基于机器学习算法的脱发原因探索与识别

学号：2341689

姓名：邬嘉恺

2024年6月

摘 要

在现如今的社会中，随着年龄的增长，越来越多的人开始关注脱发这一健康问题。根据阿里健康信息技术公司前几年公布的数据显示，我国目前大约已有超2.5亿人深受这一问题的困扰，其中男性大约占七成；而从脱发年龄段来说，30岁以下人群，已成为防脱发消费的主力。不同的研究认为，脱发与家族遗传因素、不良的生活习惯、激素分泌、精神压力等因素息息相关。

本课题利用Kaggle竞赛平台的有关脱发数据集，分析此数据集中样本的统计性分布情况，并加以描述，必要时进行相关内容的可视化展示；而在机器学习部分，主要探究如何训练分类模型，以预测个体是否会患上脱发，与此同时通过聚类算法，将不同类型的脱发群体由无监督学习的手段进行归类，最后通过算法，将所有数据集中的影响因素由其权重从高到低表示，以判断出其中最能识别出预测脱发的关键因素。

关键词：脱发，描述性统计，数据可视化，机器学习

目 录

[基于机器学习算法的脱发原因探索与识别 1](#_Toc170307955)

[1 引言 1](#_Toc170307956)

[1.1 问题的引出 1](#_Toc170307957)

[1.2 数据集介绍 1](#_Toc170307958)

[2 研究的问题与方法 2](#_Toc170307959)

[2.1 描述统计性问题 2](#_Toc170307960)

[2.2 数据可视化问题 2](#_Toc170307961)

[2.3 机器学习问题 3](#_Toc170307962)

[3 结果与分析 3](#_Toc170307963)

[3.1 描述性统计问题 3](#_Toc170307964)

[3.2 数据可视化问题 4](#_Toc170307965)

[3.3 机器学习问题 6](#_Toc170307966)

[4 问题讨论及总结 7](#_Toc170307967)

1 引言

1.1 问题的引出

脱发，正逐渐成为影响当下中国人健康的问题。阿里健康信息技术公司提示，截止到2023年，我国约有2.5亿人受到此问题的困扰，其中男性脱发人数约1.63亿，女性约0.88亿；而从脱发的年龄段而言，调查显示，30岁前脱发比例竟高达84%，比上一代的脱发年龄整整提前了约20年。

从各部分研究结果来看，能导致脱发的因素同样迥异。然而影响的主要因素体现在如下几方面：

1. 不良的生活习惯及环境

不良的生活习惯主要体现在以熬夜为代表的行为上。外部研究显示，受访者中有上述不良行为的，最终有84%正遭遇着脱发的困扰。

1. 压力过大

联想到以程序员为代表的一些职业，我们经常会衍生出一些关键词，如“发际线”等。这也从侧面反映出，这些高强度高压力的职业，通常会与脱发等健康问题相关。第三方研究表明，受访者中，凡是因各种原因遭遇情绪/精神压力过大的，有73%的样本有着脱发的问题。

1. 代谢不良

代谢，对于人类来说，也是至关重要的一种生理活动。从机构的研究中发现，受访者有61%的内分泌失调样本，正忍受着脱发的折磨。

本课题有关问题的引出，正是出于对以上背景的考虑，希望挖掘以上这些因素与脱发之间的关联，从而为个体健康管理、医疗干预以及相关健康产业的发展提供有益的参考。

1.2 数据集介绍

本次的脱发数据集，选择的是Kaggle竞赛平台中有关的内容。有关数据集的样式截图如下：

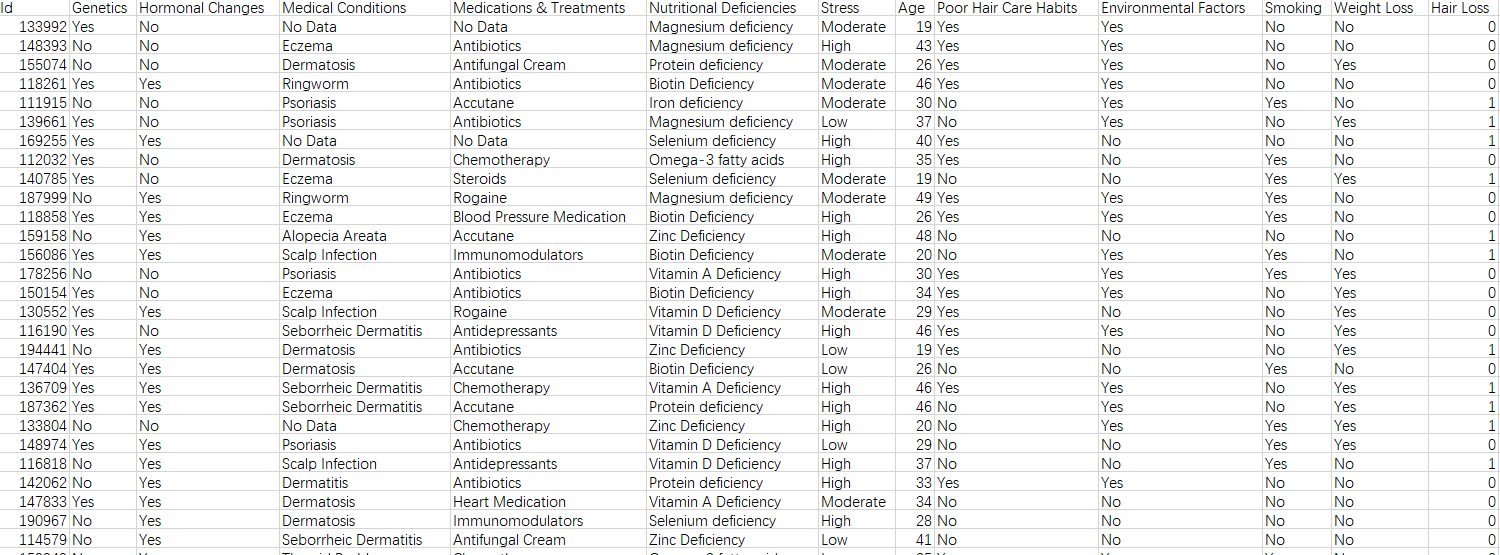


图1 脱发数据集示例

在此数据集中，共有999份受访者的信息，作为模拟非概率抽样的结果。每份信息又分为影响脱发的各类因素（即自变量）以及“是否脱发”（即因变量）。其中影响脱发的各类因素被考虑为如下几类：

（1） 定类变量

包含“遗传因素”、“激素变化”、“导致脱发的病史”、“可能导致脱发的药物治疗史”、“营养不足情况”、“是否存在不良护发习惯”、“是否暴露于可能导致脱发的环境”、“是否吸烟”、“是否经历显著体重减轻”、“年龄”等；

（2） 定序变量

包含“压力等级水平”等；

而作为因变量的“是否脱发”，自然作为定类变量。

2 研究的问题与方法

本课题主要通过Jupyter Notebook编写Python程序来解决相关问题。进行数据分析前，首先需要对数据集进行清洗，确保数据集的完善。如下图所示：

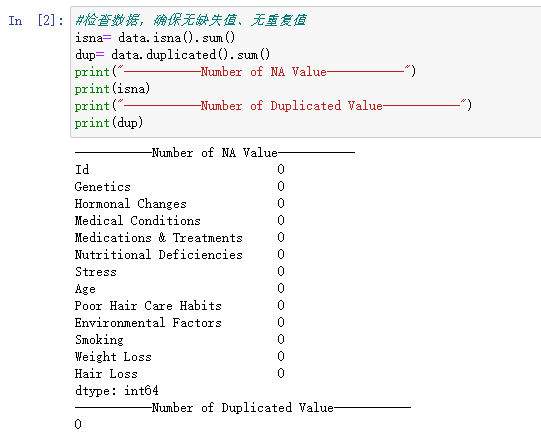


图2 检查数据集，确保无空值无重复值

研究的问题主要分为三类：

2.1 描述统计性问题

（1） 平均年龄是多少？年龄的分布情况如何？

对维度“年龄”字段进行mean函数的处理，即可得到样本中的平均年龄大小；之后对相关的字段列进行柱状图的可视化展示，横坐标以10的极差将年龄段分段。

（2） 哪些医疗条件是最常见的？它们的发生频率是多少？

对维度“可能导致脱发的药物治疗史”进行频数统计，主要用的函数为value\_counts()。发生频率则可由发生单一事件的频数除以总频数之和得到，最后使用帕累托图探究其中的主要原因。

（3） 营养缺乏的种类及其出现频率如何？

对维度“营养不足情况“进行统计，原理类似问题（2）。

2.2 数据可视化问题

（1） 不同年龄段中患者脱发的比例是多少？

使用饼图以及柱状图，将不同年龄段中脱发的比例表示出来，年龄段同上，取极差为10分段。作图时使用plt函数。

（2） 哪些因素与脱发之间存在关联？

考虑到自变量因素与因变量皆为分类数据的性质，此时考虑两者关联性首先考虑列联分析，即列联表结合卡方检验的方式；若结果不显著，则进一步考虑逻辑回归分析方法。

（3） 不同压力水平下脱发的情况如何？

本问题采取的是针对维度“压力等级水平“的所有类别（高、中、低），对不同的等级水平进行岁数分布的可视化展示的过程。具体操作为：分别分出等级水平为以上三类的数据框，使用seaborn库下的boxplot函数，以箱线图的形式，将年龄分布的情况表示出来（最小值、0.25分位数、中位数、0.75分位数、最大值），从而判断有压力条件下，年龄分布的一些规律。

2.3 机器学习问题

（1） 能否建立一个分类模型来预测个体是否会患上脱发，以基于给定的因素进行预测？

此问题我们将数据集按7：2：1的比例划分为训练集、验证集以及测试集。随后通过独热编码器将上述分类变量转换为连续型变量。尝试的过程中，我们首先通过传统机器学习方法，分别用XVM/随机森林/XGBoost的方法建立模型，并预测其准确率；之后视效果情况进一步采用神经网络算法，试图提高相关预测的准确率。

（2） 利用聚类分析，探索数据集中是否存在不同类型的脱发群体。

此问题我们可采用类似于K-Means的原理，但主要应用于分类变量的K-Modes算法。在进行无监督学习时，考虑利用肘部法则，将要聚成的类数先进行一个最优解的确定，之后通过K-Modes将不同类型的脱发群体进行聚类，

（3） 使用决策树或随机森林等算法，识别出最能预测脱发的关键因素。

此处我们主要使用XGBoost算法，之后提取各因变量在此情况下的特征重要性数值，并按照重要性降序排序，之后通过柱状图的形式将各因素权重可视化表示出。

3 结果与分析

经过上面对有关问题解决思路的分析，通过编写Python语句，实现对这些问题的解答。

3.1 描述性统计问题

（1） 平均年龄是多少？年龄的分布情况如何？

经过对数据集的数据分析发现，样本的平均年龄为34.19岁，年龄的分布情况如下图所示，可以看到在18~50岁这个年龄段内，分布还是比较均匀的：

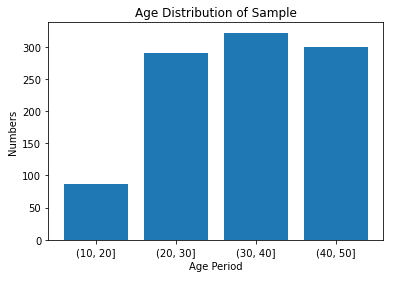


图3 数据集年龄分布情况

（2）哪些医疗条件是最常见的？它们的发生频率是多少？

经过对数据集有关“可能导致脱发的药物治疗史”进行频数统计“，显示有斑秃（10.71%）、银屑病（10.01%）、皮脂溢出性皮炎（8.81%）等病史的样本在脱发人群中较为常见。具体的分布情况如下图所示：

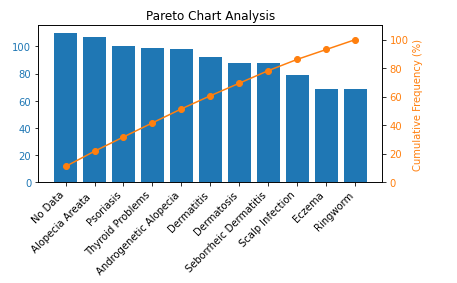


图4 数据集常见医疗条件主要构成分析

（3） 营养缺乏的种类及其出现频率如何？

根据饼图的绘制可以发现，样本数据集中对于营养缺乏的种类分布较为均匀，各部分的组成都较为平均。

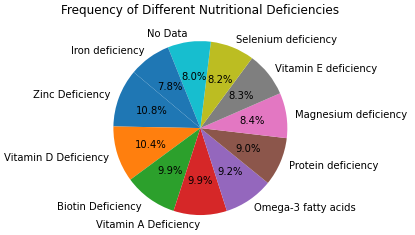


图5 数据集样本营养缺乏频率分布

3.2 数据可视化问题

（1） 不同年龄段中患者脱发的比例是多少？

此处以柱状图为例，可以看到，无论何种年龄段，患者脱发的比例在其所在的年龄段内都较为平均，占比大约在50%左右。

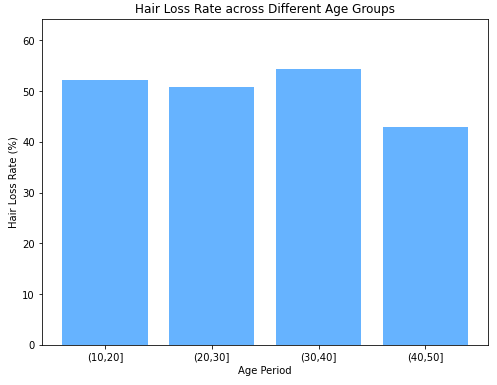


图6 数据集样本各年龄段脱发率分布情况

（2） 哪些因素与脱发之间存在关联？

此处首先进行列联分析，即构造列联表，之后进行卡方检验。通过这样的方法，对每一个自变量维度，与因变量进行相关匹配。以维度“激素变化”为例，进行列联分析后的结果如下图所示：

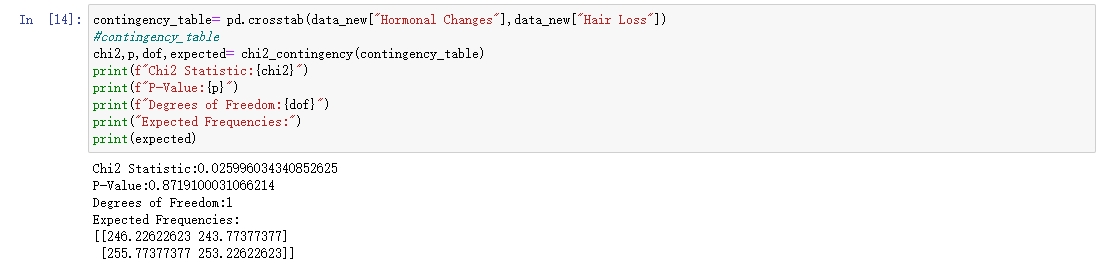


图7 列联分析代码结果示例

类似的列联分析还有很多。但与图7相似，发现实际的p值，都远远超过显著性水平α（0.05）。这表明，至少通过该类方法，是无法得出任何自变量因素与脱发这个因素相关的。

基于这样的背景，另一种判断方式应运而生，即逻辑回归分析方法。通过这种方法来确定每个自变量前的系数，并利用系数将其转化为热力图。结果中应注意以下两点：

1. 此系数的正负，代表该自变量对因变量影响的方向性；
2. 系数的绝对值大小，代表该自变量对因变量影响的大小

针对问题中的各个自变量，通过逻辑回归分析方法，我们便可得到如图8所示。从图中不难看出，与脱发关联性较大的又比如：有湿疹疾病史、癣、缺乏镁摄入等情况。这可能为病患的某些预防性措施的实施，提供了提示作用。

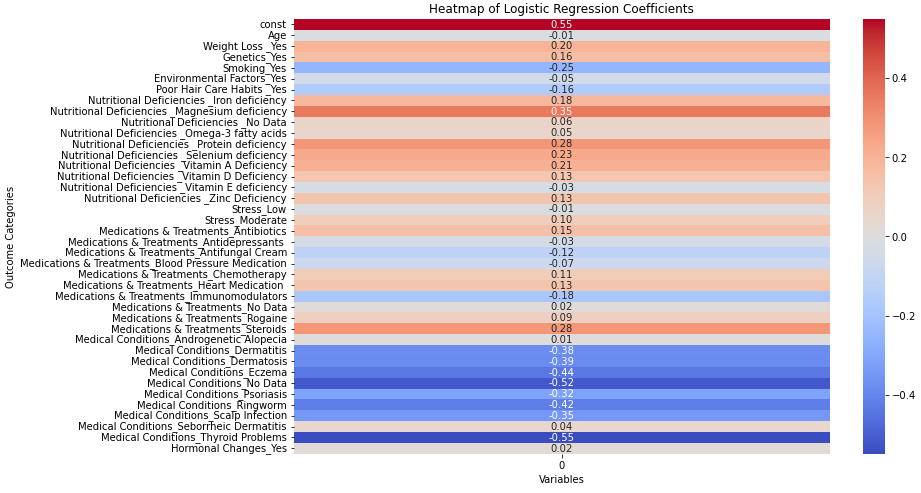


图8 逻辑回归分析下的相关性热力图

1. 不同压力水平下脱发的情况如何？

根据箱线图结果可以得知，从不同压力等级水平下的情况来看，并没有明显随着压力而产生岁数的分布改变；但可观察到，在不同等级的压力水平下，样本岁数分布中位数都集中在30~35岁的区间内，这表明，脱发水平确实在逐渐的“年轻化”。

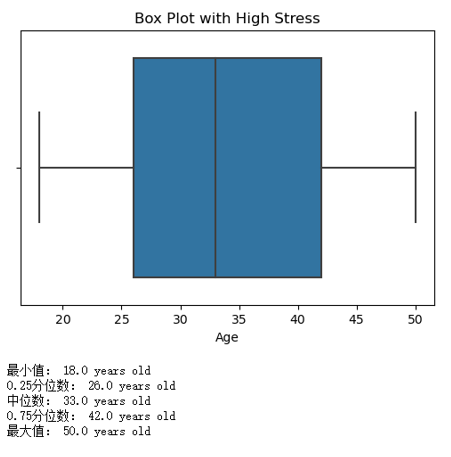


图9 不同压力等级下相似的箱线图结构

3.3 机器学习问题

（1） 能否建立一个分类模型来预测个体是否会患上脱发，以基于给定的因素进行预测？

实际操作中先用传统机器学习方法（支持向量机/随机森林/XGBoost），对其进行训练后发现预测的准确率分别为0.52，0.475，0.48。从中不难看出，通过这些方法做的准确率并不是特别高，因为就算不用机器学习方法，而用猜测的手段，在非黑即白的结果下，准确率也能有0.5。

这就激发了我们进一步采用深度学习方法提高准确率的想法。当使用一维CNN神经网络算法时，采用Adam优化器，可得到预测准确率能提升到将近0.6。

（2） 利用聚类分析，探索数据集中是否存在不同类型的脱发群体。

采用K-Modes进行聚类，首先利用肘部法则，将聚的类数选出最优解；此后，再用K-Modes算法对其进行聚类，实际得到的聚类效果图如下：

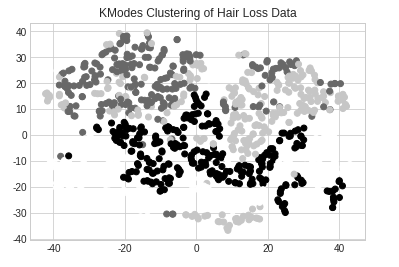


图10 不同脱发人群聚类结果

（3） 使用决策树或随机森林等算法，识别出最能预测脱发的关键因素。

采用XGBoost方法进行模型的训练，按降序顺序提取影响的重要性，如下图所示：

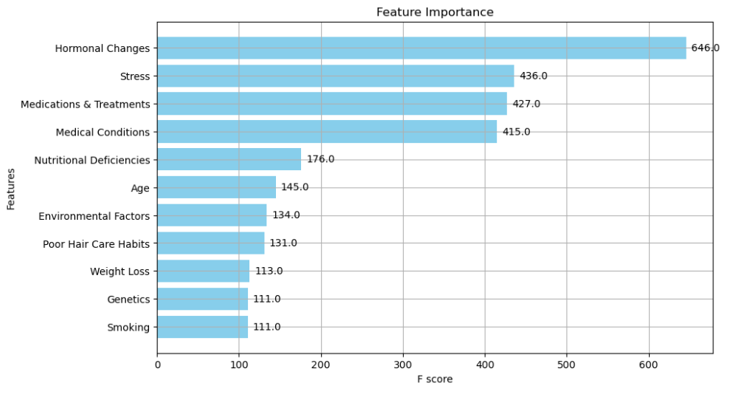


图11 影响脱发的权重降序排序

从图中可以看出，对于脱发因变量影响最大的几个因子分别是：激素变化（646.0）、压力（436.0）、可能导致脱发的药物治疗史（427.0）、可能导致脱发的病史（415.0）。以这些为代表的因素都是预测脱发的重要因素。

4 问题讨论及总结

从以上对本课题的探索中不难发现，最大的一个问题便是对于分类模型的预测准确性仍不够高。即使带上了CNN神经网络算法，对数据集进行预测，其准确率仍不足0.6。针对这个问题，我们在原有的CNN使用的基础之上，在网上寻找了一些资料作为建议，发现当其与一种叫PBT的算法结合在一起时，能使得此项目的准确率得到进一步提升，达到略超0.6的水平，这也是我们目前能找到的针对这个脱发数据集的数据分析问题，效果相对最好的一种算法方案。

其实针对这样的一个问题，我们也在思考究竟是什么原因导致了这样的结果。在对数据集本身进行检查时，我们发现，有许多数据或许存在着不太合理的取值，或者说，从某种意义上有悖于我们认知的取值。例如上文图6得出的结论，其实我们可以明显地发现，处于40~50岁年龄段的样本，其脱发率相较于更年轻的年龄段反而有一定程度的降低；又比如图5得到的相关结论中，我们发现所有的样本基本都是均匀分布，感觉像是被“设计好”的。显然，在现实的非概率抽样调研中，如果得到这样的一份数据集，首先数据本身不可能如此完善，其次也不会出现这样的“巧合”，或是说反常识结论。虽说科学有依据的方法得出的反常识结论反而对于数据分析可能是更有价值的，但假如非概率抽样的方法出了问题，那么得出的结论可能仅仅是片面的、缺乏理论支撑的。

实际上，我们在进行有关机器学习的实验中，不仅仅是分类模型预测的准确率可能受到了数据集本身的影响，之后的聚类算法等，其效果也或多或少地受到了影响。在进行肘部法则对最优聚类数的确定时，我们能观察到的效果其实并没有达到最佳。

这就尤其为我们做调研时做数据收集的方法提出了更高的要求。当然，作为一次以机器学习为主要目的的数据分析项目的尝试，我们通过此次的项目，感受到了一个简单的数据分析任务的完整流程，同时回头看写的代码，深刻意识到还有许多可以精简的地方，这都是未来我们继续努力的方向。

总而言之，通过这次课题的探究，我们进一步学习了有关脱发这一社会健康问题的影响因素，结合数据分析的各种手段，将机器学习课程中所学的各类知识运用到了实践中。无论如何，通过这样的实践，我们认识到：守护发质的健康，已经成为现代人一项愈发不可忽视的议题。让人们重视起来，可能并非一件易事，结合这样的一次课题，希望我们都能一起努力重视起来！

参考文献

# [1] 中食科康. 脱发报告|关注我国2.5亿人群的脱发困扰[DB].2023-11-28. https://mp.weixin.qq.com/s?\_\_biz=MzU2Mzg5MTM1OQ==&mid=2247494561&idx=1&sn=420770fb9efdcbd306eb32373a982cf4&chksm=fc51f651cb267f47f8901bf555b4f7f0a596fd741e03c3a6a78af01b03a0ec9db7286be721fb&scene=27