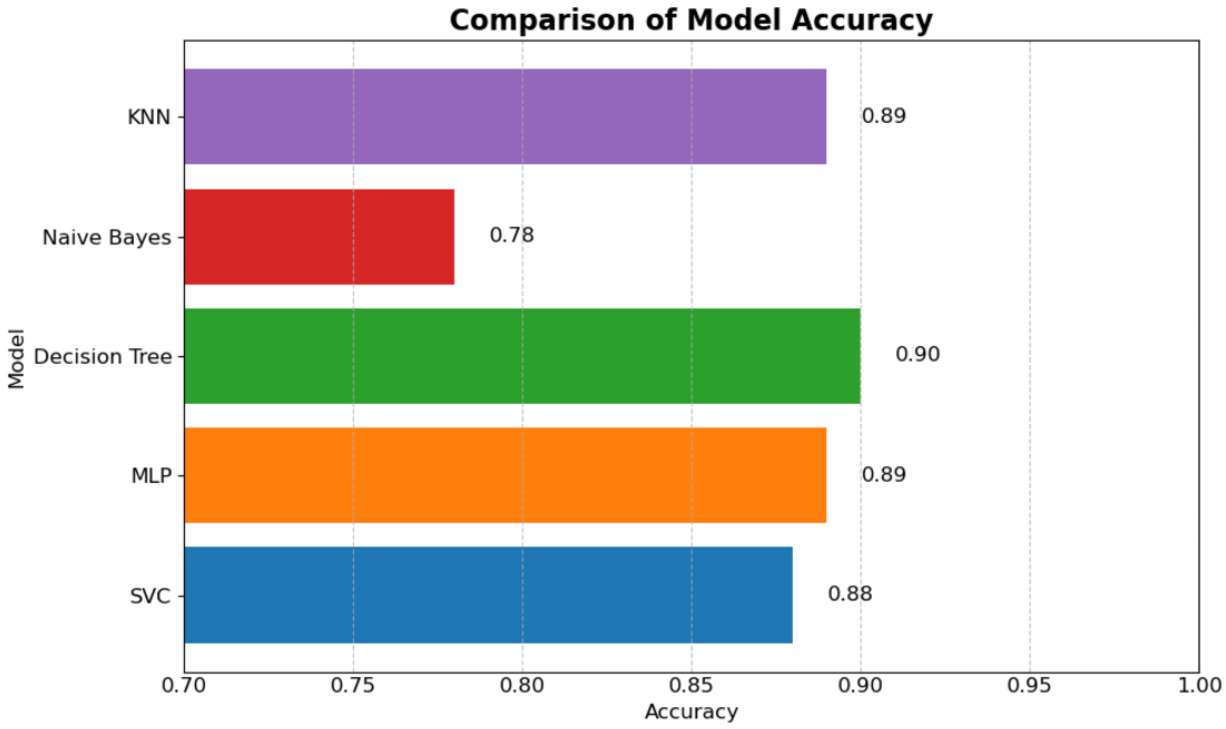



**图 4.6-1 模型准确率对比图**

从图表中可以看到，决策树模型在准确率上表现最佳，达到了 90%，优于其他模型，尤其在分类较为明确的类别上表现突出。SVC 和 MLP 的准确率分别为 88%和 89%，表现相近，而 KNN 则以 89%的准确率与前两者接近。朴素贝叶斯由于假设特征独立，无法处理特征间的相关性问题，准确率较低，为 78%。

综合比较，除朴素贝叶斯模型外其余四种模型的准确率差异不大，但决策树仍然凭借其较高的准确率和较低的过拟合风险占据了优势。考虑到本研究选取数据集较小，若数据集有所扩增，则 MLP、SVC 可能表现出更明显的优势，更适用于处理更复杂的决策边界问题。

# 结论

本研究通过利用 Kaggle 平台提供的睡眠健康和生活方式数据集，基于五种常见的机器学习算法（支持向量机 SVC、多层感知机 MLP、决策树、朴素贝叶斯和 K 近邻 KNN），构建并评估了睡眠障碍分类预测模型。通过数据预处理，包括缺失值处理、异常值检测、特征编码、SMOTE 过采样和标准化等，确保了数据的质量并有效地解决了类别不平衡问题。在此基础上，采用了超参数调优方法来优化各个模型的性能，并通过准确率、宏平均、加权平均等指标进行综合评估。

实验结果表明，五种模型中，朴素贝叶斯模型的表现较差，准确率为 78%，主要由于其假设特征之间独立，未能充分捕捉数据中的特征相关性。其他四种模型的准确率均在 90%左右，其中决策树模型表现最佳，准确率高达 90%，并且在宏平均和加权平均评估下，精确率、召回率和 F1 分数均表现优异。决策树模型不仅提供了较高的预测准确率，还能够有效消除数据不平衡的影响，且在计算效率上具有优势，最适合本研究任务。

综合考虑各模型的准确性和计算效率，决策树模型是最适合用于睡眠障碍分类预测的模型。基于决策树的模型具有较强的可解释性，能够帮助理解睡眠障碍的影响因素，尤其是 BMI 类别对睡眠障碍的影响最大。尽管朴素贝叶斯模型在本研究中表现较差，但在其他数据集上，特别是特征独立的情况下，仍可能具有应用价值。

在未来的研究中，可以尝试引入更多的特征变量，进一步优化模型，尤其是结合深度学习等先进方法，以提升模型的分类性能。同时，数据集的扩展和更新也是未来研究的关键，能够帮助更全面地评估和改进睡眠障碍预测模型。

# 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] 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.

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

[4] 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.

[5] 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.

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 袁帆,丁彩翠,宫伟彦,等.我国职业人群睡眠状况及其影响因素分析[J].中国公共卫生,2018,34(6):791-794.

[14] 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.

[15] 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.

[16] 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.

[17] 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).

[18] 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.

[19] 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.

[20] 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).

[21] 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.

附录

展示部分代码——SVC 及 MLP 模型

```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.preprocessing import StandardScaler, LabelEncoder,
label_binarize
from sklearn.svm import SVC
from sklearn.neural_network import MLPClassifier
from sklearn.metrics import classification_report, confusion_matrix,
roc_curve, auc
from imblearn.over_sampling import SMOTE
import matplotlib.pyplot as plt

# 读取数据
df = pd.read_csv('./Sleep_health.csv')
# 数据预处理

# 1. 目标变量预处理
df['Sleep Disorder'] = df['Sleep Disorder'].fillna('None')
df['BMI Category'] = df['BMI Category'].replace('Normal Weight',
'Normal')

# 2. 处理血压变量
df[['Systolic_BP', 'Diastolic_BP']] = df['Blood
Pressure'].str.split('/', expand=True)
df['Systolic_BP'] = pd.to_numeric(df['Systolic_BP'])
df['Diastolic_BP'] = pd.to_numeric(df['Diastolic_BP'])
df.drop(columns=['Blood Pressure'], inplace=True)

# 3. 类别变量独热编码（Gender, Occupation）和标签编码（BMI Category）
df = pd.get_dummies(df, columns=['Gender', 'Occupation'],
drop_first=True)
label_encoder = LabelEncoder()
df['BMI Category'] = label_encoder.fit_transform(df['BMI Category'])
# 4. 分割数据集
X = df.drop(columns=['Person ID', 'Sleep Disorder'])
y = df['Sleep Disorder']

# 将目标变量转为数值类型（0: None, 1: Insomnia, 2: Sleep Apnea）
```

```
from sklearn.preprocessing import LabelEncoder
# 创建标签编码器
label_encoder = LabelEncoder()
# 手动设置标签的顺序
target_mapping = {'None': 1, 'Insomnia': 0, 'Sleep Apnea': 2}

# 将目标变量按手动映射进行编码
y = y.map(target_mapping)

print(y.value_counts())  # 查看编码后的类别分布
# 数据划分：训练集和测试集（70% 训练集，30% 测试集）
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.3, random_state=42)

print(pd.Series(y_train).value_counts())  # 查看训练集中每个类别的分布
print(pd.Series(y_test).value_counts())   # 查看测试集中每个类别的分布

# SMOTE 过采样（训练集过采样）
smote = SMOTE(random_state=42)
X_train_resampled, y_train_resampled = smote.fit_resample(X_train,
y_train)

# 标准化（特征标准化）
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train_resampled)
X_test_scaled = scaler.transform(X_test)
print(X_test_scaled)
# 训练模型

# 1. SVC 模型（启用概率输出）
svc_model = SVC(kernel='rbf', C=1, gamma='scale', probability=True,
random_state=42)

# 超参数调优（使用 GridSearchCV）
param_grid_svc = {'C': [0.1, 1, 10], 'gamma': ['scale', 'auto']}
"""
            需要调优的超参数
            C：正则化参数，控制模型复杂度
                较小的 C 值（0.1）表示强正则化
                较大的 C 值（10）表示弱正则化
            kernel：核函数类型
```

```
                'linear'：线性核
                'rbf'：径向基函数核
            gamma：RBF 核的系数
                'scale'：1 / (n_features * X.var())
                'auto'：1 / n_features
                0.1：固定值
"""
svc_grid_search = GridSearchCV(svc_model, param_grid_svc, cv=5)
svc_grid_search.fit(X_train_scaled, y_train_resampled)

# 最优 SVC 模型
svc_best_model = svc_grid_search.best_estimator_
# 输出选取的最佳超参数（C 和 gamma）
print(f"Best C value: {svc_grid_search.best_params_['C']}")
print(f"Best gamma value: {svc_grid_search.best_params_['gamma']}")

# 使用训练集数据生成学习曲线
from sklearn.model_selection import learning_curve
train_sizes, train_scores, val_scores =
learning_curve(svc_best_model,
                                                X_train_resampled,
y_train_resampled,
                                                cv=5, n_jobs=1,
                                                train_sizes=np.linsp
ace(0.1, 1.0, 10))
# 计算训练集和验证集的平均得分
train_scores_mean = np.mean(train_scores, axis=1)
val_scores_mean = np.mean(val_scores, axis=1)
# 输出所有的参数组合及其对应的交叉验证得分
#print(svc_grid_search.cv_results_)
# 输出所有参数组合和对应的得分
results = pd.DataFrame(svc_grid_search.cv_results_['params'])
results['mean_test_score'] =
svc_grid_search.cv_results_['mean_test_score']
print(results)

# # 画出学习曲线
# plt.figure(figsize=(6, 6))
# plt.plot(train_sizes, train_scores_mean, label="Training score",
color='blue')
```

```
# plt.plot(train_sizes, val_scores_mean, label="Cross-validation
score", color='red')

# plt.title('Learning Curves (SVC)', fontsize=14)
# plt.xlabel('Training Size', fontsize=12)
# plt.ylabel('Accuracy', fontsize=12)
# plt.legend(loc='best')
# plt.grid(True)
# plt.show()

# 2. MLP 模型
mlp_model = MLPClassifier(random_state=42)
param_grid_mlp = {'hidden_layer_sizes': [(50,), (100,), (100, 50)],
'activation': ['relu', 'tanh'], 'solver': ['adam', 'sgd']}
"""
          神经网络的超参数
          hidden_layer_sizes：隐藏层的结构
            (50,)：一个隐藏层，50 个神经元
            (100,)：一个隐藏层，100 个神经元
            (50, 50)：两个隐藏层，每层 50 个神经元
          activation：激活函数
              'relu'：修正线性单元
              'tanh'：双曲正切函数
          learning_rate_init：初始学习率
               0.001：较小的学习率，训练更稳定但较慢
               0.01：较大的学习率，训练更快但可能不稳定
 """
mlp_grid_search = GridSearchCV(mlp_model, param_grid_mlp, cv=5)
mlp_grid_search.fit(X_train_scaled, y_train_resampled)

# 最优 MLP 模型
mlp_best_model = mlp_grid_search.best_estimator_
# 训练模型
mlp_best_model.fit(X_train_scaled, y_train_resampled)
#绘制损失函数图像
loss_values = mlp_best_model.loss_curve_  # 获取损失函数的变化（每个迭代
的损失）
plt.figure(figsize=(8, 6))
plt.plot(range(len(loss_values)), loss_values, color='b',
label='Training Loss')
plt.title('Loss Curve for MLP Model')
```

```
plt.xlabel('Epochs')
plt.ylabel('Loss')
plt.legend(loc='upper right')
plt.grid(True)
plt.show()

# 获取 GridSearchCV 的交叉验证结果
results = mlp_grid_search.cv_results_
# 将结果转换为 DataFrame 方便查看
results_df = pd.DataFrame(results)
# 显示每种超参数组合的表现
print(results_df[['param_hidden_layer_sizes', 'param_activation',
'param_solver', 'mean_test_score']])
# 模型评估（训练集和测试集）
import seaborn as sns
# 1. 对训练集和测试集的预测
y_pred_svc = svc_best_model.predict(X_test_scaled)
y_pred_mlp = mlp_best_model.predict(X_test_scaled)

# 2. 分类报告（Precision, Recall, F1-Score）
print("SVC Classification Report:")
print(classification_report(y_test, y_pred_svc))

print("MLP Classification Report:")
print(classification_report(y_test, y_pred_mlp))

# 3. 混淆矩阵
print("SVC Confusion Matrix:")
print(confusion_matrix(y_test, y_pred_svc))

# SVC 混淆矩阵
cm = confusion_matrix(y_test, y_pred_svc)

plt.figure(figsize=(6, 5))
sns.heatmap(cm, annot=True, fmt="d", cmap="Blues",
xticklabels=['Insomnia', 'None', 'Sleep Apnea'],
yticklabels=['Insomnia', 'None', 'Sleep Apnea'])
plt.title('Confusion Matrix')
plt.xlabel('Predicted Label')
plt.ylabel('True Label')
plt.show()
```

```
# MLP 混淆矩阵
cm = confusion_matrix(y_test, y_pred_mlp)

plt.figure(figsize=(6, 5))
sns.heatmap(cm, annot=True, fmt="d", cmap="Blues",
xticklabels=['Insomnia', 'None', 'Sleep Apnea'],
yticklabels=['Insomnia', 'None', 'Sleep Apnea'])
plt.title('Confusion Matrix')
plt.xlabel('Predicted Label')
plt.ylabel('True Label')
plt.show()
```