医疗保险欺诈监测模型项目概要

# 1. 前言

## 1.1 背景

随着医疗保险体系中欺诈行为的日益增加，对医保基金的财务损失以及对社会公正和医疗资源合理分配的负面影响日益凸显。传统的欺诈监测方法，如手动审核和基于规则的系统，由于效率低下且无法有效识别复杂欺诈模式，已不再能够满足当前的需求。鉴于此，本项目的核心目标是开发一个基于人工智能的医疗保险欺诈识别模型，该模型采用先进的机器学习算法处理大规模医疗数据，旨在自动、准确地识别各类欺诈行为，同时重视检测的准确性和模型判断的可解释性。此外，模型的设计考虑到欺诈策略的不断演变，因此强调了灵活性和可扩展性。

## 1.2 项目目标

1. 构建医疗保险欺诈识别模型：开发基于先进机器学习技术的模型，以高效准确地识别医疗保险欺诈行为。
2. 适应性和迭代：确保模型能够快速适应医疗保险欺诈手段的变化，并进行有效迭代。
3. 算法组合和测试：结合多种算法（如LightGBM、XGBoost、随机森林）以实现最优的欺诈识别性能。
4. 处理类别不平衡问题：设计数据采样和预处理策略，确保少数类别（欺诈样本）在训练中有效学习。
5. 特征构建：构建特征，以提高算法识别欺诈模式的能力并减少所需特征数量。
6. 超参数调整：使用遗传算法，选择和优化模型的超参数，以提高其性能和迭代效率。
7. 可解释的机器学习技术：采用易于理解的机器学习方法，确保专业人员可以理解模型的决策过程。
8. 全面的性能评估：使用分类报告和ROC-AUC分数等多维度指标全面评估模型性能。

# 2. 目前实现的效果

目前，我们的医疗保险欺诈检测模型通过细致的特征工程从81个特征扩展至近500个，显著提升了泛化能力和稳定性并减少了所使用的特征数量。选择了随机森林和轻量级梯度提升机作为主要分类器，实现了快速迭代和高准确性。模型AUC分数达到0.930，展示出优异的性能，同时保持了可解释性，以便专业人员和监管机构能够理解模型的决策过程。

# 3. 特色综述

1. 使用基于**贪婪思想**的方法自动选择特征。将其与**数据分析和特征重要性相结合**来选择特征。与传统的单纯使用数据分析手段或者特征重要性等手段**效率更高**，效果更好，这种方法在选择特征时还能够兼顾特征之间的潜在联系。
2. 使用模型融合的方法结合多个模型。遵循好而不同的原则，增加模型的泛化性，提高模型在未知数据集上的表现。与传统的单纯使用单一模型或者单一集成模型（RF、GBDT）不同，本方法在此基础上**进一步结合了不同模型的各自优势**，使得模型的性能进一步得到提高。
3. 使用**遗传算法**进行超参数优化。超参数优化能够进一步提高模型的表现，与传统的使用网格搜索和贝叶斯优化不同，遗传算法不仅权衡了效率和最优结果，在处理具有多个局部最优解的复杂搜索空间方面表现出色，对超参数的选择和算法本身的设计参数不太敏感，具有很好的鲁棒性和适应性。考虑到医疗保险欺诈这种需要高性能的模型，我们认为遗传算法更加合适。除此之外，我们进一步改进了遗传算法，使其能够在大规模数据集上处理**更加高效**、**添加自适应函数**使其能够处理不同分类器的参数，使用精英策略确保每一代最优个体传到下一代。

# 4. 开发工具与技术

|  |  |
| --- | --- |
| 类别 | 工具与技术 |
| 编程语言 | Python |
| 开发工具 | Jupyter Notebook、Pycharm 2023.3.2 |
| 开发环境 | Windows 11 |
| 安装包及版本 | Environment.yml文件 |

# 5. 应用对象

任何需要减少因医疗保险欺诈造成损失的保险公司以及国家社会保险机构。

# 6. 应用环境

支持Windows系统与机器学习开发工具。Environment.yml文件中包含了各种库及其版本号。

# 7. 总结

本项目成功地开发了一个高效、准确的医疗保险欺诈检测模型，显著提高了对复杂欺诈模式的识别能力。通过精细的数据处理和先进的算法选择，我们不仅提高了模型的性能，还确保了其可解释性和适应性。未来，我们计划进一步优化模型并开发出实时检测的系统，对潜在风险进行有效管理，以持续提升项目的实际应用价值。