## 基于机器学习的影响科技人员心理健康的重要预测因素

###### 一、研究的目的、意义

1. **研究目的：**筛选出影响科技人员心理健康的重要预测因素
2. **研究意义：**由于各种调查问卷涉及题目过多，需要花费太多时间去填写，导致体验比较差，为了节省科技人员填写时间，同时达到心理健康预测效果，采用机器学习方法简化调查问卷
3. **研究综述**

**3.1 当前科技人员心理健康研究现状**

当今科技的快速发展和竞争对科技人员的心理健康产生了诸多的挑战，近年来已进行了许多关于科技人员心理健康的研究，以了解这一群人的心理状况和应对政策。郭菲和陈祉妍(2019)研究发现中国科技工作者中约超过1/5的科技工作者有不同程度的抑郁问题，一半的人有一定的焦虑问题，8%以上存在中、重度焦虑表现。其主要原因：一方面，科技行业的工作要求高，工作中社会支持不足，没有价值感，得不到尊重，工作压力大等因素均会对心理健康造成损害(Yang T et al., 2016; Mark G & Smith A P., 2012; Siegrist J & Li J, 2016; 刘平平 等, 2012)。而个体如果长期处于工作压力状态下就会出现一种身心消耗过度、精力衰竭的综合症状，包括情绪衰竭、工作怠慢、个人工作效能感低三个维度。这就是美国临床心理学家Freudenberger和, H.J (1974)在研究职业压力时提出的“职业倦怠”(burnout)。研究表明，职业倦怠有可能带来较为严重的心理健康后果，职业倦怠程度较高的个体，其工作效率更低，工作绩效更差，离职意愿更强，且身心焦虑和抑郁程度也会提高(黄浪等, 2021; 赵延东等, 2012; Singh, P. et al., 2012)。另外一些研究表明工作满意度跟个体心理健康有着密切的关系，工作满意度跟个体的职场倦怠感，自尊感，抑郁和焦虑感有着强相关性(Faragher EB et al., 2005; Karim Babayi Nadinloyi et al., 2013)。另一方面，除工作之外，科技工作者也有自己的家庭角色，面临着一些现实生活问题，以往研究发现收入水平，工作家庭冲突也是影响人们心理健康的重要因素(Jenkins R et al., 2008; Gresenz C R et al., 2001; Kelly D. Chandler, 2021; Frone, M. R, 2000)。除了以上这些方面的研究外，也有通过正性与负性情绪量表(Hughes et al., 2009)，心理健康服务资源和使用的主观障碍等方面研究心理健康的影响(郭菲 等, 2020)。目前大部分的文献是通过简单的统计检验，相关性分析和回归分析方法对科技人员心理健康问题分析，这些方法大多是检验分布的一致性，只能验证影响因素与心理问题之间是否显著相关，只考虑了单个风险因素（或一小组风险因素），而不能识别出哪些因素对科技人员心理健康影响比较大。另外从研究广度来看，大部分研究都是针对某一个特定区域的样本或者针对某一个特点的角度进行分析，很少有针对全国样本的全视角分析研究。本研究将以2017年全国科技工作者心理健康问卷调查数据为数据源，采用机器学习中的随机森林方法分析出潜在影响科技人员心理建重要的因素。

**3.2 机器学习在心理学领域的研究**

机器学习涉及使用高级统计和概率技术来构建具有自动从数据中学习的能力的系统。使用机器学习方法可以使数据中的模式更容易、更准确地识别，并可以从数据源做出更准确的预测（例如更准确的诊断和预后）(Jordan &Mitchell, 2015)。与传统统计模型不同，机器学习模型不受假设的约束，特别适用于在复杂数据集中找到模式(Orrù,G et al., 2020)。目前机器学习已用于网络搜索、垃圾邮件过滤器、推荐系统、广告投放、信用评分、欺诈检测、股票交易、药品设计和许多其他应用中(Domingos, 2012)。而在心理健康领域的应用也很多，如使用机器学习随机森林方法(Breiman, L.,2001)从大量潜在的预测变量中识别出重要的预测因素。目前随机森林已经成功地应用于遗传流行病学和一般微生物学的各种问题。在很短的时间内，随机森林已经成为一种主要的数据分析工具，与许多标准方法相比表现良好(Díaz-Uriarte R & Alvarez de Andrés S., 2006)。Halim Abbas等人(2018)提出通过机器学习和特征选择方法，简化出一个新的测评工具，用来检测儿童是否患有孤独症，选取出162个样本数据进行验证，验证结果发现测评时间仅为原来的 1/4，并且提高了准确率。马林梓和李波(2021)利用机器学习方法帅选出对初中生心理健康产生重要影响的因素。S. Fodeh 等人(2019) 通过从twitter上收集大量关于自杀的推文，利用机器学习方法提取出相关特征，并将这些特征数据用机器学习算法进行训练和测试，使其可以用来有效地检测自杀风险因素。T. A. Prout 等人(2020)使用随机森林机器学习算法来确定COVID-19期间心理压力最强的预测因子，并开发回归树来确定焦虑、抑郁和创伤后压力风险更高的个体。

###### **研究方法**

1. **数据来源**

研究数据来源于2017年全国科技工作者心理状况问卷调查，调查问卷包括：匹兹堡睡眠质量指数、工作倦怠量表、工作家庭冲突量表、幸福感与工作满意度、心理健康素养问卷、心理求助障碍量表、易激惹、正性与负性情绪量表、工作时间的分配（自编）、工作满意度（自编）、收入水平（自编）、生活困扰（自编）、社会资本（自编）、职业价值观（自编）、职业评价（自编）、行政干预（自编）、职业压力指标问卷、背景信息、流调中心抑郁自评量表、广泛性焦虑筛查量表。其中流调中心抑郁自评量表、广泛性焦虑筛查量表和自杀倾向的次数将用来作为衡量科技人员心理健康的目标值，其他问卷数据作为科技人员心理健康的预测数据。

1. **数据预处理（使用python下的panda和numpy数据分析包处理）**

* 删除“流调中心抑郁自评量表”和“广泛性焦虑筛查量表”问卷含有题目未作答的样本以及自杀倾向未作答的样本
* 删除做题时间少于20分钟的样本。
* 对样本“流调中心抑郁自评量表”和“广泛性焦虑筛查量表”问卷分数求和，根据求和值的大小来进行心理健康分类，抑郁自评量表的得分以总分10 分为界来判断有无抑郁倾向，以 17 分为界来判断有无抑郁高风险，对于求和值低于10的用0表示心理健康，对于高于10分小于17分的用1表示有抑郁倾向，高于17分用2表示抑郁高风险，添加一列用来存放计算后的分类值。对于焦虑筛查量表则以得分低于5分代表无焦虑，5~9分代表轻度焦虑问题，10~14分代表中度焦虑问题，15 分以上代表重度焦虑问题，根据分数划分，用0表示低于5分无焦虑，1表示5~9分轻度焦虑，2表示10~14分中度焦虑，3表示15分以上重度焦虑，另外再添加一列用来存放该分类值。从数据中删除流调中心抑郁自评量表”和“广泛性焦虑筛查量表”所有题目对应列。
* 删除所有问卷题目中超过20%样本未作答对应列以及超过20%题未作答的样本数据。
* 删除含中文值的无意义数据。
* 删除重复列。
* 采用Scikit-learn中的”多变量插补方法”(IterativeImputer)(Pedregosa , F. et al., 2011)对缺失值进行填充。多变量插补法以迭代循环的方式进行：

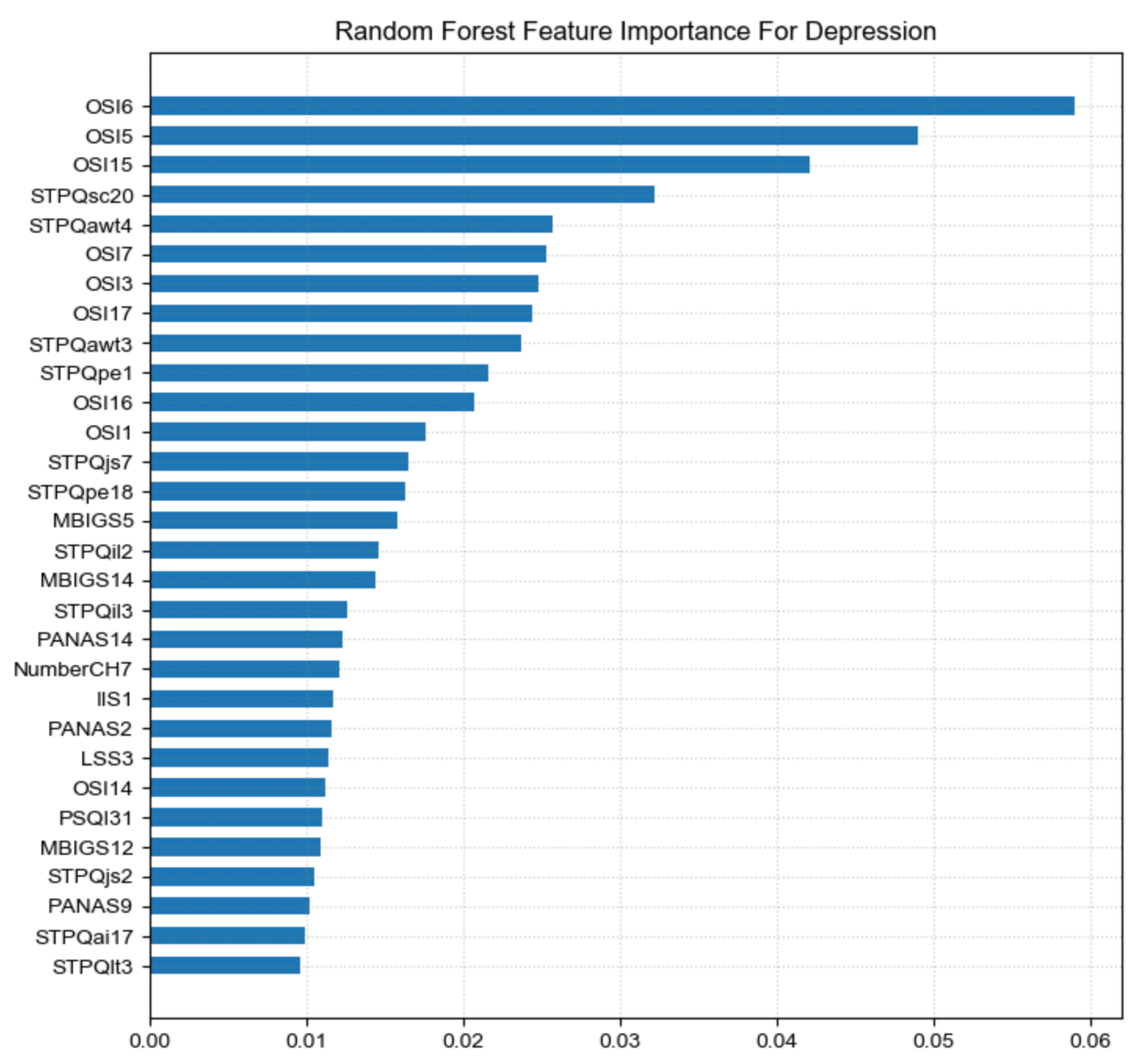
1. 在每个步骤中，将特征列指定为输出y，将其他特征列视为输入X。
2. 把回归器拟合到已知的y的(X，y)上。
3. 然后，使用回归器预测y的缺失值。
4. 针对每个特征以迭代方式完成此操作，然后在max\_iter插补回合中重复此操作。
5. 返回最后一轮估算的结果。
6. **算法选择**

选择“流调中心抑郁自评量表”和“广泛性焦虑筛查量表”得分分类作为预测目标值时，预测目标为多分类离散型数据类型，所以选择机器学习下随机森林分类算法(RandomForestClassifier)，而自杀倾向次数为连续性数据类型，故选用随机森林回归算法(RandomForestRegressor)。这两种算法都属于随机森林算法，都是通过组合多个决策树来进行分类或预测，并通过递归测试每个特征值的强度,预测哪些特征变量对模型的贡献最大。随机森林算法抗拟合能力强，不容易受数据集中噪声和异常值的影响，这能提高算法的稳定性。另外随机森林算法能捕捉特征之间的非线性关系，包括特征之间的交互效应，这对于处理大量特征的多分类问题尤为重要，因为大量特征之间的关系往往是复杂的和非线性的。总之，随机森林算法在处理大量特征的多分类问题时具有较好的性能，能提高预测的准确性和可靠性。

1. **训练模型，并选择出对科技工作者心理健康有重要影响的因素**

特征选择分别以抑郁量表得分、焦虑量表得分和自杀倾向次数为目标值对心理健康分类，其他数据为特征值。将数据随机分为两部分，70%的作为模型训练集，另外30%的数据作为测试集。采用网格搜索(GridSearchCV)从给定的一系列参数中选出最优的模型参数，然后使用交叉验证方法对模型进行验证。一下是针对不同目标值选出的前30个重要特征：

**图1. 以抑郁量表得分分类为目标选出的重要特征**



**图2. 以焦虑量表得分分类为目标选出的重要特征**

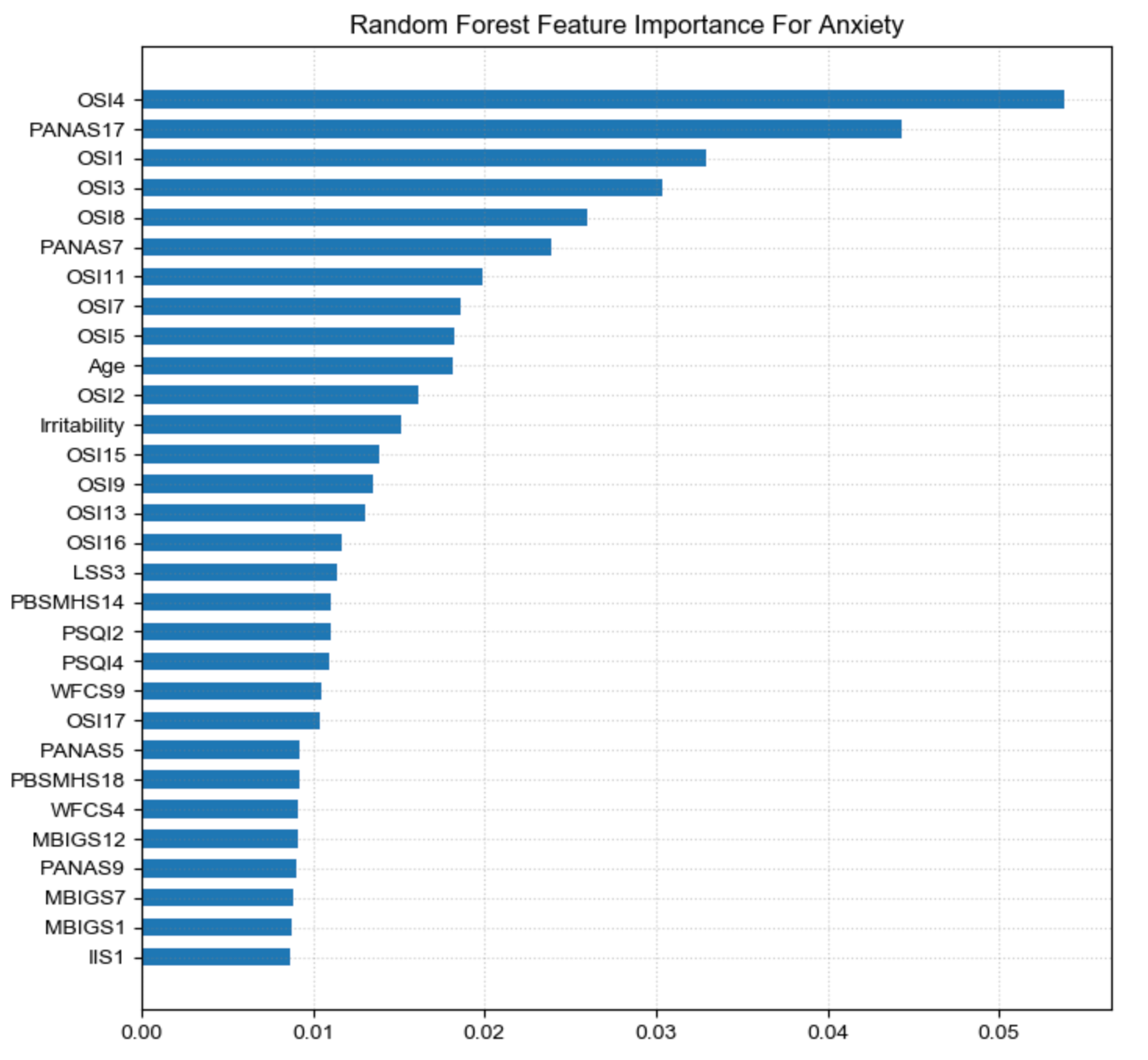
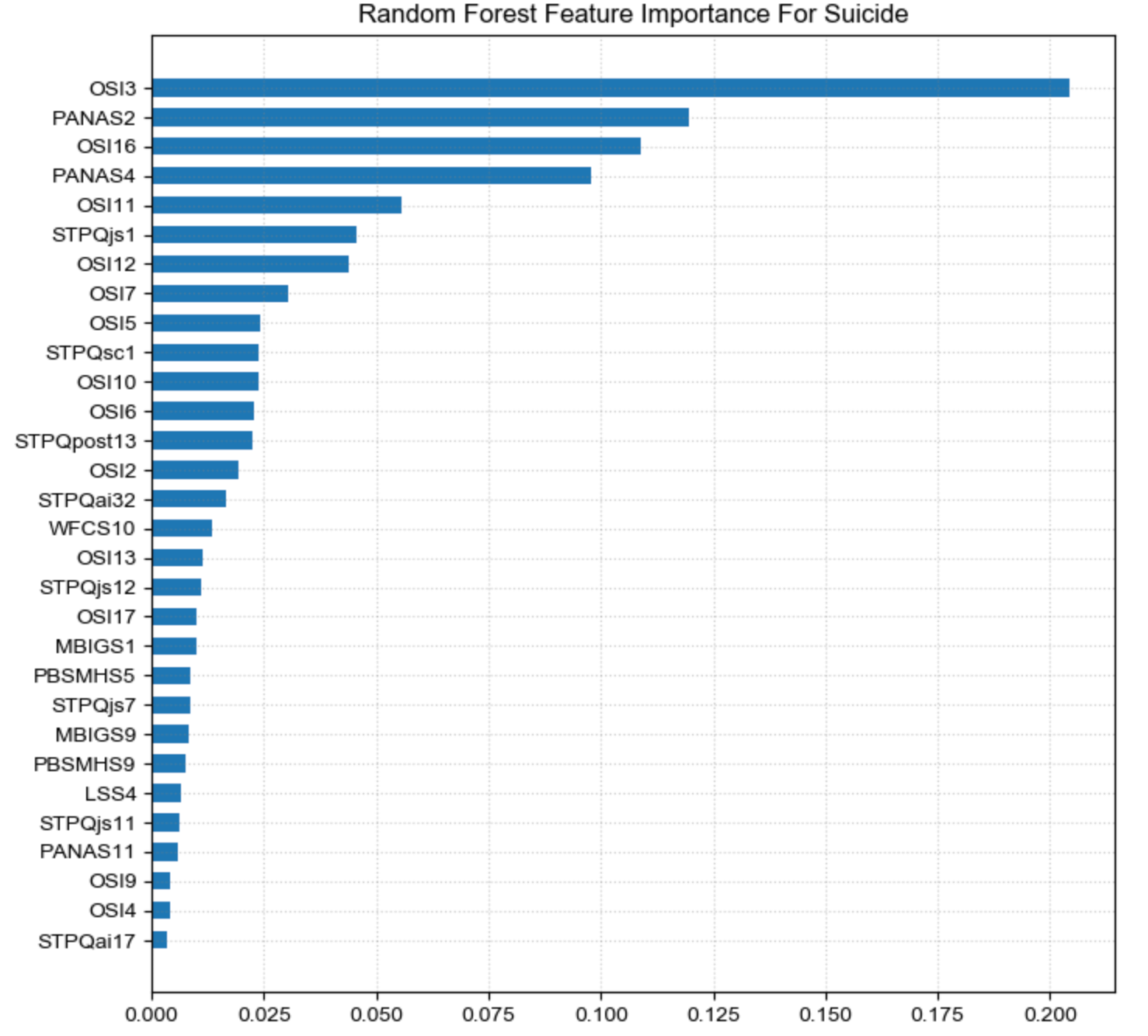
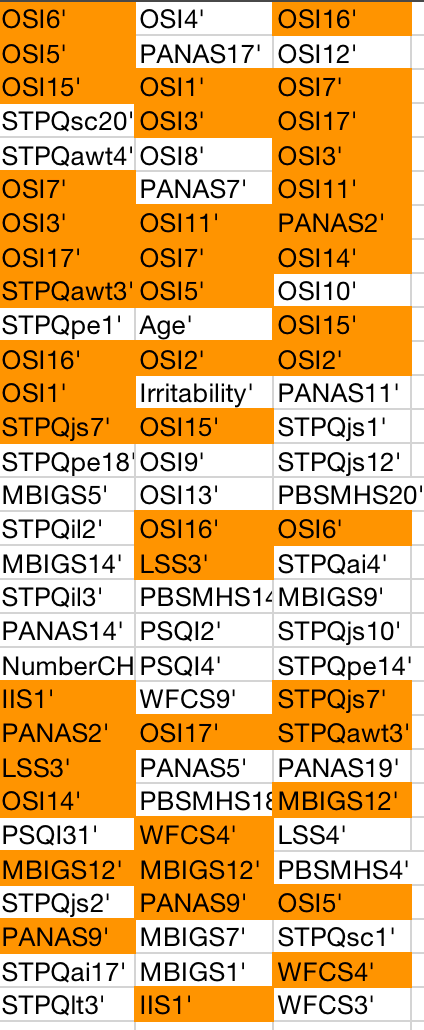


图3. **以自杀倾向得分分类为目标选出的重要特征**



从得出的重要影响因素可以看出对抑郁状态做出重要预测的因素同时也对焦虑状态的预测和自杀倾向预测也有着重要作用，可以选取出共同因素和对各自有着重要预测作用的因素组合成新的测试题目组来重新预测科技工作者的心理健康状态。

图4. 标黄的为共同预测重要因素



1. **使用选择出的重要特征重新训练模型，并采用混淆矩阵评估模型性能**

从数据源中取出选出的重要特征数据，分别对科技工作者的抑郁状态，焦虑状态和自杀倾向作出预测，并使用混淆矩阵检验随机森林分类模型的性能，使用均方根误差（RMSE）来评估随机森林回归模型的性能，使用散点图来可视化自杀倾向预测结果。

图5. 科技人员焦虑状态预测

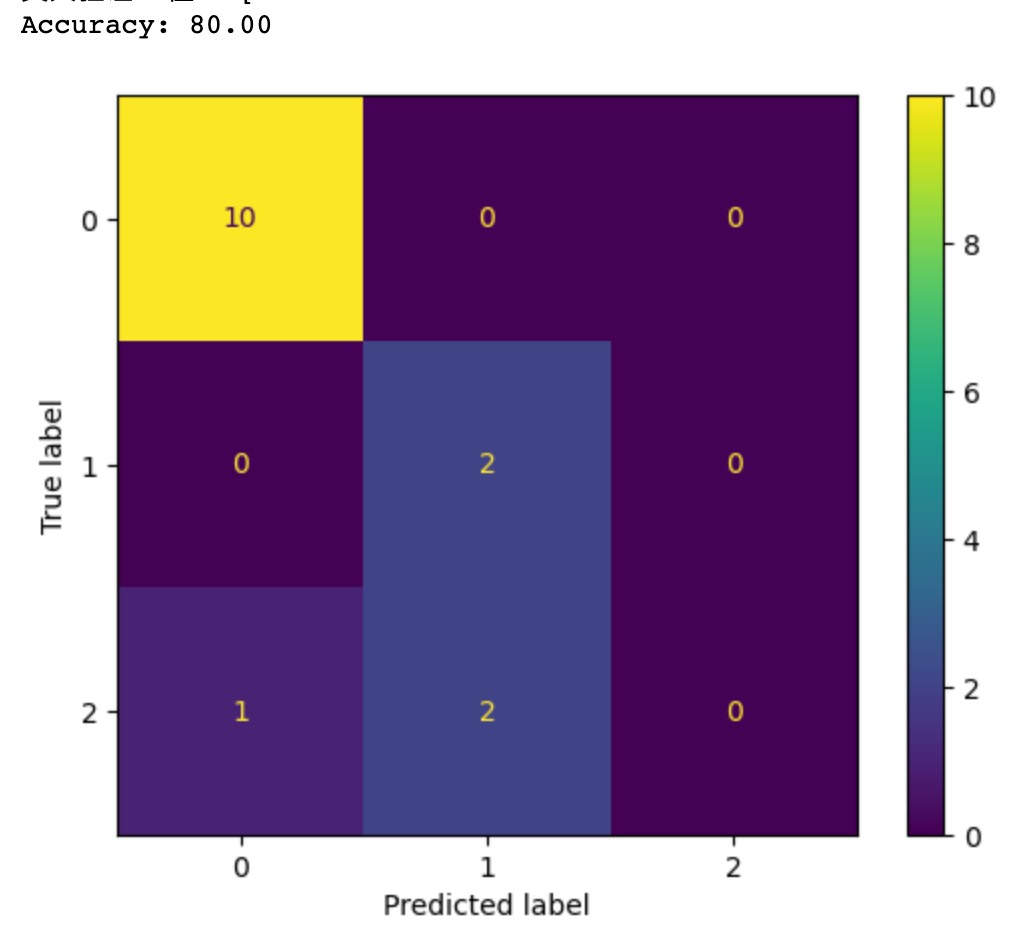


图6. 科技人员抑郁状态预测

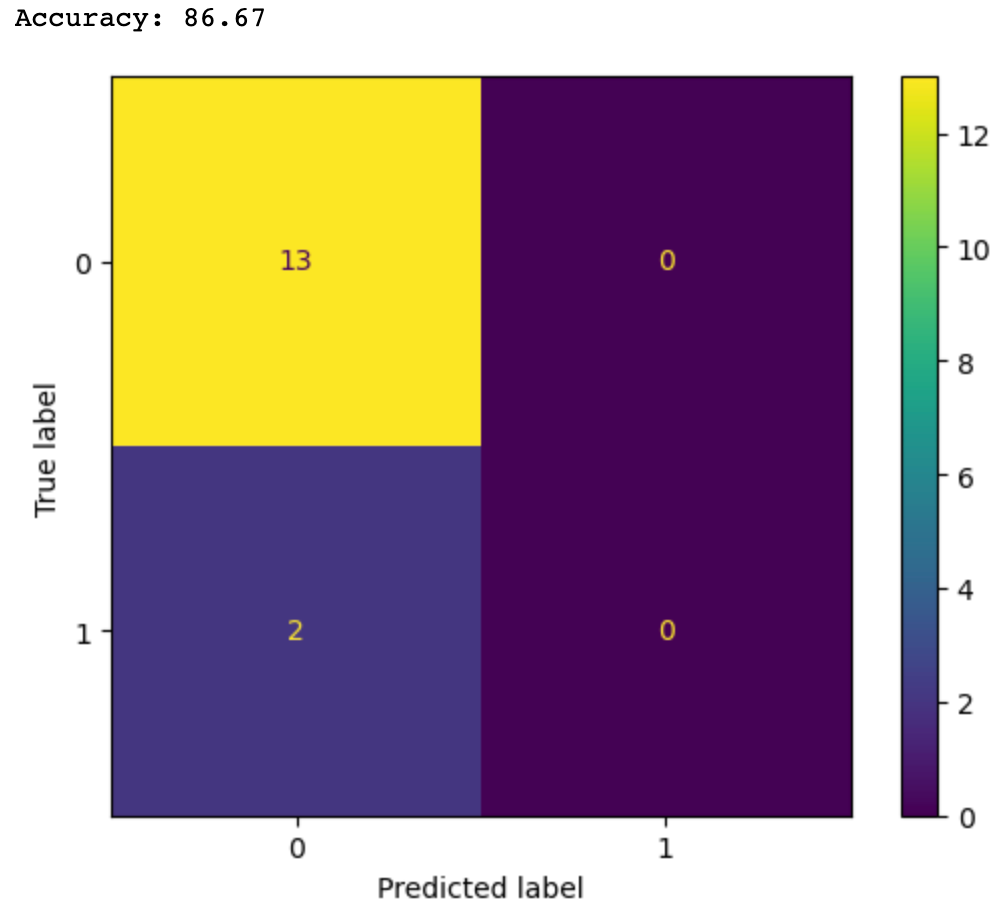
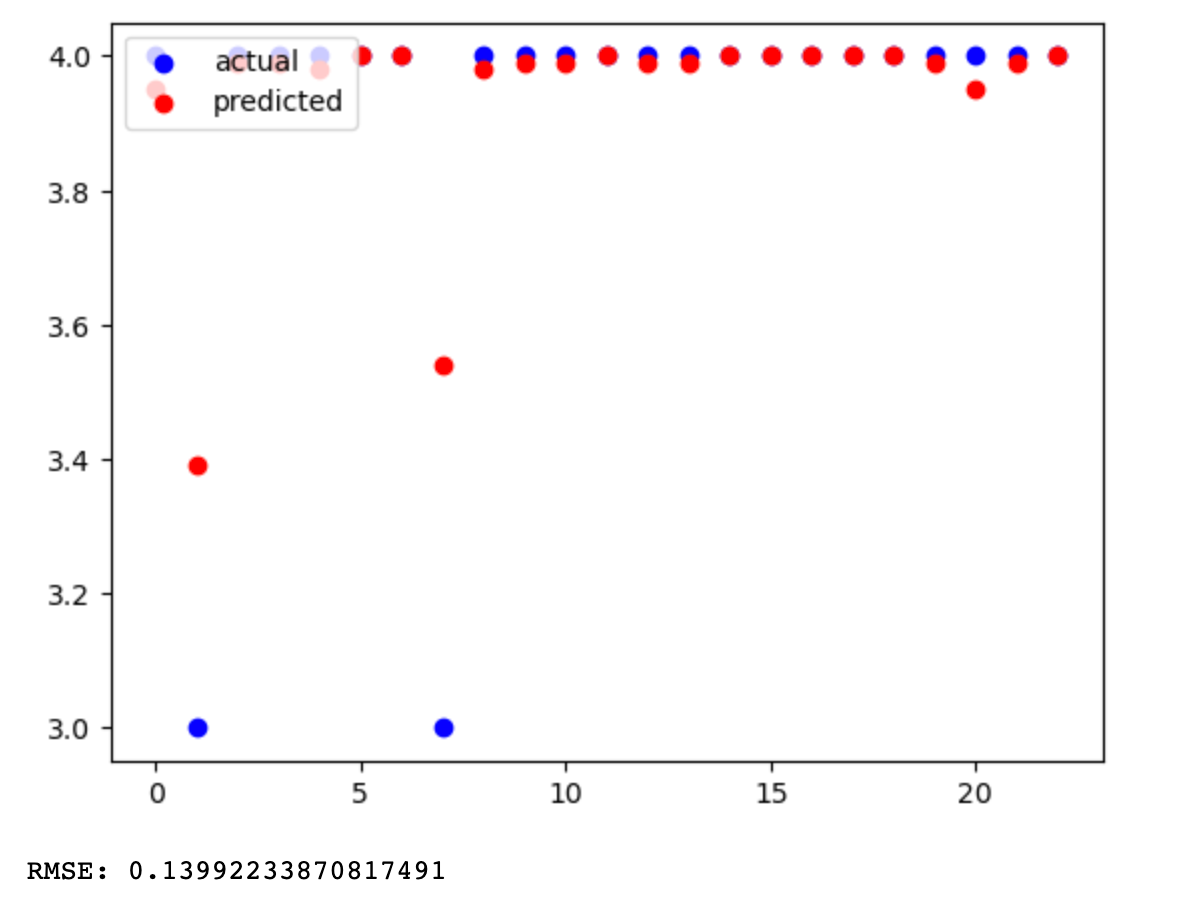


图7. 科技人员自杀倾向预测



参考文献

[1] 郭菲，陈祉妍. (2019). 科技工作者心理健康需求与服务现状 [J]. *科技导报*, *37*(11), 18–27.

[2] Yang T, Shen Y M, Zhu M, et al. (2016). Effects of co- worker and supervisor support on job stress and presenteeism in an aging workforce: A structural equation modelling approach[J]. *International journal of environmental research and public health*, *13*(1), 72.

[3] Mark G, Smith A P. (2012). Occupational stress, job characteristics, coping, and the mental health of nurses[J]. *British Journal of Health Psychology, 17*(3), 505–521.

[4] Siegrist J, Li J. (2016). Associations of extrinsic and intrinsic components of work stress with health: A systematic review of evidence on the effort- reward imbalance model [J]. *International Journal of Environmental Research and Public Health*, *13*(4), 432.

[5] 刘平平，王鹏，高峰强，等. (2012). 企业科技工作者工作压力与焦虑的关系[J].*中国健康心理学杂志*, *8*, 1172–1175.

[6] Freudenberger, H.J. (1974). Staff Burn-Out. *Journal of Social Issues, 30*, 159–165.

[7] 黄浪，陈琳，苏艺伟，等. (2021). 广州市1545名医务人员职业紧张、职业倦怠和抑郁症状的关系研究[J]. *职业卫生与应急救援, 39*(2)*,*129-135.

[8] 赵延东，石长慧，徐莹莹，黄造玉. (2020). 科技工作者职业倦怠的变化趋势及其组织环境影响因素分析[J]. *科学与社会, 10*, 62–75.

[9] Singh, P., Suar, D., & Leiter, M. P. (2012). Antecedents, work-related consequences, and buffers of job burnout among Indian software developers. *Journal of Leadership & Organizational Studies, 19*,83–104.

[10] Faragher EB, Cass M, & Cooper CL. (2005). The relationship between job satisfaction and health: a meta-analysis. *Occupational and Environmental Medicine, 62*,105-112.

[11] Karim Babayi Nadinloyi, Hasan Sadeghi, & Nader Hajloo. (2013). Relationship Between Job Satisfaction and Employees Mental Health. *Procedia - Social and Behavioral Sciences, 84*, 293-297

[12] Jenkins R, Bhugra D, Bebbington P, et al. (2008). Debt, income and mental disorder in the general population[J]. *Psychological Medicine, 38*(10), 1485-1493.

[13] Gresenz C R, Sturm R, & Tang L. (2001). Income and mental health: Unraveling community and individual level relationships[J]. *The Journal of Mental Health Policy and Economics, 4*, 197-203.

[14] Kelly D. Chandler. (2021).Work-family conflict is a public health concern. *Public Health in Practice, 2,* 100158.

[15] Frone, M. R. (2000). Work–family conflict and employee psychiatric disorders: The national comorbidity survey. *Journal of Applied Psychology, 85*(6), 888–895.

[16] Hughes, A.A., Kendall, & P.C. (2009). Psychometric Properties of the Positive and Negative Affect Scale for Children (PANAS-C) in Children with Anxiety Disorders. *Child Psychiatry Hum Dev, 40*, 343–352 .

[17] 郭菲，王雅芯，刘亚男，等. (2020). 科技工作者心理健康状况及影响因素[J]. *科技导报, 38*(10)*,* 90–102.

[18] Jordan MI and Mitchell TM. (2015). Machine learning: trends, perspectives, and prospects. *Science 349*, 255–260.

[19] Halim A, Ford G, Eric G, et al. (2018). Machine learning approach for early detection of autism by combining

questionnaire and home video screening[J]*. Journal of the American Medical Informatics Association, 25*(8),1000–1007.

[20] 马林梓，李波. (2021). *初中生心理健康影响因素的数据挖掘与预警机制研究——基于中国教育追踪调查数据*(硕士学位论文). 华中师范大学，山东.

[21] Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the Acm, 55*(10), 78–87.

[22] T. A. Prout, S. Zilcha-Mano, K. Aafjes-van Doorn, V. Békés, I. Christman-Cohen, K. Whistler, et al. (2020). Identifying Predictors of Psychological Distress During COVID-19: A Machine Learning Approach. *Frontiers in Psychology, 11*, 3063.

[23] Orrù, G., Monaro, M., Conversano, C., Gemignani, A., and Sartori, G. (2020). Machine learning in psychometrics and psychological research. *Frontiers in Psychology, 10*, 2970.

[24] S. Fodeh et al. (2019). Using Machine Learning Algorithms to Detect Suicide Risk Factors on Twitter. International Conference on Data Mining Workshops (ICDMW) (pp. 941-948).

[25] Breiman, L.(2001). Random Forests. Machine Learning 45, 5–32 .

[26] Díaz-Uriarte R, Alvarez de Andrés S.(2006). Gene Selection and Classification of Microarray Data Using Random Forest. BMC Bioinformatics 7, 3.

[27] Pedregosa , F. , Varoquaux , G. , Gramfort , A. , Michel , V. , Thirion , B. , Grisel , O. , ... Duchesnay , E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.