人工智能与数据挖掘课程设计

基于机器学习方法的医疗花费分析预测

**组员信息：**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 编号 | 姓名 | 学号 | 成员贡献 | 实践成绩 |
| 1 |  |  |  |  |
| 2 |  |  |  |  |
| 3 |  |  |  |  |
| 4 |  |  |  |  |

# **引言**

医疗保险费用的预测在健康经济学和医疗管理领域具有重要意义。了解个体的医疗保险费用是医疗政策制定、保险定价以及资源分配的关键因素。通过机器学习模型预测医疗保险费用，我们可以更准确地估计患者的成本，为个性化医疗和医疗资源规划提供支持。通过预测医疗保险费用，我们能够更准确地估算个体的医疗成本。这对于医疗机构、保险公司和政府部门来说，都是制定合理政策和费用结构的基础。也有助于制定个性化的医疗计划。了解哪些因素对费用产生影响，可以帮助医疗专业人员为患者提供更有效的医疗服务，从而提高治疗效果。此外，保险公司可以更精确地评估保单的风险，为客户提供更合理的保险费率。且有助于医疗机构优化资源分配。通过提前了解患者可能的医疗需求和费用，医疗机构可以更好地安排人力、设备和其他资源。

# **实验数据**

## 数据获取或预处理

实验数据来字Kaggle，数据下载链接为<https://www.kaggle.com/datasets/mirichoi0218/insurance>。

**2.1.1 缺失值处理**

图1展示了缺失值的缺失总体情况，missingno 是一个用于可视化缺失值的 Python 库。missingno.matrix 函数生成的矩阵图可以帮助我们直观地了解数据中缺失值的分布情况。图中的白色即为缺失值，从图中可以看出，本实验所用数据在字段bmi和children上均有缺失，显示为白色线段，其余黑色代表了未缺失。

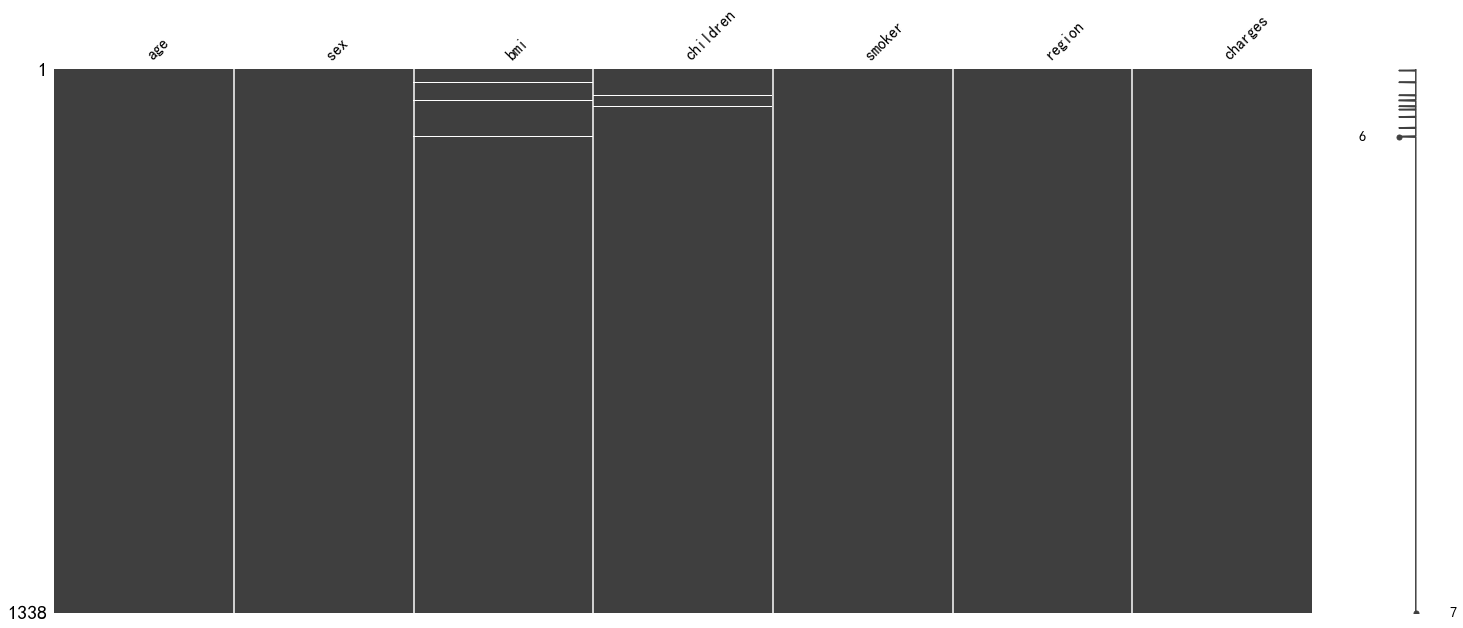


图 1 缺失值缺失情况

**2.2.2 缺失值填充**

bmi字段的数据为连续性数值，所以采用均值填充。，采用均值填充有助于保持数据集的整体分布特性。对于像BMI这样的连续性数值字段，使用均值替代缺失值可以有效避免引入过多的偏差，确保填充后的数据仍然反映了原始样本的分布特征，有助于维持数据的一致性和连续性。在实践中，这种方法无需复杂的计算或参数调整，特别适用于大规模数据集或需要高效处理的场景。这种直观而直接的填充方式使数据预处理变得更为迅速和方便。

另一方面，均值填充表现稳定且具有一定的鲁棒性。对于不同类型的数据，它都能够提供相对可靠的填充结果，并对异常值的影响相对较小。这种稳定性使得均值填充在面对多样化的数据集时都能够表现良好，而不容易受到个别异常值的干扰。此外，均值填充有助于避免信息损失。在一些基本的数据处理任务中，保留尽可能多的原始信息对于后续分析和建模至关重要。均值填充提供了一种简单而有效的手段，可以填补缺失值而不引入过多的外部信息，从而最大限度地保留了原始数据的特征。

children字段均为整型数据，因此采用均填充方法不合适，选用众数填充，适用于保持数据的整数属性，处理离散分布，且简单易行，众数填充能够有效填补缺失值而不引入连续性特性。通过选择数据中最频繁出现的整数值，众数填充保持了数据的整体分布特性，同时对异常值表现相对鲁棒，使其成为处理整数型数据缺失的合理而直观的方法。

**2.2.3类别编码**

对于字段sex，smoker，region，原始数据中为文本数据，计算机无法识别文本数据，因此选用类别编码，将原始数据转为数值数据。通过将原始文本数据转换为数值形式，类别编码使计算机能够有效处理这些信息，为机器学习算法和统计模型提供了可用的输入。这种转换不仅使数据变得更加计算机友好，还确保了文本信息的有效使用，为数据分析和建模提供了便利。

**2.2.4 数据归一化**

数据标准化是数据预处理的一项重要步骤，旨在将不同特征的数值范围调整到相同的尺度，以消除由于不同尺度而引起的偏差，从而提高模型的性能和稳定性。在数据标准化过程中，通常采用的方法是通过减去均值并除以标准差，将数据转换为均值为零、标准差为一的标准正态分布。这有助于确保不同特征的值具有相似的数值范围，使得模型在训练过程中更容易收敛，减少了由于特征间尺度不一致而导致的权重更新不平衡的情况。

# **数据统计**



图 2 性别数据统计

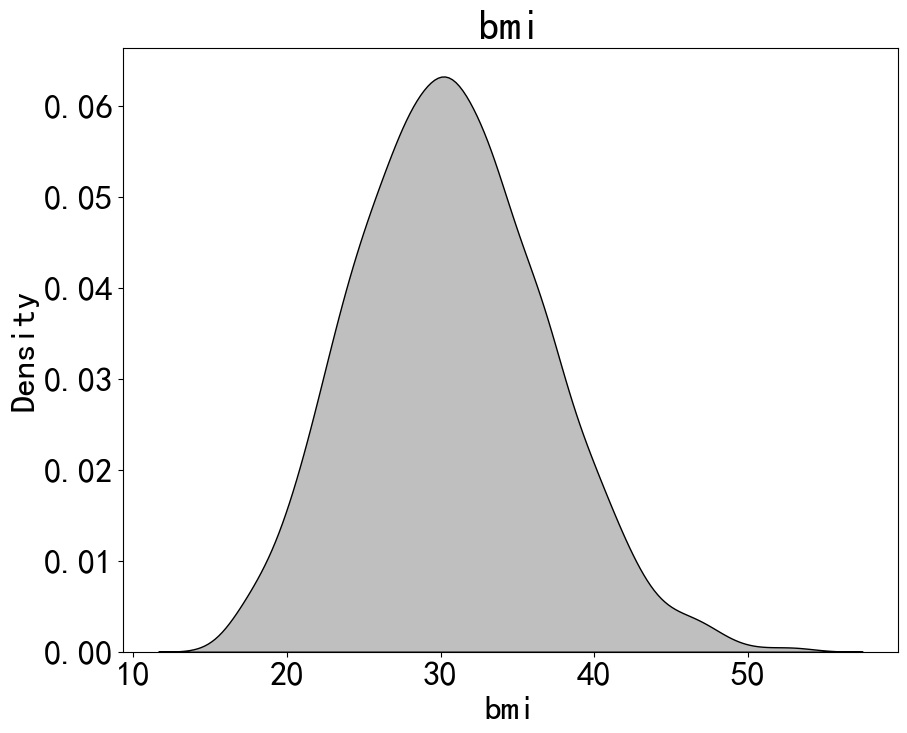


图 3 bmi数据统计

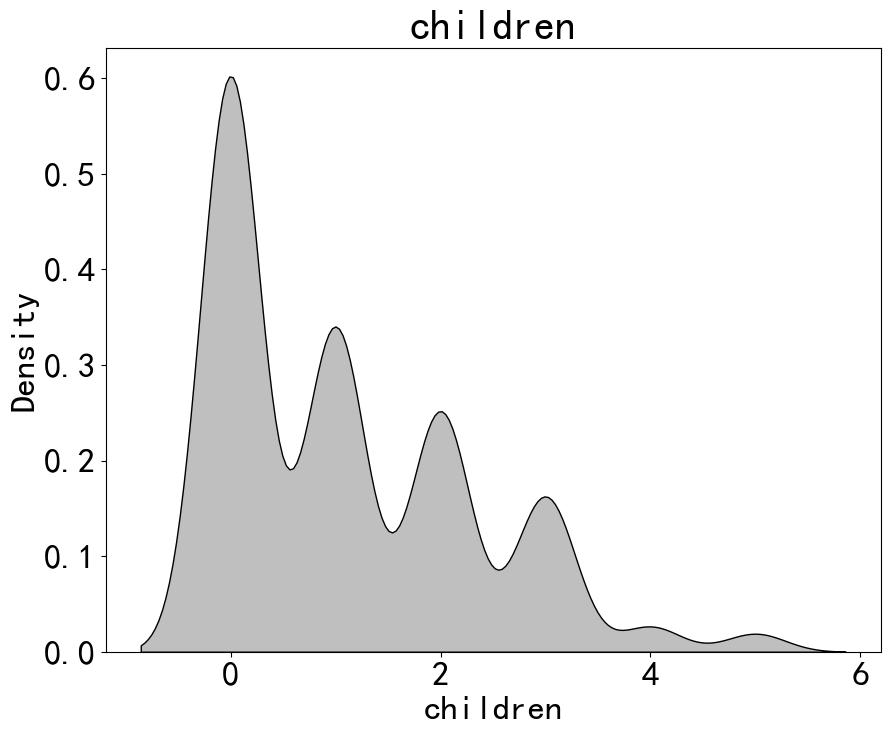


图 4 children数据统计

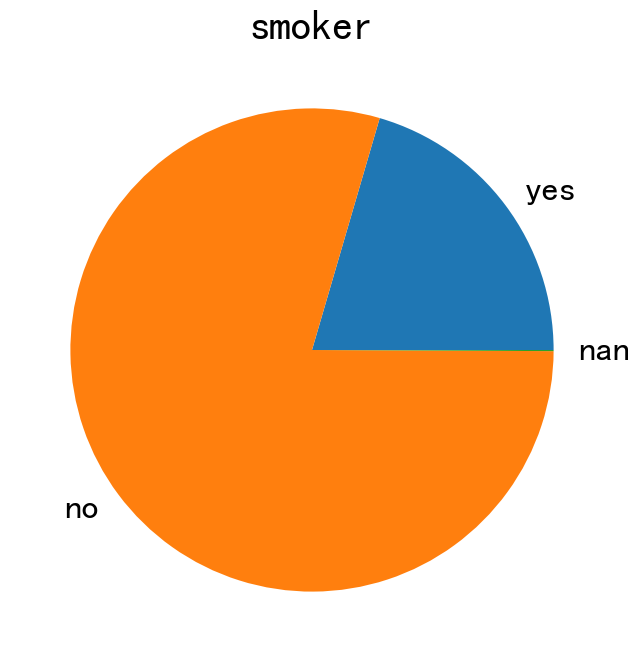


图 5 smoker数据统计

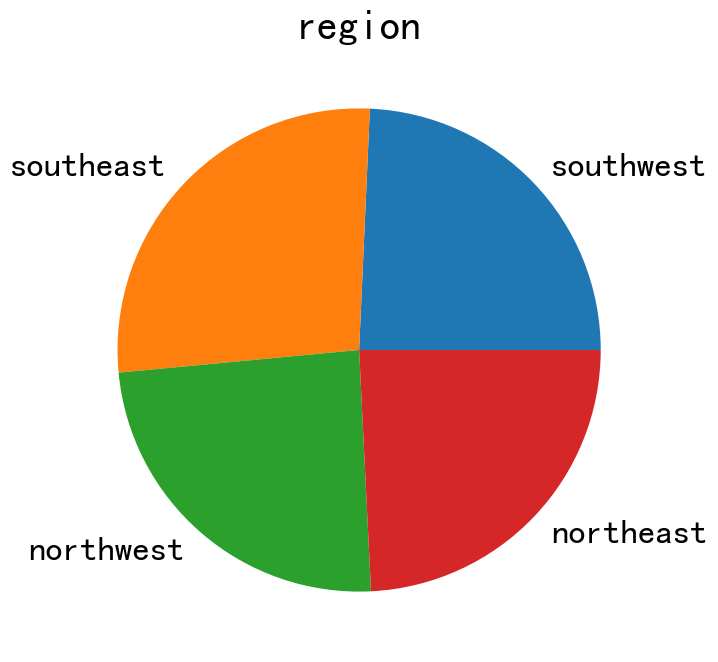


图 6 region数据统计

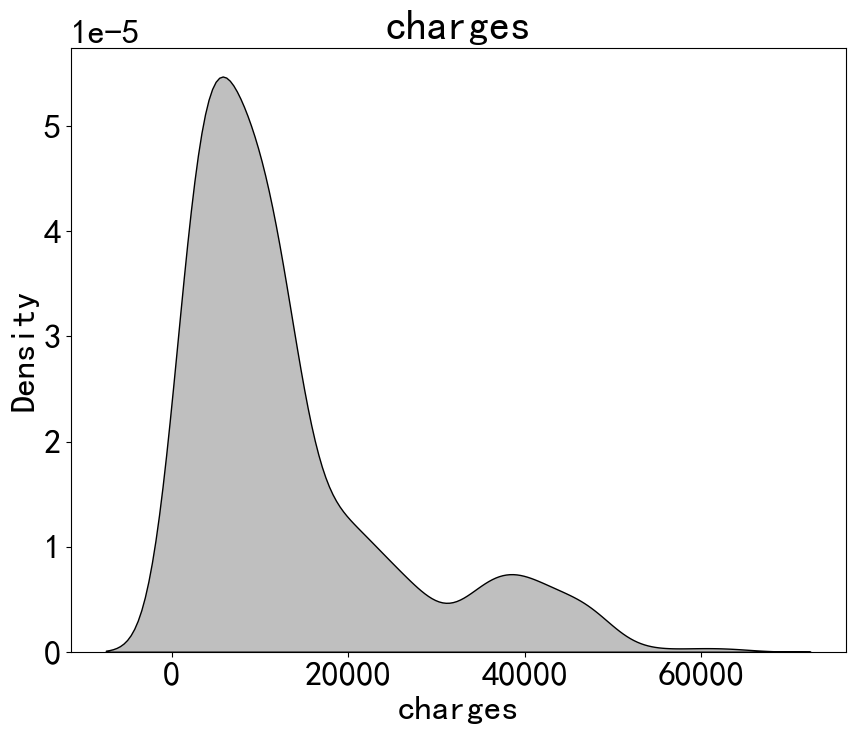


图 7 charges数据统计

# **方法/模型**

* 1. 决策树

决策树分类算法的原理基于对数据集进行逐步划分，通过对特征进行选择，使得每一次划分都能够最大程度地提高数据的纯度（同一类别的数据尽量聚在一起）。以下是决策树分类算法的基本原理：

1. **选择最佳特征：** 算法首先评估每个特征的重要性，选择能够最好地分割数据的特征。这个选择通常基于某个准则，如信息增益、基尼不纯度等。信息增益衡量了使用某个特征对数据进行划分后不确定性的减少程度。
2. **划分数据集：** 选定了最佳特征后，将数据集根据该特征的取值进行划分，形成子集。每个子集对应于特征取值的一个分支。
3. **递归建树：** 对每个子集递归地重复上述过程，选择每个子集中的最佳特征，继续划分。这个递归过程会一直持续，直到满足停止条件，例如树的深度达到预定值、节点包含的样本数小于某个阈值，或者数据已经完全分割为纯净的类别。
4. **叶子节点：** 当某个节点不能再划分时，它成为叶子节点，并表示最终的分类结果。叶子节点中的类别通常由该节点中样本的多数类别决定。

通过这样的递归过程，决策树将数据集划分为一系列的决策规则，形成一个树状结构，最终用于分类新的未见样本。决策树算法具有直观、易解释、可视化等优点，适用于许多分类问题。

* 1. 随机森林

随机森林是一种基于集成学习思想的算法，通过构建多个决策树并结合它们的预测结果来提高模型的性能和鲁棒性。以下是随机森林的基本原理：

1. **数据随机抽样：** 随机森林通过对原始数据集进行有放回的随机抽样（bootstrap抽样），创建多个不同的训练数据子集，每个子集包含原始数据集中的一部分样本。这样，对于每个子集，我们都能构建一个独立的决策树。
2. **特征随机选择：** 在构建每个决策树的过程中，随机森林对特征进行随机选择。这意味着在每个节点上，不是使用所有特征进行划分，而是从所有特征中随机选择一部分特征。这有助于确保每个决策树都是不同的，提高了模型的多样性。
3. **建立决策树：** 对于每个数据子集和随机选择的特征集，构建一个独立的决策树。这些决策树构成了随机森林的集成。
4. **投票或平均：** 在进行预测时，随机森林通过对所有决策树的输出进行投票（分类问题）或平均（回归问题），得到最终的集成预测结果。这样的集成能够降低过拟合风险，提高模型的泛化能力。

随机森林通过引入随机性，避免了单一决策树容易过拟合的问题，同时通过集成多个模型的意见，提高了整体模型的稳定性和准确性。这使得随机森林在各种分类和回归问题中都表现出色，尤其适用于处理高维数据和大规模数据集。

# **结果评测/展示**

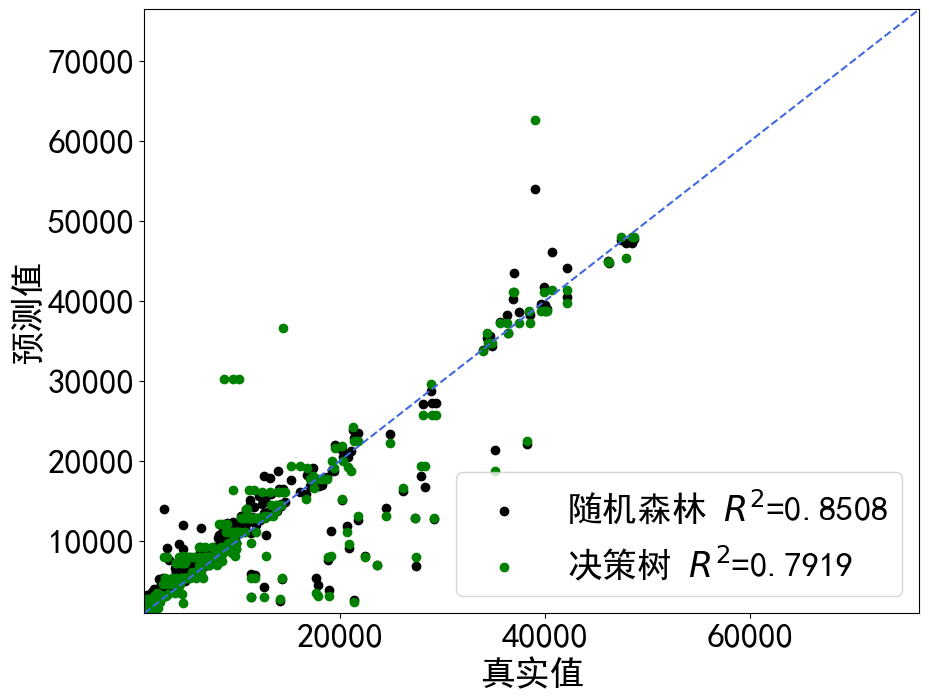


图 8 不同算法预测结果对比

R²（R-squared）是一个用于度量回归模型拟合程度的统计指标，它的取值范围在0到1之间，表示模型对数据方差的解释程度。R²越接近1，说明模型对数据的拟合越好。

在你提供的情况下，随机森林的R²为0.8508，而单一决策树的R²为0.7919。这意味着随机森林的拟合效果相对更好，更能解释目标变量的方差。而决策树的R²相对较低，说明单一决策树在拟合数据时可能存在一些局部过拟合或不足以解释数据方差的问题。

有几个原因可能解释这一差异：

1. **随机森林的集成效果：** 随机森林通过集成多个决策树，每个决策树都是在不同的数据子集和特征子集上训练的，这有助于提高模型的泛化能力，降低过拟合风险。因此，随机森林通常能够更好地适应数据，提高模型的性能。
2. **决策树的过拟合：** 单一决策树容易在训练数据上过度拟合，特别是当决策树深度较大时。这可能导致模型在训练数据上表现很好，但在未见过的数据上泛化能力较差。
3. **决策树的离散性：** 决策树是一种高方差（high variance）的模型，对输入数据的小变化可能产生较大的变化。这导致了决策树的结果较为离散，而随机森林通过多个决策树的集成能够减少这种离散性，提高整体模型的稳定性。

总体而言，R²的差异可能是由于随机森林对数据的更好拟合、更好的泛化能力以及对决策树过拟合问题的缓解所致。

图9为特征重要性分析，这个特征重要性分析是对机器学习模型中各个特征在预测医疗保险费用方面的相对贡献进行的评估。特征重要性告诉我们在模型的决策中，哪些特征对于预测目标变量的变化起到了更为关键的作用。在给定的字典中，每个特征都被赋予了一个相对权重，表示其对模型输出的影响程度。

具体而言，以下是各特征的相对重要性：

1. **'sex'（性别）: 0.0022：** 性别在预测医疗保险费用中的相对贡献较小，说明性别对于费用的变化影响较弱。
2. **'region'（地区）: 0.0088：** 地区对费用的相对贡献略有增加，但仍然相对较小。
3. **'children'（子女数量）: 0.0144：** 子女数量对费用的相对贡献有所增加，说明有一定程度的影响。
4. **'age'（年龄）: 0.1206：** 年龄是相对重要的特征，对于医疗保险费用的预测有显著的影响。
5. **'bmi'（体重指数）: 0.1949：** 体重指数在费用预测中具有较高的相对重要性，表明个体的体重状况对费用有显著影响。
6. **'smoker'（是否吸烟）: 0.6592：** 吸烟与费用的相对贡献最高，说明吸烟状态是影响医疗保险费用的主要因素。

总体而言，这个特征重要性分析强调了模型认为各个特征在医疗保险费用预测中的相对重要性。吸烟状态、体重指数和年龄等因素被认为对费用有更大的影响，而性别和地区等因素的影响相对较小。这对于理解模型的决策过程，以及指导进一步的数据收集和特征工程具有指导性意义。

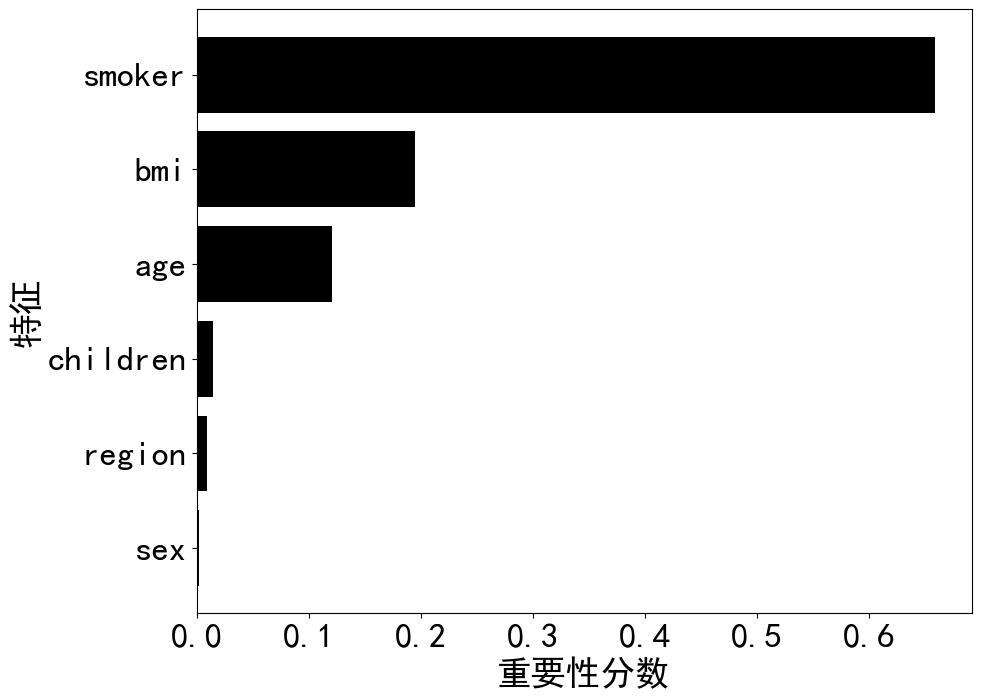


图 9 重要性分析

# **总结**

本次工作基于机器学习方法，通过对医疗保险数据集的分析和建模，旨在预测医疗保险费用。以下是工作的主要总结：

1. **数据预处理：** 进行了详细的数据预处理，包括处理缺失值、选择合适的填充方法（均值填充和众数填充）、类别编码以及数据归一化。这些步骤有助于确保数据质量和适应机器学习模型的需求。
2. **特征工程：** 对原始数据进行了特征工程，将文本数据转换为数值形式，以便模型能够处理。特征重要性分析显示，年龄、体重指数和吸烟状态等特征对医疗保险费用的影响较大。
3. **模型选择和比较：** 使用了决策树和随机森林两种不同的机器学习模型进行医疗保险费用的预测。通过比较模型的性能指标（R²值），发现随机森林相对于单一决策树表现更好，具有更好的泛化能力。
4. **特征重要性分析：** 通过特征重要性分析，识别了对医疗保险费用影响较大的特征，为进一步的解释和应用提供了有价值的信息。

**局限性分析：**

1. **数据局限性：** 结果受数据质量和覆盖范围的限制。如果数据集不够全面或者包含较多噪声，模型的泛化能力可能会受到影响。
2. **特征选取：** 虽然特征重要性分析提供了对模型的洞察，但仍需谨慎对待特征的选取。某些可能重要但未被包括的特征也可能对模型的性能产生影响。
3. **模型解释性：** 随机森林等复杂模型的解释性相对较差，可能难以直观理解其预测过程。在某些场景下，需要更加可解释的模型。
4. **实际应用：** 模型预测的结果需要在实际医疗保险领域中得到验证，以确认模型是否能够真正帮助决策制定和资源分配。

综合来看，本次工作为医疗保险费用预测提供了一种机器学习的方法，并通过数据预处理、特征工程和模型选择等步骤进行了细致的分析。然而，模型的应用仍需谨慎，并需要进一步实证研究来验证模型的实际效果。