附件 4 **重庆医科大学本科生毕业论文（设计）中期检查情况表**

|  |  |  |  |
| --- | --- | --- | --- |
| **学生姓名** | 杜兴兴 | **学 号** | 2016223592 |
| **学院名称** | 公共卫生与管理学院 | **年级专业** | 2016级应用统计学 |
| **指导教师** | 曾庆 | **职 称** | 教授 |
| **课题名称** | 机器学习模型在阿片类药物使用量预测应用中的评估研究 | | |
| **课题的进展情况（论文作者填写）：**   1. 前言    1. 研究背景和意义       1. 研究背景   研究表明，全球超过1/5的人正在遭受疼痛的折磨，其中半数患者的疼痛程度达到中度或重度水平，65岁以上的老年人慢性疼痛的发生率更是超过50%。作为有效缓解中、重度疼痛的阿片类药物，虽然许多人需要阿片类药物来控制他们的慢性和严重的疼痛，但这些治疗的一个常见后果是滥用、成瘾和升级到更恶劣的药物。阿片类药物是一种麻醉性止痛药，包括非法毒品海洛因，他们从罂粟花中提取，或人工合成，结构与其他阿片类药物相似。一些阿片类处方药的例子是吗啡、氢考酮、双氢吗啡酮和芬太尼。虽然这类药物中几乎每一种都可以用来治疗慢性疼痛，但阿片类药物的使用已经远远超出了处方药的范围，而且已经成为一种流行病。  据统计，仅2013年全球就有2800万~3800万人非法使用阿片类药物（占15~65岁之间全球人口的0.6%~0.8%）。根据美国疾病控制和预防中心的数据显示，阿片类药物的滥用已经导致美国历史上最为严重的药物过量使用，并于2014年将该问题列入五大公共卫生挑战之一。阿片类药物大体可以为三大类：非合成阿片类药物:可待因、吗啡、鸦片；半合成阿片类药物:氢可酮、羟可酮、丁丙诺啡、海洛因；合成阿片类药物:芬太尼、布托啡诺、美沙酮、丙氧芬。2013 年以来芬太尼等合成阿片类药物相关的死亡人数增加，其中2016 年芬太尼及相关药物死亡人数就超过2 万人。1999~2017 年统计数据显示，非法使用阿片类药物的群体中，男性比女性占有更大的人数比例，18~25 岁的年龄群体且更容易接触到阿片类药品。  在过去几十年里，阿片类药物滥用不仅仅是美国存在的问题，包括我国、欧洲在内的世界大部分国家和地区都存在该问题。因此，应对阿片类药物滥用是一个全球性的公共卫生问题。这种药物成瘾对社会的核心稳定发展构成了严重的威胁。   * + 1. 研究意义   在过去二十年中，人类收集，存储，处理数据的能力取得了飞速提升，人类社会的各个角落都积累了大量数据，急需能够有效地对数据进行分析利用的计算机算法，而机器学习恰能顺应时代的迫切需求，因此机器学习这门学科领域很自然地取得巨大发展，受到了广泛的关注。今天，机器学习已经与普通人的生活密切相关。例如在天气预报、能源勘探、环境监测等方面，有效地利用机器学习技术对卫星和传感器数据进行分析，是提高预报和监测准确性的重要途径。  机器学习的研究思路主要是使用计算机模拟人类的学习活动，它是研究计算机识别现有知识、获取新知识、不断改善性能和实现自身完善的方法。这里的学习意味着从数据中学习，它包括有指导学习、无指导学习、半指导学习三种类别。有指导学习是指有结果度量的监督学习过程，希望根据一组特征对结果度量进行预测，例如根据病人的饮食习惯和血糖、血脂来预测糖尿病是否发作。通过学习已知数据集的特征和结果建立起预测模型来预测并度量未知数据的特征和结果。结果度量一般分为定量和定性两种，分别对应与统计学中的回归和分类问题。常见的有指导学习包括：决策树、神经网络、支持向量机、集成学习等。无指导学习是只能观察特征，没有结果度量，此时只能从总体中给出的样本信息对总体做出某些推断以及描述数据是如何组织或聚类的。并不需要某个结果度量或训练集，例如聚类分析、关联规则分析等。半指导学习是近年来机器学习中一个备受关注的内容，已得的观察量中一部分是经由指导者鉴定并加上标识的数据，称为已标识数据，另一部分观察量由于种种原因未能标识，称为未标识数据。通过利用已标识数据去分析并指出适当的一般性规律，再利用此规律去推断得出有关未标识数据的标识。但现在半指导学习方法的性能通常不太稳定，而且半指导学习技术在什么样的条件下能够有效地改善学习性能也不够明确。比较典型的方法便是朴素贝叶斯。  目前广泛应用的机器学习算法包括神经网络、K近邻法、支持向量机、决策树、随机森林等，这些算法已经广泛应用与工程学、建筑学等领域，却很少有研究将这些算法应用在公共卫生领域，为了更好地评估这些算法是否能够有效地预测阿片类药物使用量以及寻找具有最好分类效果的分类算法，本研究比较了6种算法的分类效能。   * 1. 国内外研究现状   国内外研究表明，阿片类药物的滥用是一个多影响因素且影响因素见关系复杂的公共卫生事件。目前关于此类问题预测的方法主要以传统的logistic回归模型（logistic regression）为主。但传统模型在使用条件上较为苛刻，需要考虑数据分布是否合适、变量之间的共线性和交互作用等多种问题，特别是对多变量的问题尤为明显。因此，应用传统回归模型进行预测阿片类药物使用情况具有一定的局限性。  机器学习作为人工智能领域的“内核”，通过研究历史数据抓取事务的本质特征，以模型或算法为代表的呈现方式，实现分类、预测、回归拟合等分析行为。对于多重共线性问题，机器学习中的神经网络模型可以很好的解决，而且对于数据的分布几乎没有任何限制。对于变量属性间的交互作用，贝叶斯网络模型可以反应不同因素间的相互关系。在自变量属性众多（如，成百上千个），传统模型不稳定甚至根本无法建立模型，集成学习却可以解决，且模型稳健。而在传染病扩散、毒品传播、疾病预后等多变量、影响因素间相互作用负责的情况下，机器学习算法模型可能显著优于传统预测模型。  目前在结合机器学习算法与阿片类药物滥用的研究中，彭巧娟等学者通过灰色关联度得到阿片类药物使用趋势与社会因素的相关性，利用神经网络进行修正验证。韦可等人通过使用多项式回归的方法，借助支持向量机确定了多项式具体参数，从而预测阿片类药物滥用率。他们所提出的模型虽然具有一定的普适性并能用于预测多个县的药物使用总量，但只是使用了机器学习中一种比较常见和普遍性比较高的算法，并不能比较除不同机器学习算法中的差异优劣。  正是由于阿片类药物使用量的影响因素复杂、使用量差异较大。且自变量之间存在多重共线性和交互作用。传统的预测模型在预测这类问题时受限，为解决这个问题，本研究利用人工智能核心技术的多种机器学习算法来预测阿片类药物使用量。通过评价比较不同的机器学习算法来确定最佳的模型，以此来描述和预测分析阿片类药物使用量问题。   * 1. 主要研究思路   对不同地区、不同类型的阿片类药物进行描述性分析。通过数据可视化来描述药物使用情况，包括阿片类药物使用的数量和地域趋势，通过可视化的描述得出初步猜测，以便后续深入的建模分析研究。  提取强相关性的社会经济属性。由于所分析的与阿片类药物使用相关的各种社会经济属性太多，且考虑到过拟合问题和简化模型，因此不适合简单的将所有社会因素加入向量模型中。因此利用相关系数来探索阿片类药物使用与社会经济数据之间的相关性。找出与阿片类具有较强的相关性的社会经济属性。  通过机器学习算法建立模型。利用多种机器学习算法来预测阿片类药物使用量。将强相关性的社会经济属性纳入到不同的机器学习算法中，具体包括如下算法：K-临近算法（K-Nearest Neighbor Algorithm）、支持向量机(Support Vector Machines)、决策树(Decision Tree)、神经网络（Neural Networks）、随机森林（Random Forest）、线性回归（Linear Regression）。对数据按照7:3的比例进行分层划分训练集和测试集。  评价和比较不同算法的优劣。通过不用方法的机器学习模型的预测结果的错误率、F1评分、支持度、K折训练集等性能指标评价和比较不同模型的优劣，以期进一步完善机器学习方法在阿片类药物使用量预测分析的统计学模型。   * 1. 数据来源与说明      1. 数据来源   数据来源于美国国家法医实验室信息系统（NFLIS）与美国人口统计局（U.S. Census Bureau）。数据包括美国7年(2010-2016年)462个县的麻醉性镇痛药和海洛因的药物鉴定计数，以及各年份各县的一系类常见的社会经济因素。这462个县主要来自于这五个州：俄亥俄州，肯塔基州，西弗吉尼亚州，弗吉尼亚州和宾夕法尼亚州。   * + 1. 数据说明   采用Office Excel2016收集数据，通过Python3.6软件对数据进行进一步分析。由于数据的复杂性已经部分数据存在缺失等情况，不便于直接对数据进行统计分析，因此分析前对数据进行了一定的整理，具体如下：  删除了没有测量值的因素；  删除只在特定年份测量的因素，因为多年的趋势将较少出现，更容易受到潜在的异常值影响；  删除所有县数据不完整的因素，由于可能存在由于数据缺失而不明显的隐藏趋势而使模型的效果产生一定程度的误差；  采用均数对缺失数据进行填补。   * + 1. 缺失值处理   收集到得数据某些字段值为空得情况很多，一般有三种处理方法：删除记录、数据填补、空值处理。填补缺失值方法有：人工填补、均数填补、中位数或众数填补、多重填补、使用最接近的样本值填补等。不同情况变量的缺失值本研究采用了不同的处理方式。   * + 1. 离散化处理   连续性变量的离散化就是将连续属性的值域划分成为若干个离散的区间，最后用不同的数值或者符合代表观测值落在每个区间的属性值。连续变量的离散化的目的是减少给定的定性变量的属性值得个数。在模型解释方面，离散化数据更容易解读，且消除了某些异常值的影响起到一定作用。本次实证研究将各县的阿片类药物使用报告量离散化成6个水平：0例、1-9例、10-99例、100-499例、500-999例、1000-4999例、5000例以上。   * + 1. 变量选取处理   冗余是数据集成的一个重要问题，冗余属性出现的情况可能为同一属性因为变量名不同，出现多次导致重复。通过使用相关性分析进行检测，检测出一个变量影响另一个变量的强度。  除此之外，由于社会经济属性较多，所以选择与应变量有较强的关联性的变量有助于模型的可解释性和精简性。因此通过计算各个自变量与应变量的相关系数来选择强相关属性。考虑到模型的广泛性，选择了2010年-2016年各个年份的社会经济属性与阿片类药物报告量的相关系数均大于0.5的所有变量，共有38个变量入选。   * + 1. 归一化处理   不同的变量有不同的量纲，比如人口、家庭户数、收入等，数值间的差别比较大，归一化就是为了消除变量之间量纲的影响，对原始数据进行标准化、归一化、中心化处理。标准化或归一化实质是一种线性变换，线性变换有很多良好的性质，这些性质决定了对数据改变处理后不会造成失效，反而提高数据的表现，这些性质是归一化或者标准化的前提。  归一化就是将数据映射到某一固定区间，一般为（0,1）或（-1,1）。主要是为了数据处理方便提出来的，把有量纲表达式变成无量纲表达式，便于不同单位或量级的指标能够进行比较和加权，归一化处理过程仅由变量的极值决定。  归一化公式：  标准化会使每个特征中的数据平均值变为0、标准差变为1，标准化是依照特征矩阵的列处理数据，通过就z-score的方法转换为标准正态分布。  标准化公式：  中心化便是数据的平均值变换为0，而标准差无确切要求。  中心化公式：  考虑到量纲不同问题，故本文对所有变量均用归一化方法进行了归一化处理。   1. 模型建立准备    1. 现状分析   在对不同种类不同地区的阿片类药物使用做预测模型之前，首先对各个州各类别阿片类药物的使用量做描述性分析，以初步阿片类药物使用量的情况和分布特点，为后续研究做铺垫。   * + 1. 不同类型阿片类药物的使用情况   报告的阿片类药物共有xx种，各类阿片类的分布主要集中在海洛因、氢考酮、氢可酮这三类，分别占比为53.24%、19.79%、8.57%，海洛因的报告量达到25000多例，氢考酮的报告量为10000多例，氢可酮的报告量为4000多例；大部分类型的阿片类药物都小于0.01%；而较为常见的可卡因、吗啡等占比并没有想象中的那么高，比例分别为0.73%、1.88%；  总的来说，各个地区的报告量占比的最小值与最大值差异极大。   * + 1. 不同地区阿片类药物的使用情况   根据2010-2016年不同州的不同种类报告量绘制了热力图。可以发现，虽然不同年份的报告量有一定的差异，但是在类别上展出来的差异更加明显，肯塔基州主要为氢考酮、海洛因、氢可酮阿片药，俄亥俄州主要为海洛因、氢考酮、丁丙诺啡阿片药，宾夕法尼亚州主要为海洛因、氢考酮、丁丙诺啡阿片药，弗吉尼亚州主要为氢可酮、氢考酮、海洛因阿片药，西弗吉尼亚州主要为氢考酮、氢可酮、海洛因阿片药；可以发现大致呈现3个板块，第一个板块主要以海洛因阿片药为主；第二板块主要以氢考酮为主，第三板块主要以丁丙诺啡为主。基于这三大板块类别便可以将原本56种阿片类药物重新划分为三大类，且在不同州不同年份的报告量的相关程度更高。   * 1. 划分类别   根据阿片类药物是否为合成药物将其划分为三大类：合成类阿片药、非合成类阿片药、半合成类阿片药。基于此对2010-2016年各个州的报告量绘制折线图。可以看出不同州的不同类的阿片药报告量在2010-2016年呈现不同趋势，肯塔基州合成阿片药在2010年报告量为峰值6208例，之后逐年递减，2016年为3089例，半合成类阿片药在2010-2014年以较大的速度递增，在2014年达到峰值7151例，非合成阿片药的报告量在2010-2016年的变化趋势不明显，保持在175-378之间；俄亥俄州半合成阿片药在2015年达到峰值26674例，但从2015年开始呈下降趋势，而合成阿片药报告量在2010-2016年呈现一个先减后增的趋势，在2013年最低为5060例，2016年最高为17347例，非合成阿片药的变化趋势不明显，保持在547-626例之间；宾夕法尼亚州的半合成阿片药的报告量大致也呈现一个先增后减的趋势，在2015年达到峰值20799例，合成类阿片药报告量在2013年出现一个波谷为3537例，非合成阿片药有一个较弱的下降趋势，2010年为675例，2016年为355例；弗吉尼亚州的半合成阿片药在2011年和2013年分别出现一个波谷和波峰，波谷为3092例，波峰为7062例，非合成阿片药在2010-2016年变化趋势不明显；西弗吉尼亚州的半合成阿片药在2013年出现一个波峰，波峰为2444例，合成类阿片药在整体上呈现一个递减趋势，2011年为1672例，2016年为988例；非合成阿片药在2010-2016年变化趋势不明显；而这五个州的总体上而言，半合成阿片类药物呈现一个先增后减的趋势，而合成类阿片药为先减后增的趋势，非合成类阿片药变化趋势相对不明显。   1. 机器学习预测模型    1. K近邻模型       1. 方法   K最近邻法（KNN，K-Nearest Neighbor）是1967年由Cover和Hart P提出的一种基本分类和回归方法，它是数据挖掘分类技术中最简单的方法之一，非常容易理解应用。基本思想为：根据距离函数计算待分类样本x和每个训练样本的距离，选择与待分类样本距离最小的K个样本作为x的K个最近邻，最后根据x的K个最近邻判断x的类别。通常K是不大于20的整数，上限是数据集数量的n次开放，随着数据集的增大，K的值也要增大。依赖于训练数据集和K的取值，输出结果可能会有不同，所以需要评估算法的正确率，选取产生最小误差率的K。而关于距离的度量方法常用的主要有以下几种：   1. 欧氏距离：   计算基于各维度特征的绝对值，欧式距离需要保证各维度指标在相同的刻度级别。   1. 明式距离：   明式距离是欧式距离的推广，是对多个距离度量公式的概括性表述，当公式中p=2时就得到欧式距离。   1. 曼哈顿距离：   曼哈顿距离来源于城市区块距离，是将多个维度上的距离进行求和后的结果，即当明式距离公式中p=1时得到的距离度量公式即为曼哈顿距离。   1. 切比雪夫距离：   KNN进行分类预测的具体步骤如下：   1. 计算测试数据与各个训练数据之间的距离； 2. 按照距离的递增关系进行排序； 3. 选取距离最小的K个点； 4. 确定前K个点所在类别的出现频率； 5. 返回前K个点中出现频率最高的类别作为测试数据的预测分类。    * 1. 结果   在Python中本文使用的是sklearn包中的KNeighborsClassifier实现K最近邻算法分类，对三种阿片类药物报告量预测做KNN分类。采用精确率、召回率、F1-评分和支持度这4个指标进行简单的预测性能评价。为测试集样本预测评估结果，KNN模型在预测三类阿片类药物的使用量方面的预测性能都一般（0.5<精确度<0.7）。总体上而言，非合成阿片类药物报告量的预测效果明显优于合成类阿片类药物和半合成阿片类药物的效果。   * 1. 决策树模型      1. 方法   一般而言，一颗决策树包含一个根节点、若干个内部节点、若干个叶子节点；叶子节点对应决策结果，其他每个节点对应一个属性测试，每个节点包含的样本集合根据属性测试的结果被划分到子节点中；根节点包含样本全集，从根节点到每个叶子节点的路径对应了一个判定测试序列。决策树学习的目的是为了产生一颗泛化能力强的决策树。而决策树算法的关键在于如何选择最优划分属性。一般而言，随着划分过程不断进行，我们希望决策树的分支节点所包含的样本尽可能属于同一类别，即节点的“纯度”越来越高。度量样本集合纯度最常用的一种指标便是信息熵，一般而言，信息增益越大，则意味着使用该属性来进行划分所获得的“纯度提升”越大。因此，我们可用信息增益来进行决策树的划分属性选择。著名的ID3决策树学习算法就是以信息增益为准则来选择划分属性的。决策树算法处理过拟合的主要手段便是通过剪枝，基本策略包括预剪枝和后剪枝。剪枝处理降低了过拟合风险，但也带来了欠拟合的风险。一般情况下后剪枝决策树欠拟合风险较小，泛化能力性能往往优于预剪枝决策树。但后剪枝训练开销比未剪枝和预剪枝决策树都要大很多。  决策树进行分类预测的具体步骤如下：   1. 对数据源进行数据预处理，得到训练集和测试集； 2. 对训练集进行训练； 3. 对初始决策树进行剪枝； 4. 由所得到的决策树提取分类规则； 5. 使用测试数据集进行预测，评估决策树模型。    * 1. 结果   在Python中本文使用的是sklearn包中的DecisionTreeClassifier实现决策树分类，对三种阿片类药物报告量预测建立回归树模型。采用精确率、召回率、F1-评分和支持度这4个指标进行简单的预测性能评价。为测试集样本预测评估结果，决策树模型在预测三类阿片类药物的使用量方面的预测性能都一般（0.5<精确度<0.7）。总体上而言，半合成阿片类药物报告量的预测效果明显优于合成类阿片类药物和非合成阿片类药物的效果。   * 1. 随机森林模型      1. 方法   作为新兴起的、高度灵活的一种机器学习算法，随机森林（RF，Random Forest）拥有广泛的应用前景，从市场营销到医疗保健，既可以用来做市场营销模拟的建模，统计客户来源，保留和流失，也可用来预测疾病的风险和病患者的易感性。  随机森林是一种组成式的有监督学习方法。由多颗决策树构成，且森林中的每一颗决策树之间没有关联，模型的最终输出由森林中的每一颗决策树共同决定，对于分类问题是，森林中每颗决策树会给出最终类别，最后综合考虑森林内每一颗决策树的输出类别，最后通过投票方式来决定测试样本的类别。随机森林的随机性主要体现在两个方面：样本的随机性，从训练集中随机抽取一定数量的样本，作为每颗决策树的根节点样本；属性的随机性，在建立每颗决策树时，随机抽取一定数量的候选属性，从中选择最合适的属性作为分裂节点。  随机森林进行预测的步骤如下：   1. 从训练集中随机抽取一定数量的样本，作为每棵树的根节点样本； 2. 在建立决策树时，随机抽取一定数量的候选属性，从中选择最合适属性作为分裂节点； 3. 建立好随机森林以后，对于测试样本，进入每一颗决策树进行类型输出或回归输出；若是分类问题，以投票的方式输出最终类别，若是回归问题，每一颗决策树输出的均值作为最终结果。    * 1. 结果   在Python中本文使用的是sklearn包中的RandomForestClassifier实现随机森林分类，对不同类的阿片类药物报告量预测做随机森林回归。对于随机森林模型，合成阿片类报告量的预测效果最差，合成阿片类药物的精确度为0.63，召回率为0.62，F1-评分为0.62，支持度为961；非合成阿片类药物的效果次之，精确度为0.64，召回率为0.62，F1-评分为0.63，支持度为897；半合成类阿片药的效果最优，精确度为0.64，召回率为0.62，F1-评分为0.63，支持度为961。   * 1. 支持向量机模型      1. 方法   支持向量机（SVM，Support Vector Machine）是一种监督学习算法，可用于分类和回归问题。主要目标是找到最佳超平面，以便在不同类的数据点之间进行正确的分类。超平面维度等于输入特征的数量减去1（例如，当使用4个特征时，超平面维度为3）。为了分离两类数据点，可以选择许多可能的超平面，而支持向量机的主要目的便是取找到一个具有最大边距的平面，即两个类的数据点之间的最大距离，最大化边距的目的是最大概率的对未知的数据点进行正确分类。超平面是决策边界，有助于对数据点进行分类。落在超平面两侧的数据点可归因于不同的类。最接近超平面的数据点称为支持向量。支持向量确定超平面的方向和位置，以便最大化分类器边界（以及分类分数）。如果使用的数据不是线性可分的（导致线性支持向量机分类结果不佳）,则可以应用Kernel Trick技术，能够将非线性可分离数据映射到更高维空间，使原始数据可线性分离。然后使用这个新的维度空间支持向量机可以很容易地实现。支持向量机中有许多不同类型的内核可用于创建这种更高维的空间，例如线性，多项式，Sigmoid和径向基函数。在分类问题中，除了线性的逻辑回归模型和非线性的深度神经网络外，还可以应用一种被广泛应用的模型—支持向量机，这种算法在学习复杂的非线性方程时提供了一种更为清晰，更为强大的方式。支持向量机的主要有以下优势：高维空间有效；维度高于采样数量的情况依然有效；在决策函数中使用训练点的子集（称为支持向量），因此节约内存；功能多样：可以为决策函数指定不同的核函数。支持向量机主要有以下缺点：如果特征的数量大大多于采样的数量，在选择核函数时避免过拟合，正则化至关重要；支持向量机并不直接提供概率估计。   * + 1. 结果   在Python中本文使用的是sklearn包中的SVM实现支持向量机分类，对不同类的阿片类药物报告量预测做支持向量机模型分类。对于支持向量机模型，半合成阿片类报告量的预测效果最差，合成阿片类药物的精确度为0.67，召回率为0.60，F1-评分为0.64，支持度为961；合成阿片类药物的效果次之，精确度为0.72，召回率为0.61，F1-评分为0.66，支持度为961；半合成类阿片药的效果最优，精确度为0.84，召回率为0.64，F1-评分为0.69，支持度为897。   * 1. 神经网络模型      1. 方法   人工神经网络（ANNs，Artificial Neural Networks）起源于上世纪40-50年代，使通过模仿生物神经的行为特征，形成一种具有学习、联想、记忆和模式识别等智能信息处理的人工系统，称为人工神经网络。使通过模仿生物神经的行为特征，进行分布式并行信息处理的数学模型。这种网络依靠系统的复杂度，通过调整内部大量节点之间相互连接的关系，从而达到信息处理的目的。人工神经网络具有自学习和自适应的能力，可以通过预先提供的一批相互对应的输入输出数据，分析两者的内在关系和规律，最终通过这些规律形成一个复杂的非线性系统函数，这种学习分析过程被称作“训练”。神经元的每一个输入连接都有突触连接强度，用一个连接权值来表示，即将产生的信号通过连接强度放大，每一个输入量都对应有一个相关联的权重。处理单元将经过权重的输入量化，然后相加求得加权值之和，计算出输出量，这个输出量是权重和的函数，一般称此函数为传递函数。激活函数的作用是能使得神经网络的每层输出结果变得非线性化，进行数值转换，具有如下性质：可微性，非线性、单调性、输出值于输入值相差不会很大。   * + 1. 结果   在Python中本文使用的是sklearn包中的MLPClassifier实现支持人工神经网络分类，对不同类的阿片类药物报告量预测做人工神经网络模型分类。对于人工神经网络模型，合成阿片类报告量的预测效果最差，合成阿片类药物的精确度为0.67，召回率为0.62，F1-评分为0.64，支持度为961；半合成阿片类药物的效果次之，精确度为0.65，召回率为0.59，F1-评分为0.62，支持度为961；非合成类阿片药的效果最优，精确度为0.72，召回率为0.65，F1-评分为0.67，支持度为897。   * 1. 逻辑回归模型      1. 方法   Logistic回归（LR，Logistic Regression）模型是非线性概率回归模型。作为一种流行病学多元分析方法，被广泛应用于探索二元应变量于影响因素之间关系的研究，例如通过探索疾病的影响因素并根据危险因素取预测疾病的发生率。  Logistic回归进行预测的步骤如下：   1. 划分数据集为训练集和测试集； 2. 基于训练集建立逻辑回归模型； 3. 通过模型计算测试集的预测输出结果； 4. 计算损失函数，对比预测结果和实测结果，获得最佳。    * 1. 结果   在Python中本文使用的是sklearn包中的LogisticClassifier实现支持逻辑回归分类，对不同类的阿片类药物报告量预测做逻辑回归模型分类。对于支持逻辑回归模型，半合成阿片类报告量的预测效果最差，合成阿片类药物的精确度为0.72，召回率为0.58，F1-评分为0.63，支持度为961；合成阿片类药物的效果次之，精确度为0.73，召回率为0.62，F1-评分为0.67，支持度为961；半合成类阿片药的效果最优，精确度为0.78，召回率为0.66，F1-评分为0.69，支持度为897。   1. 模型评估与结果    1. 模型评估   本文共计使用了KNN、决策树、随机森林、支持向量机、神经网络、逻辑回归6个机器学习模型对阿片类药物报告量进行预测。并用K折交叉验证法作为参考对模型效果进行评估，从图中可以看出，三类阿片类药物报告量结果中，效果最好的是支持向量机模型，其次是人工神经网络模型，相比传统的逻辑回归模型，这两个机器学习模型在解决实际分类预测问题上占有一定的优势。从图中还可以看出，KNN、决策树、随机森林这三个机器学习模型在于传统逻辑回归模型对比不占明显优势，支持向量机、神经网络和逻辑回归模型的预测效果表现了相当高的一致性，说明模型的稳健性很好。因此本文以K折交叉验证方法为评估指标得出最优的模型为支持向量机模型。   * 1. 模型结果   与阿片类药物报告量相关的社会经济因素的重要程度除了线性关系（相关系数衡量）以外，还可能存在一些非线性相关关系。因此评估影响因素重要度的准确方法是考虑线性和非线性关系。对于线性相关的分析在前文的数据处理中便已经实现，所有模型使用的变量均是均有相对强相关性的变量，在此基础上进行的模型拟合。  然而对于非线性相关分析几乎不可能像线性相关分析那样进行量化，为了尽可能确定非线性的重要程度的存在，本文综合了上述KNN、决策树、随机森林、支持向量机、神经网络、逻辑回归6种方法来对模型的变量进行重要度计算，综合各个模型变量的重要度得到各个变量的综合重要度排序。在这个过程中尽管可能存在一些极值上的误差，但是在总体属性排序过程总只存在不到十分之一的情况，其余情况表现均较为良好。因此，最终得到的重要度排序在某些方面而言具有重要的作用和意义。  从结果来看，三类阿片类药物的报告量相关的社会经济因素重要度相关前20的变量主要属于以下几类：婚姻状况、受教育程度、家庭类型、非制度化平民的残疾状况、入学率、出生地方。除此之外，还有一些细分的类别，例如平均家庭规模是一个新的细分属性类别下的家庭类型。  结论  机器学习模型对于克服传统预测模型的多重共线性效应、变量交互作用和数据分布要求有良好的作用。支持向量机模型和人工神经网络模型在预测效果中的性能都比较理想。支持向量机模型的精确度>0.8，而人工神经网络模型的精确度>0.7。对比与传统的预测模型而言，支持向量机和神经网络模型容易调整参数，能够生成预测效能更好的模型。  对于决策树和随机森林模型，由于随机森林模型本质上是在以决策树为基学习器构建Bagging集成的基础上，进一步决策树训练过程中引入了随机属性选择。所以本文也证实其理论性，发现随机森林模型的预测效能要明显优于决策树模型。随机森林模型能够集成多颗决策树进行预测，客服了单颗决策树泛化能力不足的缺点，有助于模型的外推。同时，集成学习模型也容易调整参数，可以生成预测效能更好的模型，为类似研究提供新思路和方法。  KNN模型作为最简单的机器学习模型之一，在方法思路上具有简洁的优点，但是其实现过程所需计算机空间内存较大，且本文的实证研究发现其预测效能一般。但是相比于传统的统计预测模型而言，容易调整参数和思路简洁是一个明显的优势。  传统的logistics回归在本实证研究中的预测效能表现虽然不是最佳，但是明显优于决策树和随机森林等模型。但是传统的logistics模型的多重共线性和变量间相互关系作用的缺点。  本文从各类阿片类药物使用情况的研究背景出发，分析了不同年份和不同地区阿片类药物报告量的现状，在公共卫生事业越来越重要的情况下，成瘾性药物的研究更应该引起学者们的注意。对阿片类药物使用情况的预测不仅可以让卫生管理人员直观看到社会情况，还可以帮助相关政府部门制定必要的卫生政策。本文利用统计方法和机器学习技术，对不同类的阿片类药物做精准预测。主要工作如下：  不同地区阿片类药物报告量、相关社会经济指标数据的采集和清洗；利用数据可视化对阿片类药物报告量的描述性统计分析；相关社会经济指标数据的特征选择；不同类型的阿片类药物报告量的机器学习建模。  总结来说，本文能将多种机器学习算法运用到阿片类药物报告量的预测中，更具机器学习模型在阿片类药物报告量数据上的表现，做到分类处理的精准预测。  虽然本文对不同类的阿片类药物报告量预测做到了不错的效果，但是本文的工作还有很多的不足之处，其中包括以下几个方面：  虽然收集到2017年阿片类药物的报告量，但并没有收集到当年分的社会经济指标数据，导致大量数据失效。且在数据清洗过程中清洗掉大量数据，最终使用到的有效数据只有9000多条，相对于机器学习模型而言数据量偏少。  社会经济指标未能构造多维，本文可以通过已有变量构造其他变量。  使用的算法模型不够新。本文使用到的机器学习算法都是比较传统的机器学习算法，这些算法在一定程度上还有可以优化的空间。比如梯度提升回归树算法、CART决策树和C4.5决策树算法。 | | | |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **课题综合评价（包括存在的主要问题和改进建议）：**  已经完成初稿提纲。预计能够按照计划安排的时间来完成研究内容。  **指导教师签名：**  **2020年 6 月 6 日** | | | | | | |
| **参加检查人员** | **姓名** | **职称** | **学科、专业** | **签名** | | |
|  |  |  |  | | |
|  |  |  |  | | |
|  |  |  |  | | |
|  |  |  |  | | |
|  |  |  |  | | |
| **学院毕业论文工作领导小组审核意见：** | | | **负责人签字：**    **年 月 日** | |  |  |