Statistics and Application 统计学与应用, 2025, 14(12), 19-25
Published Online December 2025 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2025.1412341

Hans汉斯

基于随机森林算法的个性影响因素研究

盛　林

广西师范大学数学与统计学院，广西　桂林

收稿日期：2025年10月27日；录用日期：2025年11月18日；发布日期：2025年12月1日

摘　要

个性形成受到父母、学校和社会等因素影响，虽然个性一经形成，比较稳定，但并非不可改变。尤其是未成年的学生，个性的可塑性还是很强的。本文通过对个性因素数据分析，帮助老师改善学生性格，促进个性内向学生全面发展。本文引入随机森林算法对个性数据集进行研究，利用随机森林算法准确率高，处理大数据拟合效果好的优势。在建立CART决策树和随机森林模型过程中，利用网格搜索，获得最优的参数组合。同时引入ROC曲线，将CART决策树和随机森林模型的效果进行实验对比，结果显示随机森林模型的效果更好，能够准确对个性影响因素进行分析。最后，对特征变量进行重要性排序，为教师教学提供一定的参考。

关键词

个性，随机森林，决策树，社交

Research on Personal Influencing Factors Based on Random Forest Algorithm

Lin Sheng

School of Mathematics and Statistics, Guangxi Normal University, Guilin Guangxi

Received: October 27, 2025; accepted: November 18, 2025; published: December 1, 2025

Abstract

Personality formation is influenced by factors such as parents, schools, and society. Although once formed, personality is relatively stable, it is not unchangeable. Especially for underage students, the plasticity of personality is still quite strong. This paper analyzes the data of Personality factors to help teachers improve students' characters and promote the all-round development of introverted students. This paper introduces the random forest algorithm to study the personality dataset, taking advantage of its high accuracy and good fitting effect on big data. During the establishment of

文章引用: 盛林. 基于随机森林算法的个性影响因素研究[J]. 统计学与应用, 2025, 14(12): 19-25.
DOI: 10.12677/sa.2025.1412341

CART decision tree and random forest models, grid search is used to obtain the optimal parameter combination. Meanwhile, the ROC curve is introduced to experimentally compare the effects of the CART decision tree and random forest models. The results show that the random forest model performs better and can accurately analyze the influencing factors of personality. Finally, the importance of feature variables is ranked, providing a certain reference for teachers' teaching.

Keywords

Personality, Random Forest, Decision Tree, Socializing

1. 引言

具备完善的个性是适应社会的必要条件，在教师眼里，个性内向学生大多省心省力。然而，教师也因此对个性内向学生关注不够，对其个性的完善缺乏指导，可能导致个性优势不能充分发挥，出现不健康的心理问题。为了帮助教师改善学生个性，本文通过分析个性因素数据，将随机森林模型应用到研究中，为教师教学提供一定的参考。

随机森林算法结合了 Bagging 集成学习理论[1]，旨在解决决策树的性能缺陷，是一种将多棵决策树集成的学习方法，近年来，随机森林模型在国内外都有相关的研究。马红迪[2]在研究食品安全风险预警问题时，建立决策树和随机森林预警模型，为了达到食品安全风险指数较为理想的预测效果，引入可以快速实现参数调优的网格搜索法。最后，对比决策树和随机森林模型的结果，改进后的随机森林模型明显更好。曲艳婷[3]在对 P2P 网络借贷违约研究时，使用几种常见的决策树算法，通过混淆矩阵评价各算法模型的性能，再利用 ROC 曲线去选择最佳分类阈值，突出了随机森林模型的优势，整体的预测结果比较理想。孙文轩[4]利用随机森林模型对公司人才流失问题进行研究，将随机森林模型优化后，对属性特征进行重要性排序，结果直观清晰，表明了随机森林模型可以用于人才流失原因的分析。

Yan G 和 Chen X 等[5]利用多元线性回归和随机森林分析东北地区萎缩城市的因素及其效应。Wang Q 等[6]使用随机森林模型，对地表温度的重要性和相互影响进行评估，选取上海中心城区作为研究区域。Carter 等[7]利用广义随机森林研究尼加拉瓜的农村发展，发现不同类型的家庭从该政策中受益的程度存在较大差异。Wang 等人[8]提出选择后提升随机森林的算法，巧妙地将传统的随机森林与 Lasso 方法相结合。Qiu X 和 Wang H 等人[9]探索了男性与女性患者，在卒中后 3 个月时 PSD 的影响因素，并对其重要性进行排序分析。Parkhurst 等[10]在生物信息学领域方面，使用随机森林研究五个海滩的细菌密度与其他变量之间的关系。通过对随机森林的研究现状进行总结归纳，可以判断，随机森林模型在各领域都能够很好地应用，具有很强的适应性。此外随机森林模型处理数据速度快、拟合效果好、能够可视化展示，因此本文采用随机森林模型对个性影响因素进行研究。

2. 主要理论

2.1. 决策树

决策树是一种常见的分类算法，其规则是 if-else 的思想。其分类结果是一个树形结构，其主要组成部分有根节点、子节点、叶子节点。通过树形结构可直观反映各节点属性，位于树最上面的是根节点，

其分支是子节点和最终的叶子节点。首先将分类最好的特征变量作为根节点的划分条件，通过不断的递归分类，最后的分类结果即为最终的预测结果。决策树算法中，由于选择的方法不同，可以分为 ID3、C4.5 以及 CART 算法，算法之间的联系比较紧密。CART 算法最早由 1984 年 Breiman [11]提出，其字段选择指标是基尼系数，基尼指数的取值范围是 0 到 1，具体公式如下。

$$\mathrm{Gini}(A)=\sum_{k=1}^{n}p_k(1-p_k)=1-\sum_{k=1}^{n}p_k^2$$

其中，p_k 表示某样本中第 k 个可能值的概率。基尼指数还可以写为：

$$\mathrm{Gini}(A)=1-\sum_{k=1}^{n}\left(\frac{|C_k|}{|A|}\right)^2$$

其中，$|A|$ 是指所有的样本，$|C_k|$ 是事件中第 k 个值出现的次数，$\frac{|C_k|}{|A|}$ 值即为 p_k。

CART 决策树在分类和回归问题上都能适用，因此选择 CART 决策树，通过计算基尼指数最小值，其对应的特征变量作为节点的划分条件，递归地对当前数据进行划分，直到满足终止条件为止，最终生成一棵 CART 决策树。

2.2. 随机森林

由于决策树算法对样本进行分类时，可能会出现过拟合问题。而随机森林算法属于集成学习，该算法的核心是决策树的投票机制，从单棵决策树变成多棵决策树的组合预测，降低产生过拟合的可能性。随机森林的随机性表现在两个方面：一方面是训练集随机抽取，另一方面是特征随机抽取。对于分类问题，将多数决策树的结果，作为最终分类结果。对于回归问题，取多棵决策树的平均值，作为预测的最终结果。

3. 数据来源和变量描述

3.1. 数据来源

本文选择 Kaggle 网站资料，查看数据集，该数据集中含有 2900 个数据，包含独处时间、外出次数、朋友圈人数等 8 个特征变量。其中，因变量为 Personality，使用 replace 函数，将个性变量由离散型数据转换为数值型数据，其中个性外向用 1 表示，内向用 0 表示，转换后的各变量含义和变量类型如表 1 所示。

Table 1. Data introduction
表 1. 数据介绍

变量名称	变量含义	变量类型
Stage_fear	是否存在怯场	离散型
Drained_after_socializing	社交后是否感到疲惫	离散型
Social_event_attendance	参加社交活动次数	数值型
Going_outside	外出次数	数值型
Drained_after_socializing	社交后是否感到疲惫	数值型
Friends_circle_size	朋友圈人数	数值型
Post_frequency	社交媒体更新次数	数值型
Personality	个性	数值型

3.2. 变量说明

由于数据量比较多，可能会出现缺失的情况，本文数据比较齐全，无需删除缺失值。由于数据集中因变量位置不在第一列，为了方便后续的数据处理，通过 drop 和 insert 函数，将因变量个性这一列转换到第一列。当数据集中存在离散型变量时，通常对其进行重编码，将离散型特征变量转换成数值型。本文选择的数据处理方法是哑变量处理。本文对是否存在怯场(Stage_fear)、社交后是否感到疲惫(Drained_after_socializing)两个特征变量进行哑变量处理，需要用到的函数有 get_dummies，变量一般取值 0 或 1。通过哑变量处理，Stage_fear 被分为 Stage_fear_Yes 和 Stage_fear_No。Drained_after_socializing 分为 Drained_after_socializing_Yes 和 Drained_after_socializing_No。经过哑变量处理后，数据集由 8 个特征变量扩充为 10 个特征变量。通过上述处理，提高了数据的质量，便于后期模型的构建。

4. 实证分析

4.1. 模型的构建

在构建 CART 决策树和随机森林模型的过程中，将训练集和测试集按照 7:3 的比例进行划分，同时通过网格搜索的途径，全面搜索整个参数空间，从而获取各参数组合下的最优值，其中需要引入 GridSearchCV 函数。下面是对随机森林模型中所需要的一些参数进行具体介绍，参数含义如表 2 所示。

Table 2. Parameter meaning
表 2. 参数含义

参数	具体含义
n_estimators	随机森林中决策树个数
max_depth	随机森林中每棵树的最大深度
min_samples_split	随机森林中根节点或子节点能够分割的最小样本数
min_samples_leaf	随机森林中叶子节点最小样本数
max_features	最大特征数量

本文对 CART 决策树模型和随机森林模型进行分析，其中 criterion 参数采用默认设置为“gini”，在对 CART 决策树和随机森林模型参数优化过程中，在 CART 决策树模型中分别设置最大深度、能够分割的最小样本数、叶子节点最小样本数的数值。经过网格搜索，CART 决策树模型的最佳参数组合为 9、12、10。在随机森林模型中分别设置最大深度、最大特征数量、叶子节点最小样本数、能够继续分割的最小样本数、决策树个数，经过网格搜索，随机森林模型最佳参数组合为 2、2、2、2、50。在建立 CART 决策树和随机森林模型后，对模型进行预测，计算两种模型的准确率、精确率、召回率和 F1 分数，结果如表 3 所示。

Table 3. Model comparison results
表 3. 模型对比结果

模型	Accuracy	Precision	Recall	F1-score
CART 决策树	0.9402	0.9543	0.9289	0.9414
随机森林	0.9414	0.9544	0.9311	0.9426

其中 Accuracy 是准确率，代表预测正确的比例。Precision 是精确率，代表预测为正例的个数中，实际结果也为正例的比例。Recall 是召回率，代表实际结果为正例的个数中，模型预测正确的比例。F1-score 反映了模型的稳健性，是综合精确率和召回率的指标。从结果来看，随机森林模型的准确率、精确率、召回率、F1 分数均高于 CART 决策树模型，即随机森林模型更好，取得预期的效果。

4.2. 模型评估

评估分类模型的方式有很多，常见的方法有混淆矩阵、ROC 曲线、K-S 值，本文引入 ROC 曲线对模型进行评估。ROC 曲线由真正例率(TPR)、假正例率(FPR)构成。在二分类问题中，每个预测有四个不同的结果，如表 4 所示。

Table 4. Confusion matrix
表 4. 混淆矩阵

	正例	负例
正例	真正例(TP)	假负例(FN)
负例	假正例(FP)	真负例(TN)

混淆矩阵中每一行代表真实所属的类别，每一列代表预测所属类别。真正例(TP)代表实际结果为正，预测结果也为正的个数。假负例(FN)代表实际结果为正，预测结果为负的个数。假正例(FP)代表实际结果为负，预测结果为正的个数。真负例(TN)代表实际结果为负，预测结果也为负的个数。其中真正例率(TPR)和假正例率(FPR)公式如下。

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}$$

ROC 曲线距离左上角越近，表示分类器效果越好。当 AUC 的面积大于 0.8 时，可以认为模型的效果较好。为了进一步验证 CART 决策树和随机森林模型的性能，分别计算 AUC 值，两种模型结果如图 1、图 2 所示。

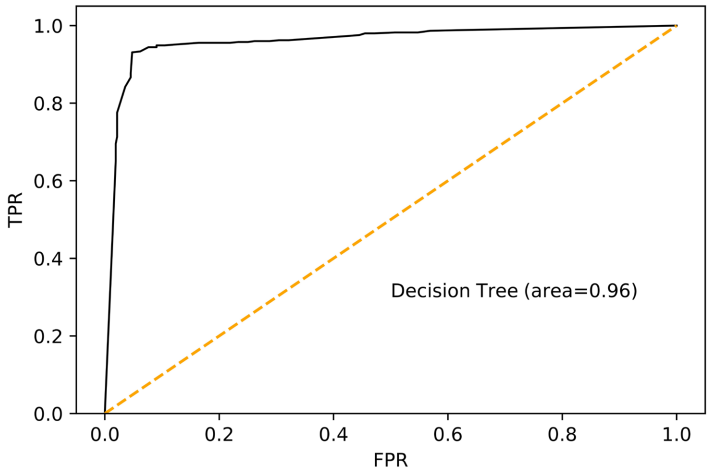

Figure 1. CART decision tree ROC curve
图 1. CART 决策树 ROC 曲线

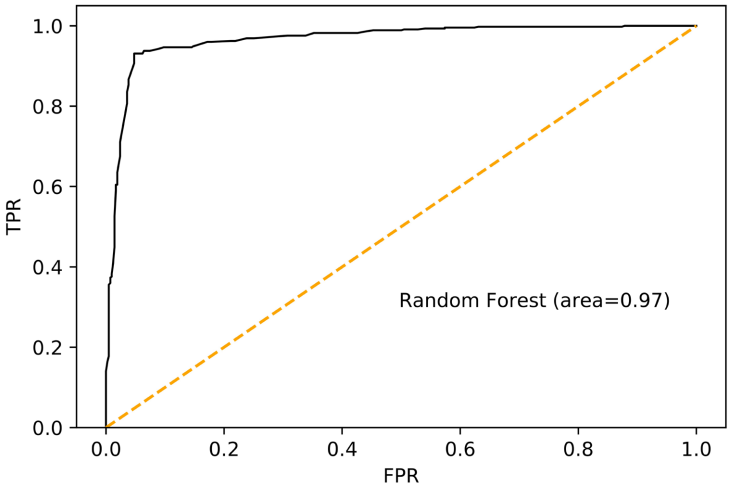

Figure 2. Random forest ROC curve
图 2. 随机森林 ROC 曲线

通过比较分析 CART 决策树和随机森林的模型，CART 决策树模型的 AUC 值为 0.96，随机森林模型的 AUC 值为 0.97，随机森林模型比 CART 决策树模型高出 0.01。即随机森林模型性能比 CART 决策树模型更好，模型拟合达到了预期的效果。接下来选择随机森林模型，计算各个类别下的精确率、召回率等指标，进一步了解各个类别下的预测效果，其中 support 代表各类别的实际个数，avg/total 是各个指标的综合水平。0 类代表个性内向，1 类代表个性外向，结果如表 5 所示。

Table 5. Test results of category 0~1
表 5. 0~1 类测试结果

类别	precision	recall	f1-score	support
0	0.93	0.95	0.94	420
1	0.95	0.93	0.94	450
avg/total	0.94	0.94	0.94	870

根据模型结果，0 类和 1 类的各项指标均较高，预测效果较好。未来可以考虑增加样本个数、结合其他分类模型、改进模型等方法，提高模型性能。

4.3. 特征重要性分析

为了了解哪些因素是比较重要的，通过随机森林模型的 feature_importances 方法对特征进行重要性排序，结果如图 3 所示。

从模型的特征重要性排序结果可知，对个性影响最大的是社交后是否感到疲惫、是否存在怯场等。据此本文认为社交是影响个性的主要因素，其余特征变量影响力较小。

根据结果，为教师教学提供建议。在日常教学中。教师帮助内向学生社交、改善性格。首先是给足安全感、提供低压力参与机会。在课堂分组时，让内向学生先和熟悉的同学搭档，减少陌生环境的紧张感。任务分配明确，不要求主导。其次用小步反馈建立自信，而非强迫改变性格。让内向学生意识到温和的互动也是社交，参加各种有意义的活动，积累社交正体验。最后指导内向学生学习社交能量管理，帮助他们识别并参与高质量、低消耗的社交活动，并重视社交后的独处恢复时间。

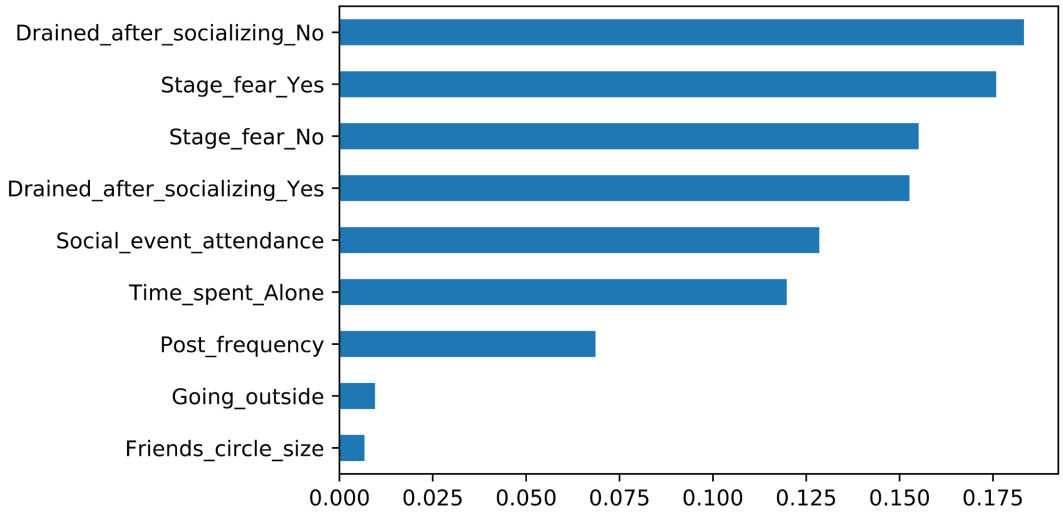

Figure 3. Ranking of feature importance
图 3. 特征重要性排序

5. 结果与讨论

本文探讨影响个性的各种因素，以期为教师改善学生的性格、促进其全面发展提供参考。首先构建 CART 决策树和随机森林模型，以预测个性的内向或外向。通过计算 CART 决策树和随机森林模型的准确率、召回率等指标，引入 ROC 曲线评估模型，对比得出随机森林模型的效果更好。之后引入随机森林模型下各类别的准确率、召回率等指标。最后根据重要性程度，对教师教学提供建议。研究也存在一些不足之处，如数据的代表性等。今后可以进一步完善随机森林模型，删除一些重要性程度低的特征、扩大参数的取值范围、引入其他分类模型等，进一步提高模型性能。

参考文献

[1] Kwok, S.W. and Carter, C. (1990) Multiple Decision Trees. *Machine Intelligence and Pattern Recognition*, **9**, 327-335.

[2] 马红迪. 基于决策树和随机森林模型的食品安全风险预警[D]: [硕士学位论文]. 大连: 东北财经大学, 2020.

[3] 曲艳婷. P2P 网络借贷违约的随机森林预测模型[D]: [硕士学位论文]. 重庆: 重庆大学, 2018.

[4] 孙文轩. 基于随机森林算法的公司人才流失问题与对策研究[D]: [硕士学位论文]. 长春: 吉林大学, 2021.

[5] Yan, G., Chen, X. and Zhang, Y. (2021) Study on the Distribution Pattern and Influencing Factors of Shrinking Cities in Northeast China Based on the Random Forest Model. *Journal of Geography and Cartography*, **3**, 41-51. https://doi.org/10.24294/jgc.v3i1.1305

[6] Wang, Q., Wang, X., Zhou, Y., Liu, D. and Wang, H. (2022) The Dominant Factors and Influence of Urban Characteristics on Land Surface Temperature Using Random Forest Algorithm. *Sustainable Cities and Society*, **79**, Article ID: 103722. https://doi.org/10.1016/j.scs.2022.103722

[7] Carter, M.R., Tjernström, E. and Toledo, P. (2019) Heterogeneous Impact Dynamics of a Rural Business Development Program in Nicaragua. *Journal of Development Economics*, **138**, 77-98. https://doi.org/10.1016/j.jdeveco.2018.11.006

[8] Wang, H. and Wang, G. (2020) Improving Random Forest Algorithm by Lasso Method. *Journal of Statistical Computation and Simulation*, **91**, 353-367. https://doi.org/10.1080/00949655.2020.1814776

[9] Qiu, X., Wang, H., Lan, Y., Miao, J., Pan, C., Sun, W., *et al.* (2022) Explore the Influencing Factors and Construct Random Forest Models of Post-Stroke Depression at 3 Months in Males and Females. *BMC Psychiatry*, **22**, Article No. 811. https://doi.org/10.1186/s12888-022-04467-0

[10] Parkhurst, D.F., Brenner, K.P., Dufour, A.P. and Wymer, L.J. (2005) Indicator Bacteria at Five Swimming Beaches—Analysis Using Random Forests. *Water Research*, **39**, 1354-1360. https://doi.org/10.1016/j.watres.2005.01.001

[11] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32. https://doi.org/10.1023/a:1010933404324