基于机器学习的患者健康风险分析与个性化医疗推荐

岑　通
广州南方学院　广州从化　440181

摘　要：本论文聚焦于医疗数据展开相关研究，希望能够探寻机器学习于医疗领域所有的应用价值。不过该研究依旧存在一些局限性，比如数据质量以及完整性有待提升，以及模型的适应性以及优化空间等问题，未来的研究需要收集数据、优化模型，推动精准医疗的发展，为医疗决策提供更为有力的支持。

关键词：机器学习医疗；健康风险分析；个性化医疗推荐

1 基于机器学习的患者健康风险分析模型构建

1.1 特征提取与选择

数值型特征涉及了年龄、体重、身高以及检查结果等连续变量，这些特征可直接体现出患者的生理状况以及健康指标。类别型特征涉及了诸多方面，有性别、地区、疾病类型等等，这些特征借助标签编码也就是 LabelEncoding 转换成为数值形式，以此方便模型进行处理。在处理缺失值时，针对数值型特征，采用均值来填充其中的缺失值，而对于类别型特征，则运用众数去填充缺失值。

1.2 模型选择

本研究选用逻辑回归、随机森林和神经网络三种模型进行患者健康风险分析，各模型具有独特优势与理论依据。

1.3 模型训练与验证

对于逻辑回归模型，设置最大迭代次数为 1000，随机种子为 80，利用训练集进行拟合训练，并在测试集上进行预测。随机森林模型同样设置随机种子为 80，使用训练集进行训练，然后对测试集进行预测。神经网络模型将训练集和测试集转换为张量形式，定义了包含输入层、两个隐藏层和输出层的神经网络结构，采用交叉熵损失函数和 Adam 优化器进行训练，训练 100 个 Epochs，在训练过程中不断调整模型参数，最后在测试集上进行预测。

1.4 模型评估与优化

1.4.1 模型评估

采用准确率（Accuracy）、精确率（Precision）、召回率（Recall）和 F1 分数（F1-score）等指标对三种模型的性能进行评估。具体总结如下：

准确率方面：随机森林模型的准确率最高，为 0.6290；逻辑回归模型准确率为 0.5716；神经网络模型准确率是 0.5762。随机森林模型在整体预测正确的比例上表现相对更优。

精确率方面：随机森林模型精确率为 0.6571，高于逻辑回归模型的 0.5930 和神经网络模型的 0.6042，说明其预测为正例中实际正例的比例相对较高。

召回率方面：随机森林模型召回率为 0.6000，逻辑回归模型为 0.5638，神经网络模型为 0.5377，随机森林模型在实际正例被正确预测的比例上领先。

F1 分数方面：随机森林模型的 F1 分数为 0.6273，逻辑回归模型是 0.5789，神经网络模型为 0.5690。F1 分数综合了精确率和召回率，随机森林模型在这一综合评估指标上同样表现最佳。

总体来看，在这四种评估指标下，随机森林模型的性能在三个模型中相对更优。同时，使用这些模型进行患者健康风险预测时，随机森林模型在准确率指标上显著高于其他两个模型，在综合评价指标 F1 值上也表现较好；神经网络模型的召回率最低。

1.4.2 模型优化

为了进一步提高随机森林模型的性能，采用 Optuna 库对随机森林模型的超参数进行优化。定义了目标函数，在目标函数中设置超参数的搜索范围，包括决策树数量、决策树最大深度、内部节点再划分所需最小样本数、叶节点最少样本数和考虑的最大特征数等。

通过运行 100 次试验，寻找最优的超参数组合。经过优化后，得到最优的超参数。使用最优参数重新训练随机森林模型，并对其性能进行评估。优化后的随机森林模型，呈

现了最优随机森林模型的性能指标和分类报告，具体如下：

模型性能指标：准确性 0.6667，即模型预测的正确率约 66.67%。精确率 0.7123，预测正例中实际正例比例高。召回率 0.6009，模型识别出的正例占实际正例的 60.09%。F1 值 0.6531，平衡了精确率和召回率。分类报告类别 0：精确率 0.63，召回率 0.74，F1 值 0.68，样本数 636，识别时召回效果较好。类别 1：精确率 0.71，召回率 0.60，F1 值 0.65，样本数 690，精确率高但召回率低。整体准确性 0.67，是所有样本的正确预测比例，和模型性能指标的准确性稍有不同。宏平均：精确率、召回率、F1 值都是 0.67，是各类别指标的算术平均。加权平均：也是 0.67，按类别样本数量加权计算得出。

借助超参数优化，随机森林模型的性能有了一定程度的提高，结果说明优化过后的模型在准确率、精确率以及召回率等指标方面都有不同程度的提高，超参数调优对模型性能的改善有着积极的影响，对优化后模型的特征关键性展开分析，绘制出特征关键性柱状图，可直观地呈现出各个特征对模型预测结果的贡献程度，为理解模型决策过程以及特征选择提供参考。

2 个性化医疗推荐系统设计与实现

2.1 个性化医疗推荐系统架构

2.1.1 整体架构概述

个性化医疗推荐系统依据用户输入的疾病类型、身体指标比如身高体重以及遗传病史等信息，来给用户提供个性化医疗建议以及治疗方案，系统整体运用客户端 – 服务器架构，客户端主要承担用户界面交互工作，借助 Pyside6 联合 QTDesigner 开展设计，便于用户输入信息以及查看推荐结果，服务器端（在本示例里未涉及实际服务器端处理，主要于客户端本地开展计算和推荐）负责处理用户输入的数据，开展相关计算以及逻辑判断，最终生成个性化医疗推荐内容并展示给用户。

2.1.2 各模块功能模块详细设计（部分）

（1）数据处理模块

在 MainWindow 类的 recommend_treatment 方法中，通过 self.comboBox_disease_type.currentText() 获取用户在疾病类型下拉框中选择的疾病类型，self.lineEdit_height.text() 获取用户输入的身高，self.lineEdit_weight.text() 获取体重，self.comboBox_genetic.currentText() 获取遗传病史信息。对获取到的身高和体重数据进行简单的有效性检查，确保输入为数字格式，若为空或非数字则进行相应的提示处理。

（2）推荐引擎模块

依据用户所选择的遗传病史状况来生成相应建议，要是遗传病史显示为“无”，那么推荐的内容是“虽然当前无遗传病史，然而依旧需要留意家族疾病史，要定期去做体检”，要是遗传病史是“不清楚”，则给出的建议是“遗传病史的信息处于不清楚的状态，建议去了解家族疾病的具体情况，以此来优化治疗方案”，要是遗传病史为“有”，则给出的建议是“存在遗传病史，建议开展更为详细的基因检测以及家族疾病风险评估，制定出个性化的治疗方案”。

2.2 系统实现步骤与代码示例

2.2.1 系统集成与测试

（1）系统集成

将生成的 ui_medical_recommendation.py 文件和编写的 main.py 文件放在同一目录下，确保代码中导入模块的路径正确。运行 main.py 文件，即可启动个性化医疗推荐系统的客户端界面。

（2）功能测试

打开系统后，依次进行以下测试：

界面组件测试：检查各个标签、下拉框、文本框、按钮和文本浏览器是否正常显示，位置和大小是否符合设计要求，组件之间的布局是否合理，交互是否流畅（如点击下拉框是否能正常弹出选项，点击按钮是否有响应等）。

数据输入与处理测试：在疾病类型下拉框中选择不同的疾病，在身高和体重文本框中输入合法的数值（如身高 170cm，体重 60kg）以及不合法的数值（如空值、非数字字符等），观察系统是否能正确获取数据并进行相应的处理。

2.3 系统性能评估与优化

本医疗推荐系统旨在根据用户输入的疾病类型、身高、体重和遗传病史等信息，为用户提供个性化的医疗建议，包括疾病治疗方案、药物推荐、BMI 相关健康指导以及遗传病史应对策略。

然而，在实际测试中发现，当输入“高血压”这一疾病类型时，系统未能正确识别并给出相应的准确推荐，而是显示“未识别的疾病类型，请重新输入”。

这清晰地显示出系统于疾病类型识别的完整性方面存有不足之处，说不定致使用户难以获取针对该疾病的有效建

议，对系统的实用性以及可靠性均产生了影响。

3 实验与结果分析

3.1 个性化医疗推荐实验结果及可视化展示

在个性化医疗推荐方面，构建了一个包含疾病类型、身高、体重和遗传病史等信息的推荐系统。通过用户输入疾病类型、身高、体重和遗传病史等信息，系统根据预定义的疾病信息字典和遗传病史建议规则生成个性化的医疗推荐。

3.2 结果分析与讨论

3.2.1 关联分析

从对疾病相关因素的分析结果可看出，年龄、性别、疾病类型以及遗传病史等诸多因素，和疾病严重程度之间有着紧密的联系，比如说，一些疾病在特定的年龄段，其发病率以及严重程度都比较高，而且性别差异也有可能致使疾病的表现以及发展出现不同，在治疗这方面，不同的疾病类型与治疗方案以及药物使用之间存在着较为十分突出的对应关系，这就意味着在临床实践当中，需要依据患者具体的疾病状况来制定个性化的治疗方案，要充分考虑这些关联因素，以此来提升治疗效果。

在个性化医疗推荐领域，凭借身高以及体重所计算得出的 BMI，与疾病风险以及治疗建议之间同样存在着关联，不同的 BMI 范围，分别对应着不一样的健康风险以及建议措施，举例来说，处于低 BMI 范围时，有可能暗示着营养不良或者存在潜在的健康问题，此时就需要增加营养的摄入量并且进行适当的锻炼，而当 BMI 处于高范围时，则有可能使患某些疾病的风险有所增加，这种情况下就需要对饮食加以控制，同时还要加强运动等。这样的情况，为在综合考量患者身体状况的基础上开展医疗推荐提供了相应依据。

3.2.2 整体性能评估

对于患者健康风险分析，借助多种可视化手段以及评估指标对分析结果给予了全面评估，数据预处理保证了数据质量，疾病相关因素和治疗相关分析为深入了解疾病特征与治疗规律给予了丰富信息，在模型训练过程中，随机森林模型在优化前就呈现出相对不错的性能，经过超参数优化后，准确率和 F1 分数提升，该模型在疾病严重程度分类预测方面有一定的可靠性。

3.2.3 改进建议

为了让患者健康风险分析变得更加准确可靠，可去收集更多医疗数据，让数据来源更为丰富，样本量也得以增加，提高模型的泛化能力，在对数据进行预处理时，可以采用更为先进的异常值检测处理办法以及数据清洗技术，以此来保证数据质量，在模型方面，可以尝试把多种模型的优势结合起来，运用集成学习方法提升预测性能，就像构建融合逻辑回归、随机森林以及神经网络的模型那样。

参考文献：

[1] 张乐 . 智慧医疗领域的机器学习分析及应用 [J]. 数字技术与应用 ,2024,42(2):51–54.

[2] 胡芬 . 机器学习在医疗行业的应用 [J]. 大众标准化 ,2020,(07):61–62.

[3] 张炼文 . 云计算与机器学习技术在智慧医疗的应用策略 [J]. 科学与信息化 ,2024(7):154–156.

[4] 何远 . 基于机器学习的医疗数据分析与挖掘研究 [J]. 微型计算机 ,2024(2):46–48