Identifying Factors Influencing China Junior High Students' Cognitive Ability through Educational Data Mining: Utilizing LASSO, Random Forest, and XGBoost

YIMING LUO

yluo5754@uni.sydney.edu.au

The University of Sydney, Camperdown NSW 2006, Sydney, Australia

Abstract. The study innovatively applied educational data mining techniques to the China Education Panel Survey, using LASSO regression, Random Forest, and XGBoost algorithms to identify factors influencing students' cognitive ability. Experimental results indicated that the XGBoost and Random Forest algorithms significantly outperformed the baseline model in assessing Chinese junior High Students' cognitive ability. The main factors identified as influencing cognitive ability include parental and student educational expectations, reading habits and reading and math skills, school environment, and myopia. Eventually, The study concludes by emphasizing the central role of parents and schools in the development of student's cognitive ability and the importance of physical health to students.

Keywords: Educational data mining, machine learning, cognitive ability, random forest, LASOO XGBoost, CEPS.

1 Introduction

With the vibrant development of the Internet and information technology, Educational Data Mining (EDM) has attracted considerable attention as an emerging cross-disciplinary research area. Specifically, EDM utilizes various information technologies to scrutinise the Multidimensional data of teachers and students generated by educational processes. including student academic performance[1], patterns of student behavior in learning[2], student social network[3], and finally achieve diverse objectives such as enhancing student outcomes and assessing their teaching programs continually[4]. Further, it offers fresh perspectives and practical approaches for pedagogical reforms[5].

While EDM technology has seen some application, its application in students' cognitive ability has not been extensively explored and researched yet, Especially using big data and information mining technology. By using the China Education Panel Survey (CEPS) and utilizing methodologies such as lasso regression, random forest, and XGBoost, This study explores indices related to the cognitive ability of Chinese junior high school students and represents a fresh approach to analyse cognitive levels using machine learning. Additionally, this study quantified and visualized the significance of features across various models and

ICMEIM 2023, September 08-10, Wuhan, People's Republic of China

DOI 10.4108/eai.8-9-2023.2340189

conducted comprehensive insights into factors influencing cognitive capability with practical recommendations.

The paper is structured as follows: It first presents the concept of EDM, and then briefly introduces the application of EDM, Research on Students' Cognitive Ability, and CEPS. Next, This paper describes the data pre-processing in detail, the data mining models and the model hyper-parameter tuning process. Experimental results explain the factors that have strong correlations with the cognitive ability of Chinese junior high school students. Eventually, it summarises the characteristics of students with high cognitive ability and makes recommendations for future research.

2 Literature review

Traditionally, Cognitive ability has been considered a crucial measure of academic performance, as proven by the research of Yen and his team[6],which assessed that tests of cognitive ability consistently reflect variations in academic achievement.Carroll[7] defined cognitive ability as the ability that involves some kind of cognitive task and perceived it as an instrumental tool to enhance learning and understanding. This perspective is reinforced by the insights from Robertson et al[8]. which assert that cognitive ability influences the development of students at every cognitive level.

One advantage of cognitive ability is the global comparability of such tests. This notion is further supported by the findings of Reynolds et al[9], which shows that different races and ethnic groups have analogous cognitive ability structures. Given these insights, there is strong reason to believe that this study's attempt to identify factors potentially related to students' cognitive ability through data mining holds universal significance. Given these insights, this study's attempt to identify factors potentially related to students' cognitive ability through data mining holds universal significance and global relevance.

Currently, EDM is employed in predicting students' academic performance and other areas. For instance, researchers like Qwaider have developed an artificial neural network (ANN) model to predict students' academic performance and establish an academic monitoring and early warning model[10]. Similarly, Li and Liu have utilized computational methods to investigate the effects of adolescents' attendance at remedial classes on their emotional health[11]. Notably, Jin and team applied machine learning techniques In CEPS to identify factors that influence academic grades[12]. Mayilvaganan and team utilized data mining techniques to categorize students' cognitive ability with a smaller dataset[13]. Based on the foundation of educational mining technology, this study seeks to identify factors influencing students' cognitive ability through a comprehensive analysis and gives guidance for improvement suggestions.

3 Methodology

3.1 Data Overview

The China Education Panel Survey (CEPS), conducted by the Renmin University of China, is a national tracking research project that randomly selects from 28 county-level units based on population education levels and resident population ratios, covering approximately 20,000 students.This study draws data from CEPE(2014-2015) and focuses on all 10,279 student's sample in junior high school. The questionnaire covers a wide range of aspects related to the students, including their basic personal information, growth experiences amounting to a total of 311 features.

In addition to the questionnaire, CEPS designed a cognitive ability test for students based on a three-parameter IRT model, which yields a standardized overall score that serves as the target variable for assessing students' cognitive ability in this study and the content of CEPS is used as input variables. Though the 3RT model has been widely accepted in the past, it may still contain potential biases due to regional cultural differences or inherent limitations of the logistic regression algorithm itself.This paper used Python 3 and various packages such as NumPy, Scikit-learn, and Matplotlib for data analysis and visualization.

3.2 Data preprocessing

The data preprocessing significantly influences the generalization performance of a supervised machine learning algorithm[14]. Firstly, We removed perfectly collinear variables and meaningless variables and it was observed that most missing values occur due to further investigation of a minority group or due to privacy concerns of students. After consideration, we decided to drop variables with a blank value of more than 10% to ensure the privacy and reliability of the data. While encoding the questionnaire features, we denoted ordinal variables with '1', and categorical ones with '0', and used '-1' for unknowns. What should be pointed out is that high-cardinality categorical variables need to be merged into fewer categorical or binary variables. For instance, the question about parents' educational expectations of students, which previously consisted of 10 different levels, is reorganized into 4 categories based on educational level. We then eliminated data related to students lacking follow-up or cognitive test scores as they could not be evaluated. Subsequently, all variables are standardized to ensure that they can be considered on the same scale. Finally, We used the KNNImputer[15] from the Python and employed K-means clustering for imputation, and missing values were filled using the average of the two nearest neighbours[16].

3.3 Model design

The specific structure of the model and the process of hyperparameter tuning will be introduced in the model design. We refer to linear regression as the baseline model, which can be used to further improve the generalization capabilities. The MSE of the linear model is 0.42216. Given the abundance of binary variables in educational data, we utilized LASSO to simplify the feature set. In order to further explore non-linear relationships,we selected both the Random Forest and XGBoost models, both of which belong to robust and powerful decision tree model categories that can capture complex data patterns. The evaluation of the model and parameters is optimized using cross-validation and the evaluation metric for

Furthermore, to delve deeper into feature importance within the models, it is essential to elucidate how this significance is measured in the employed algorithm.The feature importance in the Lasso regression is determined by the absolute value of the weights in the equation. Meanwhile, In both the Random Forest and XGBoost algorithms, feature importance is measured differently: while Random Forest evaluates it based on the reduction in the Gini index, which measures how each feature reduces impurity during node splits, XGBoost quantifies it by counting the frequency with which a feature serves as a splitting attribute across all trees.All models is measured by the Mean Squared Error (MSE) of the test set.

3.3.1 LASSO regression

The linear model is considered one of the fundamental predictive models but still has limitations. In contrast, Lasso regression applies regularization by imposing a penalty(hyperparameter alpha) on the linear model parameters. This process effectively shrinks the coefficients of less important features to zero, providing an effective defence against overfitting and improving the model's generalization capability[17]. Additionally, the survey might contain some similar questions. LASSO could mitigate the problems of multicollinearity among variables to some extent, leading to more reasonable prediction results. A grid search was conducted on the penalty hyperparameter **(Figure 1)**, revealing a hyperparameter alpha of 0.04 as the optimal choice.

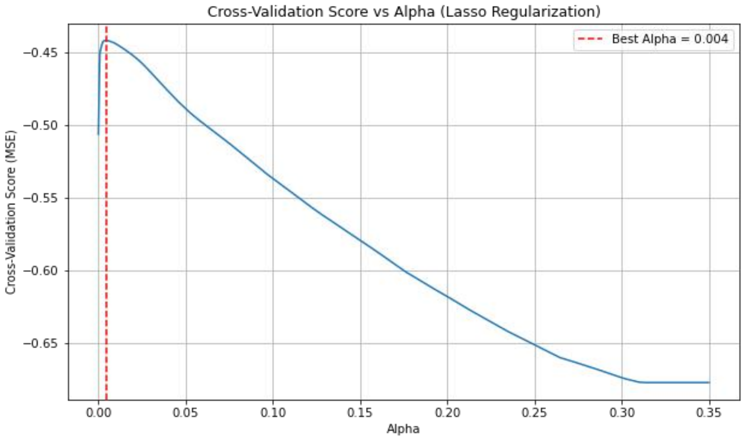

Fig. 1. Grid search on the hyperparameter alpha

3.3.2 Random Forest

The Random Forest algorithm is a representative of the ensemble method within bootstrap aggregation. It combines multiple Classification and Regression Trees (CART) decision trees, each trained on randomly selected data subsets. The composite model results from integrating the outputs of these CART trees through averaging in regression, ensuring excellent generalizability.

We constructed a random forest model based on the scikit-learn framework. Our first step in hyperparameter optimization focused on evaluating the optimal number of trees. As illustrated in **Figure 2**, with other parameters set to their defaults, the performance of the random forest model converges upon reaching a certain threshold, and the MSE metric also stabilizes. Recognizing the importance of more thorough hyperparameter tuning, we utilized the Hyperopt library, a part of Python's vast ecosystem. This library employs the Tree-Based

Parzen Estimators algorithm for automated hyperparameter selection, making it an ideal choice for our needs[18]. The final hyperparameters were determined to be ‘n_estimators’at 325, ‘max_depth’at 32, ‘min_samples_split’at 3, and ‘min_samples_leaf ‘at 4.

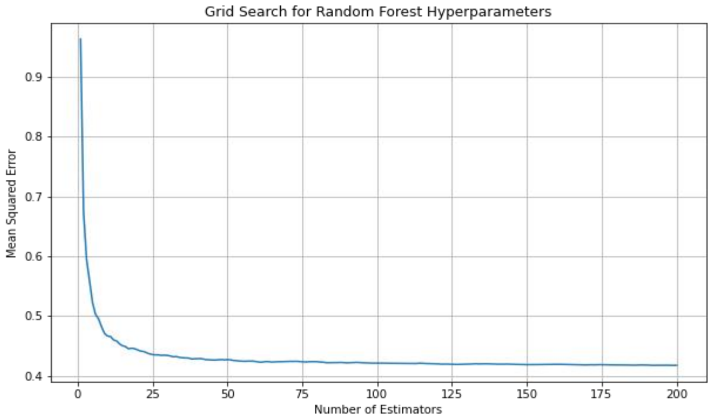

Fig. 2. Grid search on the hyperparameter number of trees

3.3.3 XGBoost

XGBoost is an ensemble algorithm in the Boosting category. It builds upon the Gradient Boosting Decision Tree (GBDT) framework, utilizing the second-order Taylor expansion for loss function optimization and regularization to prevent overfitting. These measures enhance the efficiency and performance of the model. To reduce computational costs, we first fixed a set of defaults to roughly determine the range for each hyperparameter. According to the Figures, the learning rate should be around 0.1(**Figure 3**), and the curve for the number of decisions tends to flatten after the number of trees equal to 12(**Figure 4**). In the end, we also used Hyperopt to finalize the hyperparameters as 'colsample_bytree': 0.59, 'learning_rate': 0.06, 'max_depth': 3.0, 'n_estimators': 770, 'subsample': 0.62.

4 Experimental Results

In order to visually and clearly illustrate the impact of various algorithms on model performance, we used a "Percentage Improvement" metric. This metric was determined by calculating the MSE difference between the baseline model and the other models, which was then expressed in a percentage format relative to the MSE of the baseline model. As depicted in **Figure 5**, it can be concluded that the LASSO regression, compared to the linear model, achieved a performance improvement of 0.4%, reducing the MSE to 0.420478. The Random Forest achieved a 1.58% improvement with an MSE of 0.415501. Most significantly, the XGBoost showed a performance improvement of 8.71% with an MSE of 0.38538, indicating a considerable advantage when dealing with this dataset. Based on the results of feature importance from the three models in **Table 1**, several key factors influence students' cognitive test performance, as follows:

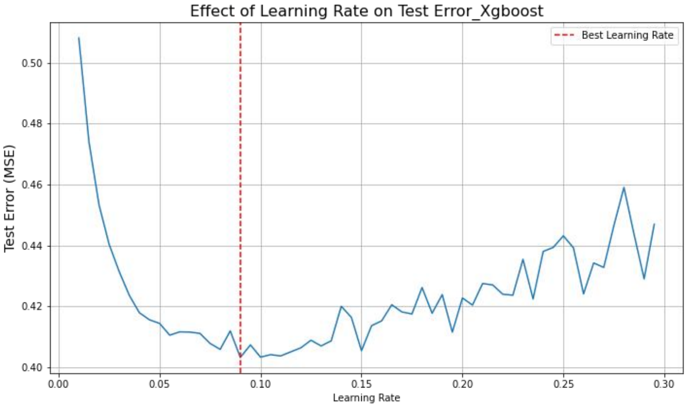

Fig. 3. Grid search on the hyperparameter learning rate(XGBoost)

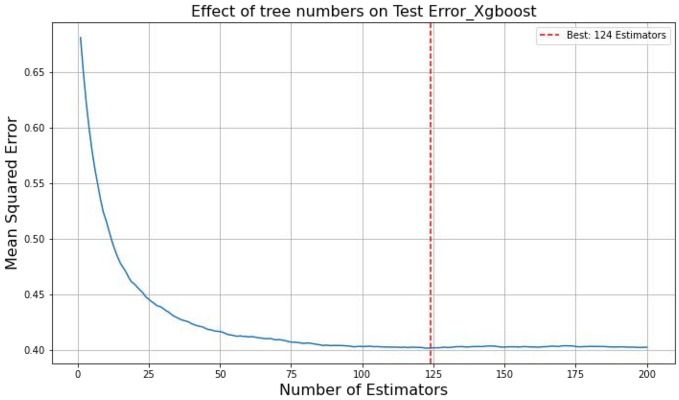

Fig. 4. Grid search on the hyperparameter tree numbers (XGBoost)

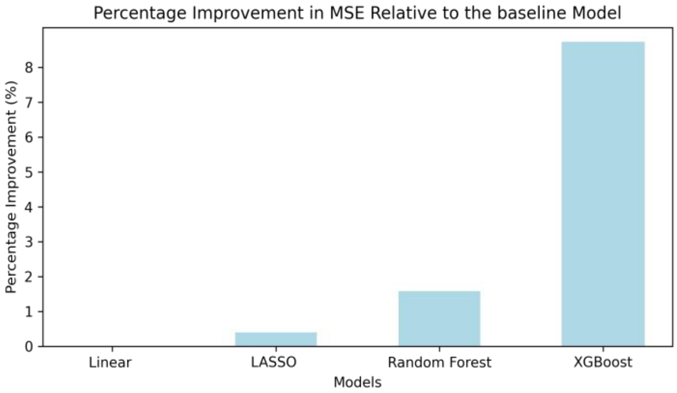

Fig. 5 Results of the percentage improvement in MSE Relative to the baseline model

4.1 Parental and Student Educational Expectations

One major factor, that hugely contributes to cognitive ability and is evident in all three models, is the educational expectations set by both parents and students themselves. Children with high-expectation parents not only receive more resources and support in honing their abilities but also tend to achieve superior academic performances due to the positive influence of parental educational expectations on their cognitive level[19]. Such parental expectations, either directly or indirectly, shape children's views of education and reflect the parents' convictions. These convictions may act as a catalyst for children's logical thinking and problem-solving abilities, inducing intrinsic motivation within them, and consequently leading them to set higher cognitive requirements for themselves.

4.2 Reading and Mathematics

The results from all three algorithms show that having abundant books at home, frequently visiting libraries or museums, and reading in one's free time have a significant feature of importance in influencing cognitive ability. A plausible explanation for this might be that for mentally active and creative youngsters, reading extracurricular books is an essential and effective method to quickly broaden their horizons and efficiently shape their worldview. Additionally, The question "Do You Currently Find Learning Mathematics Difficult?" had a significant feature of importance, which also notably influenced cognitive representation. This suggests that mathematics might be a crucial tool for assessing abstract logical abilities.

4.3 School environment

The results of both the LASSO and XGBoost models highlight the significant influence of a positive school environment on students' cognitive ability, including frequent praise from teachers and openness to diverse campus cultures. This is further supported by neuroscience research. It suggests that certain brain connections, which are called resting-state functional connectivities (rsFC), are correlated with cognition and that an active-minded school environment can positively modulate these connections[20].

4.4 Myopia

A worrying fact revealed by all three models is the significant weight of student myopia in influencing cognitive ability. This indicates certain prices, such as the neglect of eye protection while reading. Another example is that, While access to the Internet has greatly expanded the channels through which children can obtain information, the radiation from the use of electronic devices can also affect their vision.

Table 1. Top 10 Features in three models

Top 10 Features by Absolute Coefficient Value (Lasso)			
Code	Feature Name	Feature Type	Coefficient
w2a27	What Are Your Parents' Expectations Regarding Your Academic Performance?	Ordinal	0.1245220
w2a12	Do You Have a Computer and Internet Connection at Home?	Binary	0.0995724
w2b02	Do You Currently Find Learning Mathematics Difficult?	Binary	0.0833465

w2b18	What Level of Education Would You Like to Achieve?	Ordinal	0.0819200
w2c09	Are You Nearsighted?	Binary	0.0644415
w2a28	What Are Your Parents' Educational Expectations for You?	Ordinal	0.0582280
w2b0610	Regarding School Life, Do You Agree with the Following Statements - I Would Like to Go to Another School	Binary	0.0540431
w2b1405	What Do You Typically Do During Summer and Winter Vacations - Staying at Home	Binary	0.0529204
w2d0102	In the past year, were you able to do the following - Observe orders and queue up voluntarily.	Binary	0.0523765
w2b0603	Regarding School Life, Do You Agree with the Following Statements - Class Teacher Often Praises Me	Binary	0.0517522

Top 10 Feature Importance in Random Forest			
Code	Feature Name	Feature Type	Importance
w2b02	Do You Currently Find Learning Mathematics Difficult?	Binary	0.0869988
w2b18	What Level of Education Would You Like to Achieve?	Ordinal	0.0822471
w2a28	What Are Your Parents' Educational Expectations for You?	Ordinal	0.0477347
w2a27	What Are Your Parents' Expectations Regarding Your Academic Performance?	Ordinal	0.0382041
w2a12	Do You Have a Computer and Internet Connection at Home?	Binary	0.0275858
w2c02	What Is Your Current Weight?	Continuous	0.021011573
w2c01	What Is Your Current Height?	Continuous	0.0176437
w2c09	Are You Nearsighted?	Binary	0.0153071
w2a11	Do You Have Many Books at Home (Excluding Textbooks and Magazines)?	Ordinal	0.0107946
w2a25	How Often Do You Visit Museums, Zoos, Science Museums, etc. with Your Parents?	Ordinal	0.0089556

Top 10 Feature Importance in XGBoost			
Code	Feature Name	Feature Type	Importance
w2b18	What Level of Education Would You Like to Achieve?	Ordinal	0.073198058
w2a28	What Are Your Parents' Educational Expectations for You?	Ordinal	0.046537966
w2b02	Do You Currently Find Learning Mathematics Difficult?	Binary	0.030753771
w2a12	Do You Have a Computer and Internet Connection at Home?	Binary	0.028095279
w2a27	What Are Your Parents' Expectations Regarding Your Academic Performance?	Ordinal	0.024927625
w2c09	Are You Nearsighted?	Binary	0.018021179
w2b1405	What Do You Typically Do During Summer and Winter Vacations - Staying at Home	Binary	0.010315914
w2b1104	What Interests and Hobbies Do You Have - None	Binary	0.01019203
w2d15	Do you think that students from your county (district) in your class would be willing to be friends with students	Binary	0.009892219

	from other counties (districts) or cities?		
w2b1301	Apart from the Interests and Hobbies Mentioned Above, What Other Activities Do You Like to Do - Reading Books	Binary	0.008428613

5 Conclusion

In this study, we applied EDM techniques to the CEPS using LASSO regression, Random Forest, and XGBoost algorithms. The test errors of these models showed a significant improvement compared to the baseline model. Our findings indicate that factors such as the school environment, educational expectations from both parents and students, reading habits, a supportive environment for reading, and myopia strongly influence cognitive ability.

The first is to increase parent-child involvement. Parents play a critical role in influencing students' cognitive ability. Parents should foster reading habits in their children and ensure open communication in a supportive environment. Secondly, Schools should prioritize fostering a positive, and open school environment. It's essential to continuously improve teaching methodologies and curriculum management to enhance student interaction. Furthermore, while cognitive ability is vital, the importance of physical health, especially vision protection, cannot be ignored. Regular vision check-ups are recommended.

In this study, we applied EDM techniques to identify factors influencing students' cognitive ability. While our findings provide valuable insights, the causal relationship between cognitive abilities and these identified factors still needs deeper investigation. To gain multi-dimensional perspectives, it would be beneficial for future research to deploy a wider variety of educational mining techniques, such as association rule analysis or lag sequence analysis.

References

[1] Kaunang, F. J., Rotikan, R. (2018) Students' academic performance prediction using data mining. In 2018 Third International Conference on Informatics and Computing (ICIC). Palembang. pp. 1-5. https://doi.org/10.1109/IAC.2018.8780547

[2] Luckin, R. (2007) Modeling learning patterns of students with a tutoring system using Hidden Markov Models. In 2007 conference on Artificial Intelligence in Education: Building Technology Rich Learning Contexts That Work. Netherlands. pp. 238-245. https://doi.org/10.5555/1563601.1563642

[3] Palazuelos, C., García-Saiz, D., Zorrilla, M. (2013) Social network analysis and data mining: An application to the e-learning context. In Computational Collective Intelligence. Technologies and Applications: 5th International Conference, ICCCI 2013. Berlin. pp. 651-660. https://doi.org/10.1007/978-3-642-40495-5_65

[4] Sheard, J., Ceddia, J., Hurst, J., Tuovinen, J. (2003) Inferring student learning behaviour from website interactions: A usage analysis. Education and Information Technologies, 8: 245-266. https://doi.org/10.1023/A:1026360026073

[5] Zhang, Y., Yun, Y., An, R., Cui, J., Dai, H., Shang, X. (2021) Educational data mining techniques for student performance prediction: method review and comparison analysis. Frontiers in Psychology, 12: 698490. https://doi.org/10.3389/fpsyg.2021.698490

[6] Yen, C.J., Konold, T.R., Mcdermott, P.A. (2004) Does learning behavior augment cognitive ability as an indicator of academic achievement. Journal of School Psychology, 42: 157-169. https://doi.org/10.1016/j.jsp.2003.12.001

[7] Carroll, J.B. (1997) Psychometrics, intelligence, and public perception. Intelligence, 24: 25-52. https://doi.org/10.1016/S0160-2896(97)90012-X

[8] Robertson, K.F., Smeets, S., Lubinski, D., Benbow, C.P. (2010) Beyond the threshold hypothesis: Even among the gifted and top math/science graduate students, cognitive abilities, vocational interests, and lifestyle preferences matter for career choice, performance, and persistence. Current Directions in Psychological Science, 19: 346-351. https://doi.org/10.1177/0963721410391442

[9] Reynolds, C.R., Altmann, R.A., Allen, D.N. (2021) The problem of bias in psychological assessment. In Mastering Modern Psychological Testing: Theory and Methods. Springer International Publishing, Cham. pp. 573-613. https://doi.org/10.1007/978-3-030-59455-8_15

[10] Qwaider, S.R., Abu-Naser, S.S., Zaqout, I.S. (2020) Artificial Neural Network Prediction of the Academic Warning of Students in the Faculty of Engineering and Information Technology in Al-Azhar University-Gaza. Aug Repository. http://dspace.alazhar.edu.ps/xmlui/handle/123456789/634

[11] Li, S., Liu, A. (2023) Does cram school participation bring about negative emotions? Causal inference based on Chinese Education Panel Survey (CEPS) data. Chinese Journal of Sociology, 9: 219-249. https://doi.org/10.1177/2057150X231165145

[12] Jin, Y., Yang, X., Yu, C., Yang, L. (2021) Educational Data Mining: Discovering Principal Factors for Better Academic Performance. In 2021 the 3rd International Conference on Big Data Engineering and Technology (BDET). New York. pp. 1-8. https://doi.org/10.1145/3474944.3474945

[13] Mayilvaganan, M., Kalpanadevi, D. (2015) Cognitive skill analysis for students through problem solving based on data mining techniques. Procedia Computer Science, 47: 62-75. https://doi.org/10.1016/j.procs.2015.03.184

[14] Alexandropoulos, S.A.N., Kotsiantis, S.B., Vrahatis, M.N. (2019) Data preprocessing in predictive data mining. The Knowledge Engineering Review, 34: e1. https://doi.org/10.1017/S026988891800036X

[15] Pedregosa et al. (2011) sklearn.impute.KNNImputer. https://scikit-learn.org/stable/modules/generated/sklearn.impute.KNNImputer.html

[16] Pedregosa et al. (2011) sklearn.examples. https://scikit-learn.org/stable/auto_examples/index.html

[17] Ranstam, J., Cook, J. (2018) LASSO regression. Journal of British Surgery, 105: 1348-1348. https://doi.org/10.1002/bjs.10895

[18] Bergstra, J., Yamins, D., Cox, D.D. (2013) Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms. In Proceedings of the 12th Python in Science Conference. pp. 20. https://doi.org/10.25080/Majora-8b375195-003

[19] Phillipson, S., Phillipson, S.N. (2012) Children’s cognitive ability and their academic achievement: The mediation effects of parental expectations. Asia Pacific Education Review, 13: 495-508. https://doi.org/10.1007/s12564-011-9198-1

[20] Rakesh, D., Seguin, C., Zalesky, A., Cropley, V., Whittle, S. (2021) Associations between neighborhood disadvantage, resting-state functional connectivity, and behavior in the adolescent brain cognitive development study: the moderating role of positive family and school environments. Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 6: 877-886. https://doi.org/10.1016/j.bpsc.2021.03.008