Preliminary Personal Trait Prediction from High School Summer Vacation e-learning Behavior

Kyosuke Takami[1], Brendan Flanagan [1], Rwitajit Majumdar[1] and Hiroaki Ogata [1]

[1] *Academic Center for Computing and Media Studies, Kyoto University, Japan*

Abstract

Previous studies have shown that e-book interaction logs can predict students at-risk of academic failure-students whose academic performance is low. However, in the context of an individualized e-learning system, it is very important to predict personality traits to realize the well optimized and suited assist, intervention and feedback based on personality bases. Here we examine the extent to which individuals' Big Five personality traits can be predicted on the basis of learning log data harvested K-12 e-learning system. Taking a machine-learning approach, we predict conscientiousness (R= 0.38), which is related to academic achievement, based on behavioral data collected from 129 high school students' summer vacation learning log. This result is preliminary but the first step to open the prediction of personality from K-12 learning log.

Keywords

Big Five personality trait, Educational data mining, Learning analytics, Personality trait prediction

1. Introduction

Previous studies have shown that e-book interaction logs can predict students at-risk of academic failure-students whose academic performance is low [1,2]. However, in the context of individualized educational systems as represented by Intelligent Tutoring System (ITS) [3], it is important to optimize individual learning comprehension, preference or personality. Some research used profiling or personality in e-learning systems [4,5,6]. Thus, it is very worthwhile to predict personality traits from e-learning logs and to reveal which personalities are associated with what behaviours in the e-learning log. We hypothesized that personality could be predicted from the pattern of learning times during the summer vacation, so we tried to conduct feature engineering and compared the prediction accuracy of different models by Pycaret (Automated Machine learning pipeline package [7]).

2. Related Work

Personality inventories are psychological questionnaires that reveal personality traits of participants with the purpose of better understanding their behavior in applied settings. Big Five inventory [8] is one such model which describes an individual's personality across five dimensions: Openness to experience (O), Extraversion (E), Agreeableness (A), Conscientiousness (C), and Neuroticism (N). Previous studies showed these personality traits are predictable from cyberspace digital footprints such as Facebook '*like*' data predict Openness (R=0.43) [9], twitter social network data predict Extraversion (R=0.44) [10]. Also, sensor-rich smartphone data predict these five personalities overall R_{median}=0.34 [11]. These previous researches show that personality can be predicted from digital logs. Some research attempted to predict personality traits from e-learning logs (game-based learning environment [12], learner's network behaviors [13]). However, these studies have been limited in their sample size (about fifty participants) and have mainly been conducted in higher education. Therefore, personality was not sufficiently attempted to be predicted from education learning log data especially in K-12 education. In this study, the following research question was posed for the preliminary investigation.

Proceedings of the 4th Workshop on Predicting Performance Based on the Analysis of Reading Behavior, March 21-22, 2022
https://sites.google.com/view/lak22datachallenge

RQ: *Can personalities be predicted from K-12 e-learning learning logs by machine learning?*

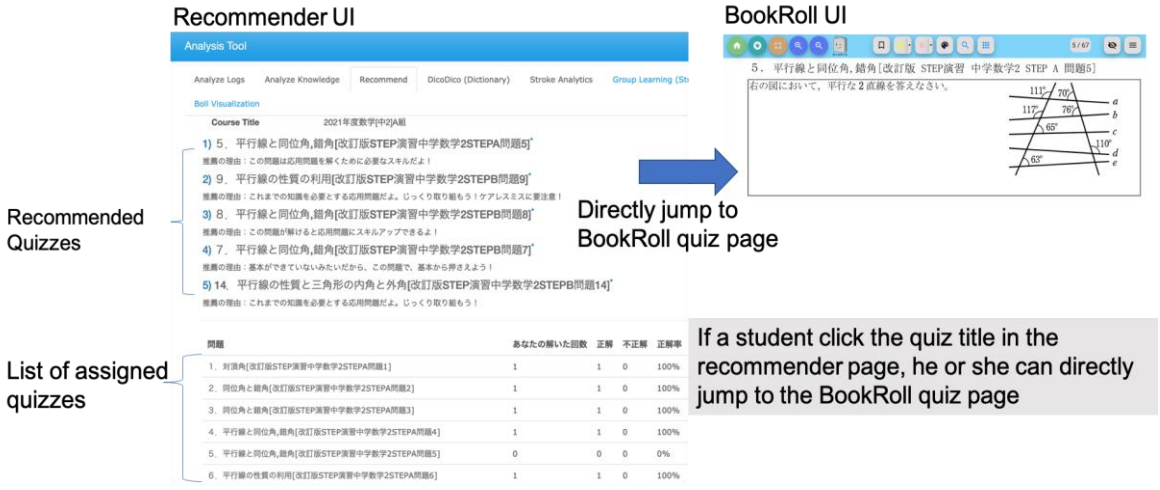
3. Methods

3.1 Participants and dataset

In this study, data were collected from an eBook system named BookRoll that was developed by Ogata et al [14]. The Bookroll reading system provides learning material and quiz exercises to access these materials inside or outside of the classroom. This system also has several features including navigation functions such as NEXT, PREVIOUS, BOOKMARK, etc. for navigating between different pages. The BookRoll system works within the Learning Analytics framework [15] to enable the collection of learning log data. We assumed that there is more diversity in learning patterns during the summer holiday than during the regular school year, and we hypothesized that this diversity is related to personality. So, we decided to use data from the summer vacation period for our analysis. The Big five inventory consists of 70 questions, with a total of 60 questions relating to the five personality factors with 2 options (yes or no). From these, twelve representative questions relating to each of the five main factors are listed in appendix 1. Subjects answer these questions with a choice of yes or no. The overall personality score is calculated by taking the sum of each item and used as the target label for the prediction algorithm.

Before summer vacation, we conducted a Big Five personality questionnaire for first grade high school students and data on personality with no missing values were obtained from 129 students. These students were given the assignment of solving 54 or 58 mathematical quizzes as homework during summer vacation from July 20th to August 23rd, 2021. The students were highly recommended to solve the quizzes and report their answers in the new recommendation system [15,16] included in Bookroll. This recommendation system shows recommended quizzes and all the assigned quizzes as a list in a web page as shown in figure 1. When a student clicks on a quiz, he/she can jump to that quiz on the bookroll. Thus, there is little need to use the NEXT, PREVIOUS and BOOKMARK buttons on the bookroll. Therefore, we did not use these navigation interaction logs for any prediction and directly focused on the reading time of each event as shown in Table 1. Note that we did not filter any extremely long or short reading times, as such behaviours can be characteristics of individuality as shown r_time_max and r_time_min.

Figure 1: Screenshots of Recommender and BookRoll UI



3.2 Data preprocess and prediction

For data analysis, we used Pycaret[7] which is an open source low-code machine learning library in Python. It simplifies the model learning process. This also includes the data pre-processing stage. As a result, the PyCaret library is able to process these functions automatically. Pycaret also automatically creates a model, performs cross validation and evaluates regression metrics, tunes the hyperparameters of a regression model and analyzes model performance using various plots. We performed all analyses

using the default settings, for example, test/hold-out set was 70/30, 10-fold cross validation for model compere.

Table 1
Description of features used in prediction

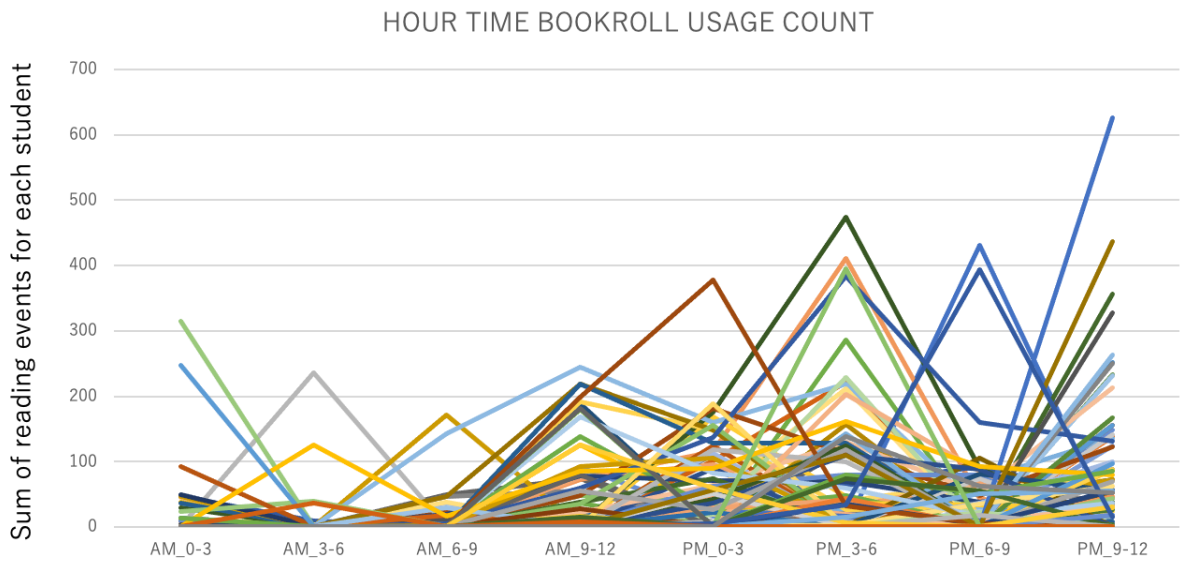
Features	Description	Statistics for all students (N= 129)
r_count	Total number of reading event	25243 (total reading event log)
r_time_sum	Total reading time during summer vacation	1300616 minutes (total reading time for all students)
r_time_mean	Mean reading time (sum of reading time / total reading event)	45.51 minutes (average of mean reading time)
r_time_max	Maximum reading time	6595 minutes (top reading time in one reading event)
r_time_max	Minimum reading time	0
r_time_std	Standard deviation of reading time	mean 139.49, max 1239.5 min 4.45 std 177.81
AM_0-3	Number of events between 0am and 3am	1052 (total events for all students)
AM_3-6	Number of events between 3am and 6am	453
AM_6-9	Number of events between 6am and 9am	677
AM_9-12	Number of events between 9am and 12am	3622
PM_0-3	Number of events between 0pm and 3pm	4294
PM_3-6	Number of events between 3pm and 6pm	5808
PM_6-9	Number of events between 6pm and 9pm	2879
PM_9-12	Number of events between 9pm and 12pm	6458

4. Result

4.1 Predicting personality from e-learning logs

Figure 2 shows the sum of the number of times the bookroll is used in each three-hour period. Each plot shows each student's sum of reading events. As you can see, together with table 1, the midnight to early morning hours, from AM_0-3 to AM_6-9, are used less frequently, but the daytime and evening hours, after 9am, are used more frequently. The low number of use in pm6-9 may be due to having dinner or relaxing.

Figure 2: Hour Time Bookroll usage count for each student



4.2 Data set for prediction

Figure 3: Example of data frame for predicting personality trait

	student_id	r_count	r_time_sum	r_time_mean	r_time_max	r_time_min	r_time_std	AM_0-3	AM_3-6	AM_6-9	AM_9-12	PM_0-3	PM_3-6	PM_6-9	PM_9-12	Target Personality
0	1848	432	4146	9.61948956	507	0	31.6451104	0	0	0	0	0	1	431	0	6
1	1849	141	2627	18.6312057	213	0	28.9110486	0	0	0	0	0	0	2	139	10
2	1851	5	190	38	76	14	22.8363745	0	0	0	5	0	0	0	0	7
3	1852	4	932	233	915	0	454.685972	0	0	0	0	0	4	0	0	2
4	1857	4	21	5.25	12	0	6.18465844	0	0	0	0	0	4	0	0	0
5	1858	323	3123	9.66873065	714	0	41.2928322	0	0	0	37	0	286	0	0	0
6	1859	146	5725	39.2123288	1648	0	159.425486	0	0	0	20	49	10	67	0	3
7	1860	62	1044	17.4	49	0	10.8833569	0	0	25	11	0	26	0	0	5
8	1864	359	13137	37.6418338	2683	0	178.366875	14	0	49	74	0	120	14	88	4

We calculated the reading time from the raw data, and then summarized the number of times it was used at each time of day to get a dataset for prediction, as shown in figure 3. We tried to analyze if the target five personality scores could be predicted from this data set.

4.3 Comparing models

We compared 24 regression models to evaluate performance by Pycaret. This function trains all the 24 models in the model library and scores them using k-fold cross validation for metric evaluation. The table 2 shows the top three algorithms, average Mean Squared Error **(**MSE) and Mean Absolute Percentage Error (MAPE) across the 10 folds along with training time for each five personality scale. We chose the model with the smallest MAPE as the best model and conducted tuning the best model to optimize the parameter.

Table 2

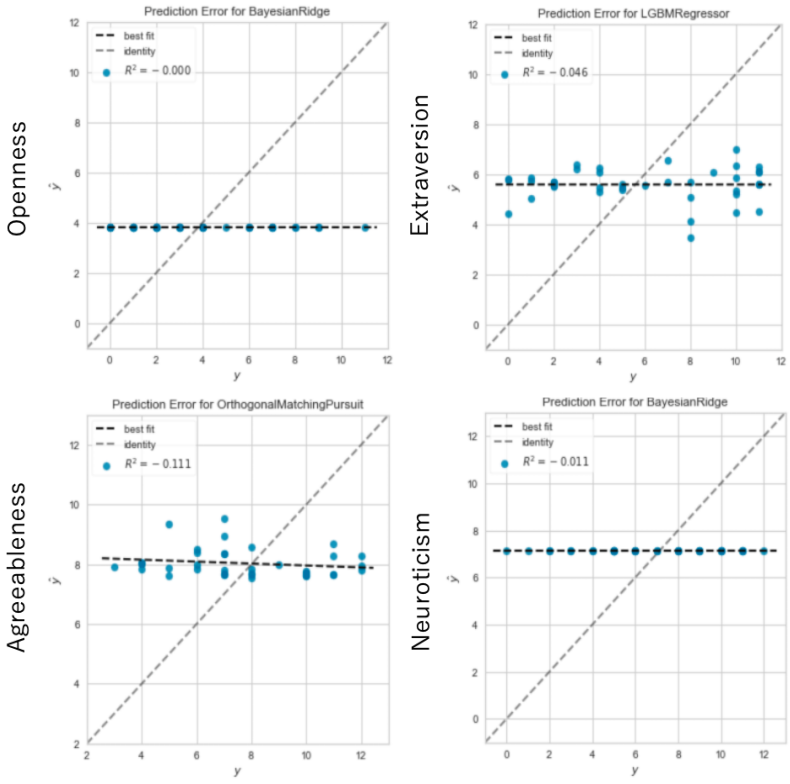
Comparison of models in terms of MSE and MAPE

Personality trait (target)	Top three algorithm	MSE	MAPE
Openness to experience (O)	Bayesian Ridge	2.3228	0.7168
	Orthogonal Matching Pursuit	2.3621	0.7217
	Lasso Regression	2.3839	0.7283

Conscientiousness (C)	Light Gradient Boosting Machine	2.3062	0.5266
	Random Forest Regressor	2.3976	0.5372
	Ada Boost Regressor	2.4264	0.5369
Extraversion (E)	Light Gradient Boosting Machine	3.3398	0.8399
	Bayesian Ridge	3.3981	0.8746
	Matching Pursuit	3.5964	0.9093
Agreeableness (A)	Orthogonal Matching Pursuit	2.4282	0.4649
	Bayesian Ridge	2.4679	0.4653
	K Neighbors Regressor	2.5800	0.4820
Neuroticism (N)	Bayesian Ridge	3.5498	0.8399
	Lasso Regression	3.3921	0.8746
	Elastic Net	3.4615	0.9093

Figure 4 shows prediction error plot of Openness to experience, Extraversion, Agreeableness and Neuroticism. These plots show the actual targets from the dataset against the predicted values generated by each best model. From these figures, it can be seen from the reading learning log that these four indicators did not predict well.

Figure 4: Prediction Error in four personality scale



In contrast, we found that Conscientiousness was predictable (R^2=0.147, R=0.383) from learning material reading logs (Figure 5). Previous study [11] reported overall predictive correlation was about 0.34, thus our result was better score. Residuals plot (the difference between the observed value of the target variable (y) and the predicted value (ŷ)) shows that the points are randomly dispersed around the horizontal axis and error was normally distributed around zero in the histogram. This means this linear model was performing well. We also checked the feature importance in this model.

Figure 5: Prediction for Conscientiousness

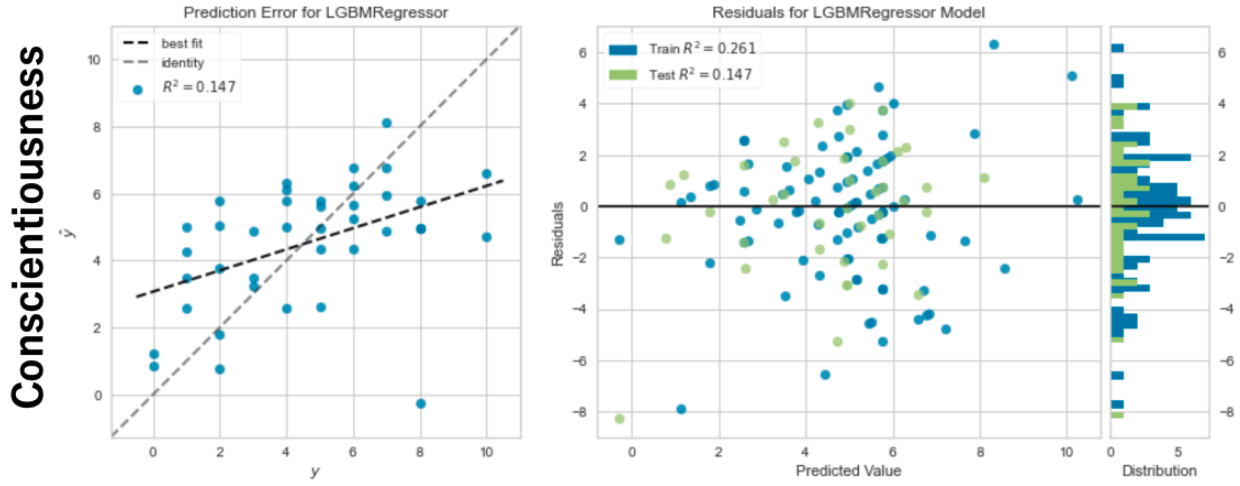
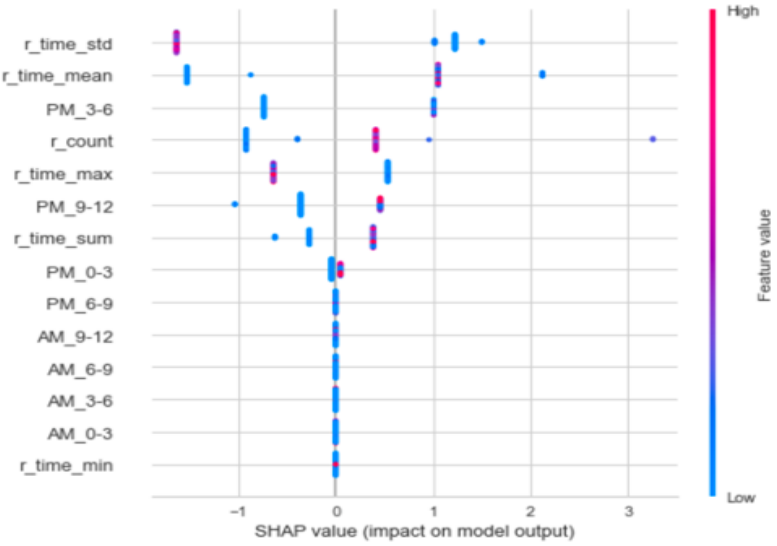


Figure 6 shows the SHAP (SHapley Additive exPlanations[18]) plot sorts features by the sum of SHAP value magnitudes over all samples, and uses SHAP values to show the distribution of the impacts each feature has on the model output. The color represents the feature value (red high, blue low). This reveals that a high r_time_std (reading times were not constant and were highly dispersed) lowers the predicted conscientiousness and a low r_time_std (always the same reading time) and a high r_time_mean (long reading time per event) highers the predicted conscientiousness. These results are consistent with Items of Conscientiousness expressed in the “I like order” or "I follow a schedule”.

Figure 6: SHAP plot of Conscientiousness prediction



5. Summary and future perspectives

We examine the extent to which individuals' Big Five personality traits can be predicted on the basis of learning log data harvested K-12 e-learning system. Taking a machine-learning approach,we compared 24 regression models to evaluate performance. As a result, using the Light Gradient Boosting Machine model we predict conscientiousness (r = 0.38), which is related to academic achievement [19], based on behavioral data collected from 129 high school students' summer vacation learning log. This result is preliminary but the first step to open the prediction of personality from K-12 learning log.

In this study, we were only able to predict conscientiousness. This may be because we only used limited log data from the summer vacation period. Previous study reported extraversion and openness personality were predicted from game-based learning logs [12], agreeableness and extraversion were automatically detected from learner's network behaviors [13]. Thus, it might be possible to predict other personality dimensions if we use long term and various logs i.e., learning logs throughout the year, logs about group learning, students' interaction log in discussion forum etc. If it becomes possible to predict personality to some extent from learning log data, it would be possible to automatically segment people's personalities without the need for questionnaires, and to provide optimal feedback and interventions for each segmentation to realize individualized educational systems.

6. Acknowledgements

This work was partly supported by JSPS Grant-in-Aid for Scientific Research (B) 20H01722, JSPS Grant-in-Aid for Scientific Research (Exploratory) 21K19824, JSPS KAKENHI Grant-in-Aid for Early-Career Scientists 20K20131, JSPS Grant-in-Aid for Scientific Research (S) 16H06304 and NEDO JPNP20006 and JPNP18013.

7. References

[1] JUNCO, Reynol; CLEM, Candrianna. Predicting course outcomes with digital textbook usage data. *The Internet and Higher Education*, 2015, 27: 54-63.

[2] DANIEL, David B.; WOODY, William Douglas. E-textbooks at what cost? Performance and use of electronic v. print texts. *Computers & Education*, 2013, 62: 18-23.

[3] ALKHATLAN, Ali; KALITA, Jugal. Intelligent tutoring systems: A comprehensive historical survey with recent developments. *arXiv preprint arXiv:1812.09628*, 2018.

[4] TZOUVELI, Paraskevi; MYLONAS, Phivos; KOLLIAS, Stefanos. An intelligent e-learning system based on learner profiling and learning resources adaptation. *Computers & Education*, 2008, 51.1: 224-238.

[5] FROSCHL, Christoph. User modeling and user profiling in adaptive e-learning systems. *Graz, Austria: Master Thesis*, 2005.

[6] FATAHI, Somayeh. An experimental study on an adaptive e-learning environment based on learner's personality and emotion. *Education and Information Technologies*, 2019, 24.4: 2225-2241.

[7] PyCaret Available at URL:https://pycaret.org/

[8] JOHN, Oliver P.; DONAHUE, Eileen M.; KENTLE, Robert L. Big five inventory. *Journal of Personality and Social Psychology*, 1991.

[9] KOSINSKI, Michal; STILLWELL, David; GRAEPEL, Thore. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the national academy of sciences*, 2013, 110.15: 5802-5805.

[10] MORI, Kazuma; HARUNO, Masahiko. Differential ability of network and natural language information on social media to predict interpersonal and mental health traits. *Journal of personality*, 2021, 89.2: 228-243.

[11] Stachl C, Au Q, Schoedel R, Gosling SD, Harari GM, Buschek D, Völkel ST, Schuwerk T, Oldemeier M, Ullmann T, Hussmann H, Bischl B, Bühner M. Predicting personality from patterns of behavior collected with smartphones. Proc Natl Acad Sci U S A. 2020 Jul 28;117(30):17680-17687. doi: 10.1073/pnas.1920484117. Epub 2020 Jul 14. Erratum in: Proc Natl Acad Sci U S A. 2021 Jul 20;118(29): PMID: 32665436; PMCID: PMC7395458.

[12] Denden, M., Tlili, A., Essalmi, F. *et al.* Implicit modeling of learners' personalities in a game-based learning environment using their gaming behaviors. *Smart Learn. Environ.* **5,** 29 (2018). https://doi.org/10.1186/s40561-018-0078-6
[13] Ghorbani, Fatemeh, and Gholam Ali Montazer. "E-learners' personality identifying using their network behaviors." *Computers in Human Behavior* 51 (2015): 42-52.
[14] Ogata, H., Yin, C., Oi, M., Okubo, F., Shimada, A., Kojima, K., & Yamada, M. (2015, November). E-Book-based learning analytics in university education. In *International conference on computer in education (ICCE 2015)* (pp. 401-406).
[15] Flanagan, B., Ogata, H. (2018). Learning analytics platform in higher education in Japan, Knowledge Management & E-Learning: An International Journal, 10(4), 469-484.
[16] TAKAMI, K., FLANAGAN, B., DAI, Y., & OGATA, H. Toward Educational Explainable Recommender System: Explanation Generation based on Bayesian Knowledge Tracing Parameters.In: *International conference on computer in education (ICCE 2021)*. 2021. p. 532-537.
[17] TAKAMI, K., FLANAGAN, B., DAI, Y., & OGATA, H. 2022.Educational Explainable Recommender Usage and its Effectiveness in High School Summer Vacation Assignment. In LAK22: 12th International Learning Analytics and Knowledge Conference (LAK22), March 21–25, 2022, Online, USA.ACM, New York, NY, USA, https://doi.org/10.1145/3506860.3506882
[18] SHAP Available at URL:https://github.com/slundberg/shap
[19] O'CONNOR, Melissa C.; PAUNONEN, Sampo V. Big Five personality predictors of post-secondary academic performance. *Personality and Individual differences*, 2007, 43.5: 971-990.

Appendix 1

Big Five constructs and items

Constructs	Items
Openness to experience	● I have a rich vocabulary. ● I have a vivid imagination. ● I have excellent ideas. ● I am quick to understand things. ● I use difficult words. ● I spend time reflecting on things. ● I am full of ideas. ● I am an important person. ● If only I had the opportunity, I could do so much for the world. ● I have difficulty understanding abstract ideas. (*Reversed*) ● I am not interested in abstract ideas. (*Reversed*) ● I do not have a good imagination. (*Reversed*)
Conscientiousness	● I am always prepared. ● I pay attention to details. ● I like order. ● I follow a schedule. ● I am exacting in my work. ● I am a lazy person.(*Reversed*) ● I often work on something and stop halfway through.(*Reversed*) ● I am three-day monk with no patience.(*Reversed*) ● I am a bored person.(*Reversed*) ● I tend not to consider a problem in detail, but to put it into practice.(*Reversed*) ● I make decisions and act rashly.(*Reversed*) ● When things don't go well, I want to throw up immediately.(*Reversed*)

Extraversion	● I am the life of the party. ● I feel comfortable around people. ● I start conversations. ● I talk to a lot of different people at parties. ● I do not mind being the center of attention. ● I am a proactive person. ● I do not talk a lot. (*Reversed*) ● I keep in the background. (*Reversed*) ● I have little to say. (*Reversed*) ● I do not like to draw attention to myself. (*Reversed*) ● I am quiet around strangers. (*Reversed*) ● I am not a good public speaker. (*Reversed*)
Agreeableness	● I am interested in people. ● I sympathize with others' feelings. ● I have a soft heart. ● I take time out for others. ● I feel others' emotions. ● I make people feel at ease. ● I like to take care of children and the elderly. ● I don't want to help if it's against me, even if everyone else has decided.(*Reversed*) ● There is not much to be gained by working with integrity.(*Reversed*) ● I can't really trust even my closest colleagues.(*Reversed*) ● When people are nice to me, I tend to be wary of them because I think they have ulterior motives.(*Reversed*) ● People's words can be deceptive, so it's best not to believe them.(*Reversed*)
Neuroticism	● I get stressed out easily. ● I worry about things. ● I am easily disturbed. ● I get upset easily. ● I change my mood a lot. ● I have frequent mood swings. ● I get irritated easily. ● I often feel blue. ● I am sure that I worry about things that I don't need to worry about myself. ● I'm often nervous and frustrated. ● I am relaxed most of the time. (*Reversed*) ● I seldom feel blue. (*Reversed*)