Proceedings of the 2023 International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/8/20230266

Analysis on the accuracy of different classification prediction models in the field of the impact of online classes on students during the COVID-19

Yucen Qian

School of Mathematical Sciences, Jiangsu University, Zhenjiang, Jiangsu Province, 212013

Jackal1230520@163.com

Abstract. Some students are pleased, while others are sad, as a result of the introduction of online education. Different attitudes, levels of physical health, family economic conditions, and so on have resulted from online classes. Consequently, it is necessary to determine what has caused various students to have different states and whether or not parents of school-aged children took the appropriate response measures. In this paper, the author preprocesses relevant datasets from Kaggle, then uses naive Bayesian, random forest, K-neighbor, SVC, logistic regression, and neural networks to classify and predict the dataset, analyses their accuracy and confusion matrices to determine the best classification and prediction model, and conducts importance analysis based on the best classification and prediction model. Therefore, it is possible to determine the significance of each parameter and to make suggestions based on the significance of various parameters. This can be used to determine whether there is room for development and errors in the next plans implemented in the era of online classes, and in the future, it can be used to reduce the health problems of students caused by the next pandemic in the era of online classes.

Keywords: correlation matrix, confusion matrix, naive Bayesian, random forest.

1. Introduction

In Mr.find in the code's Covid learning loss has been a global disaster, he primarily investigated the inconvenience caused by the suspension of COVID-19.,, The family economy is in deficit, and students' ability to learn declines. As noted in the study on Psychological Health Problems and Countermeasures of College Students in the Context of Epidemic Prevention and Control [1], negative emotions accumulated during long-term online classes at home can cause emotional health problems. Numerous people have conducted pertinent investigation on this topic. In Mumand's article survey [2], he focused more on the epidemic era, changes in students' sleep time, and whether or not they desire to learn independently, as well as vaccination, school closures, and the relationship between the economy and population. Because physical health is more essential than learning, the focus of this paper is on how the epidemic has affected students' physical health in the era of online classes; which factors have led to different outcomes in students' physical health; and which factors are the most significant.

DOI: 10.54254/2755-2721/8/20230266

The author of this paper imported the COVID-19 and Its Impact on Students dataset from the Internet to a local machine and then performed preprocessing [3]. The author examined the dataset's parameters. The author chose to fill in the missing values of a few factors with the median and mean values, respectively. The author then decided to use data visualisation in order to display the digital characteristics of the various columns of data. This can also be used to address the first issue raised by the author. The author chose the correlation matrix as his research technique in order to examine the correlation of related column data [4]. This can also include the correlation coefficients between which factors, which can be used to provide pertinent decision-making suggestions and to determine if certain parameters can be deleted due to excessive correlation. The author then divided the dataset into 70% training sets and 30% test sets, processing the data with various classification and prediction models, including naive Bayesian, random forest, K-neighbor, SVC, logistic regression, and neural networks. By observing the accuracy of various classification and prediction f1-scores, the author determined which model is best suited for this dataset and created confusion matrices for future prediction. Through importance analysis, physicians or teachers can determine which factors can be altered to reduce the likelihood of illness among students during an epidemic by determining the importance of each parameter.

This study has demonstrated, to some extent, which factors correlate with the physical health of students. In addition, various classification prediction models can be used to predict whether other students will experience physical health issues during the online course of the epidemic, and relevant medication-based recommendations can be made.

2. Methodology

The author examined the dataset's parameters. The author chose to fill in the missing values of a few factors with the median and mean values, respectively. The author then decided to employ data visualisation in order to display the digital characteristics of distinct data columns. This can also be used to address the first issue raised by the author. The author chose the correlation matrix as his research technique in order to examine the correlation of related column data [4]. This can also include the correlation coefficients between which factors, which can be used to provide pertinent decision-making suggestions and to determine if certain parameters can be deleted due to excessive correlation. The author then divided the dataset into 70% training sets and 30% test sets, processing the data with various classification and prediction models, including naive Bayesian, random forest, K-neighbor, SVC, logistic regression, and neural networks. By observing the accuracy of various classification and prediction f1-scores, the author determined which model is best suited for this dataset and created confusion matrices for future prediction. Through importance analysis, physicians or teachers can determine which factors can be altered to reduce the likelihood of illness among students during an epidemic by determining the importance of each parameter.

3. Results and analysis

First, for the original dataset, the author conducted a screening, deleted some columns that were not inconsistent with the research direction, and digitized the text data, as shown in the following table 1.

Table 1. Digitization of text (original).

	study condition	study device	Change in your weight	health issue	time utilize	do you find more connected
1	excellent	laptop/desktop	Increase	yes	yes	yes
2	good	smart phione	Decrease	no	no	no
3	average	tablet	Remain constant			
4	poor	any gadget				

Then, the missing values were filled in with the average and mode values based on their various factor components[5]. The author filled in the absent values of Medium for online class and Rating of

DOI: 10.54254/2755-2721/8/20230266

online through-class experience with mode values and the remaining missing values with average values. Figure 1 demonstrates that data integrity testing was performed and that the processed dataset was complete.

Figure 1. Data integrity testing (original).

The author then used the health issue during confinement as the final Y value and the remaining variables as parameters to begin researching this topic. In order to prevent it from impeding the progress of the project, the author would first remove each column individually, conduct data visualization, and display its digital characteristics to determine whether any columns contain abnormal values. Figure 2 depicts some of the most significant illustrations.

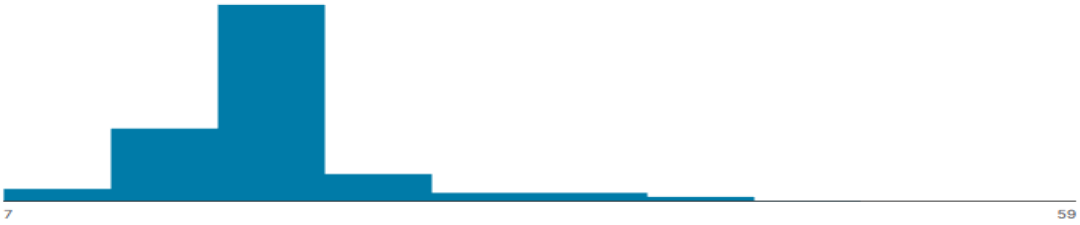

Figure 2. Age distribution.

The majority of the dataset is comprised of middle school students, while the proportion of adult students and elementary school students is relatively small; consequently, the results are skewed towards middle school students.

Figure 3. The health issue during lockdown.

From the fuguer 3,it can be seen that most students have no health problems, but they still cannot be ignored.

DOI: 10.54254/2755-2721/8/20230266

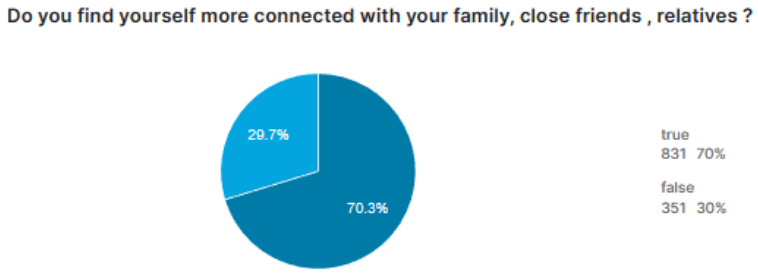

Figure 4. The result of the question "Do you find your self more connected with your family, close friends, relatives?"

Figure 4 indicates that the vast majority of students do not feel more connected to their families, which is a negative phenomenon.

The author then utilised a correlation matrix to corroborate the correlation between different parameters in order to determine whether or not the relevant factors are related. The correlation matrix is depicted in the figure below: Figure 5 depicts a 14 x 14 square, and the colour intensity can approximate the relationship between the two parameters. The information depicted by the tiny square indicates the precise range of correlation values. (0,1). The correlation is higher the larger the value.

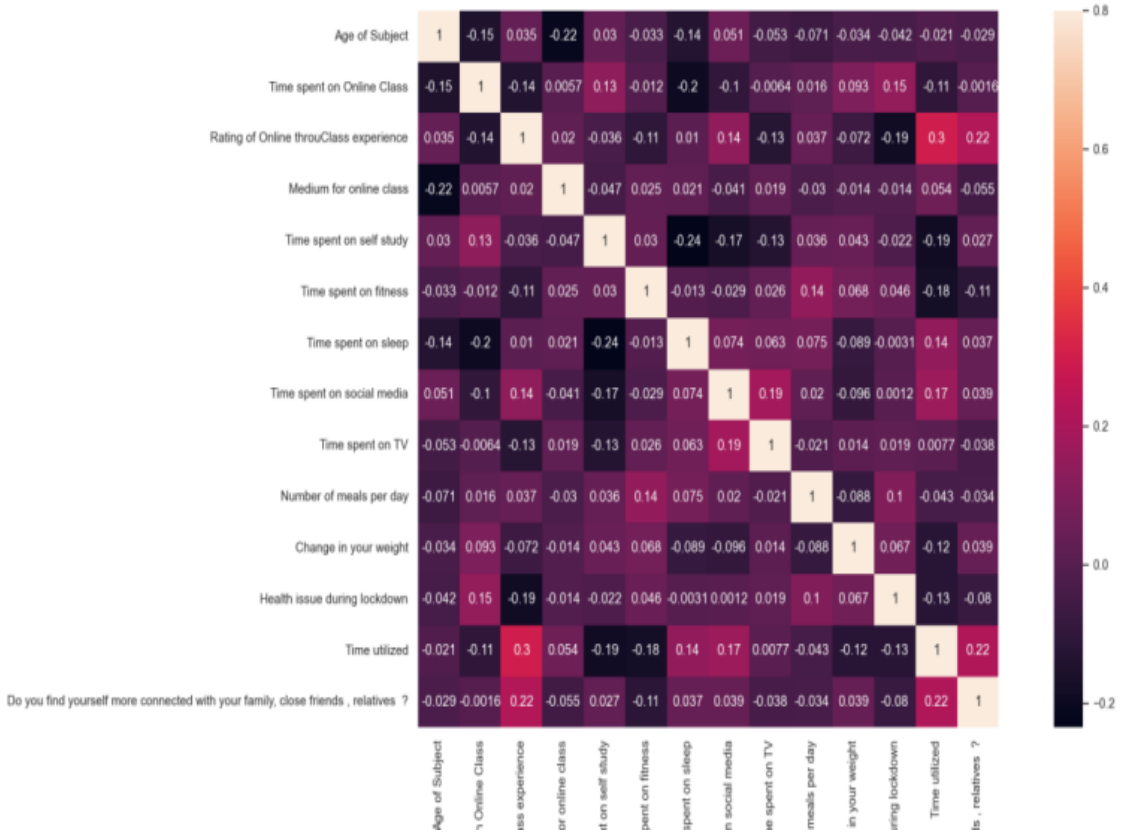

Figure 5. Correlation matrix between 14 factors.

Observing the correlation matrix reveals that the correlation between the various parameters is very low; therefore, it is not necessary to eliminate the parameters in order to conduct an in-depth analysis of the subsequent data.

Using various classification and prediction models to analyse the processed dataset, the author compared the accuracy of naive Bayesian, SVC, random forest, K-neighbor classification, logistic regression, and neural networks to determine the most accurate model.

DOI: 10.54254/2755-2721/8/20230266

Table 2. The accuracy of different classificaiton and prediction models (original).

Model	Accuracy
model: GaussianNB()	accuracy 0.83
model: RandomForestClassifier()	accuracy 0.85
model: SVC()	accuracy 0.85
model: KNeighborsClassifier()	accuracy 0.81
model:LogisticRegression()	accuracy 0.85
model:neural network()	accuracy 0.84

The K-neighborhood classification prediction model has a reduced accuracy, whereas random forest, SVC, and logistic regression have the same accuracy or a higher accuracy, as shown in Table 2. Consequently, they are more appropriate for this dataset. The paper utilised these three classification prediction models to simulate the likelihood that other students will experience physical disease problems during the epidemic. The paper determined the significance of each factor through an analysis of their relative weight. Because the author is more acquainted with random forest models, he chose them for the subsequent problem analysis. Obviously, the author double-checked the confusion matrices of each model using the confusion matrix concept to ascertain their accuracy [6].

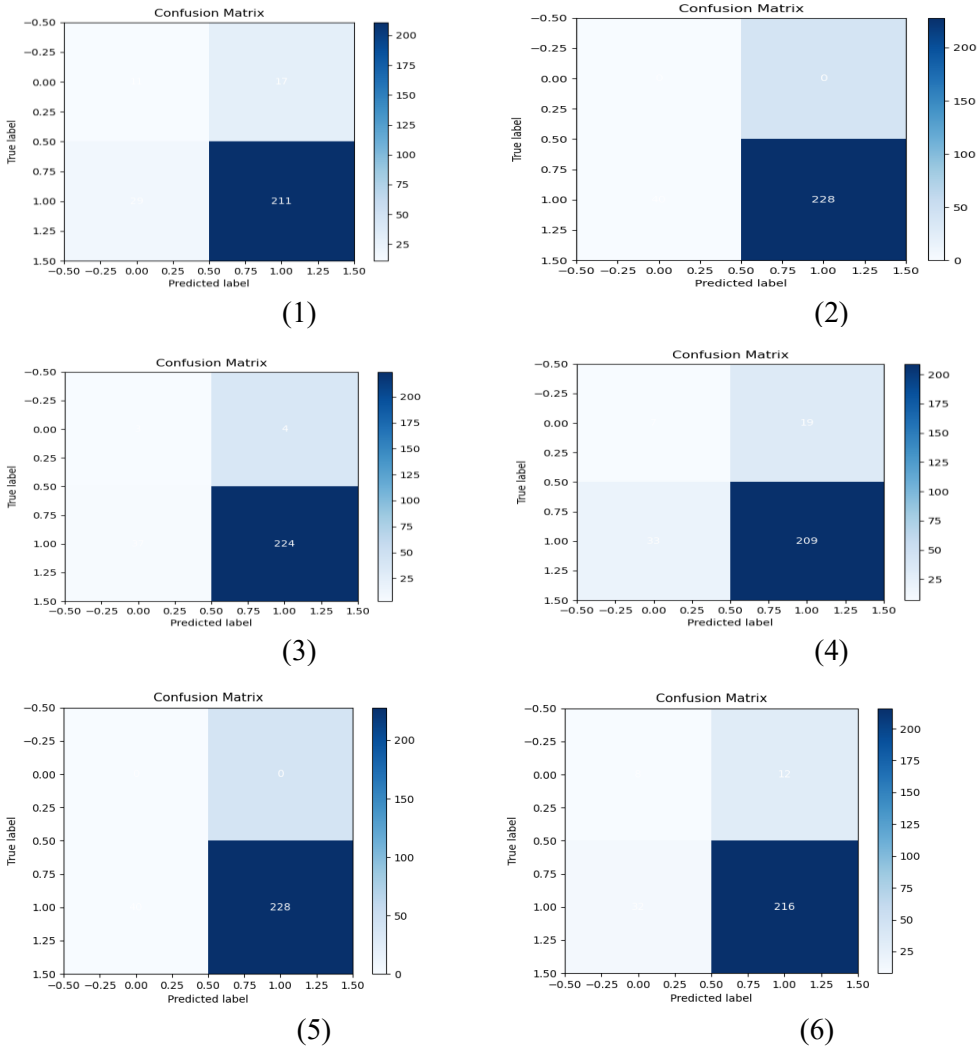

Figure 6. confusion matrixs.

DOI: 10.54254/2755-2721/8/20230266

Figure 6 depicts, in order, the confusion matrices of naive Bayesian, SVC, random forest, K-neighbor classification, logistic regression, and neural network. A consistent accuracy inference can be derived through comparison. Therefore, it was determined to use random forests for the subsequent analysis of importance.

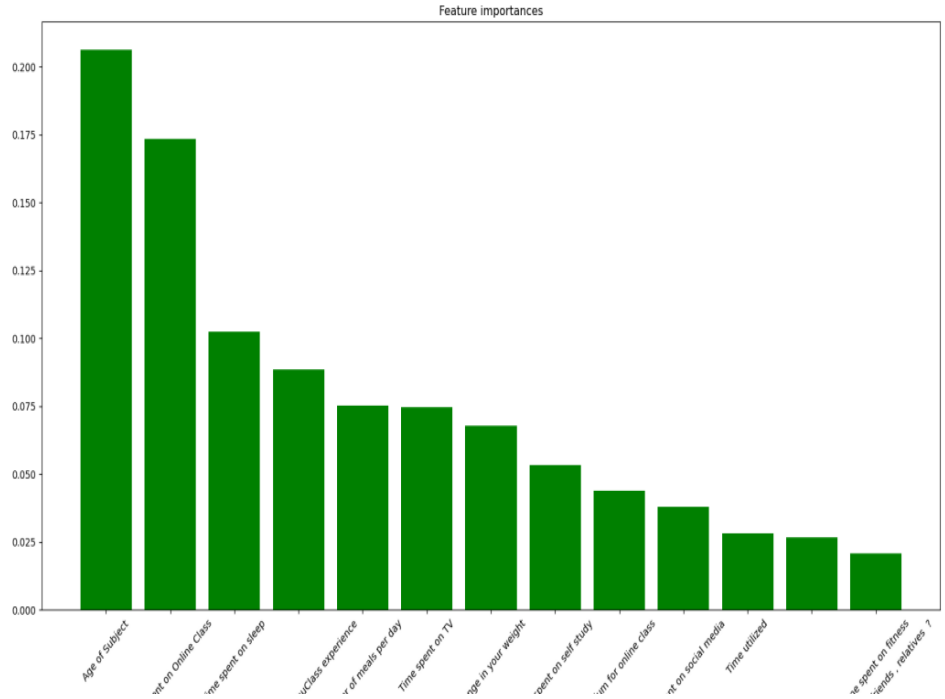

Figure 7. Importance analysis.

Figure 7 demonstrates that based on the random forest model's importance analysis, the online class time and sleep time of students position second and third, respectively. As a supplement to "Research on the Effects and Countermeasures of the Epidemic on College Students' Physical Fitness"[7], students should increase outdoor exercise, schools should reduce online class time and assign appropriate homework, and parents should encourage students not to stay up late. These techniques effectively reduce the incidence of physical health issues. Obviously, suggestions for students of different ages should vary marginally. In the same way that the article "The negative impact of COVID-19 on students' academic performance" divides students into four phases and discusses them separately [8], it is possible to draw more specific conclusions and recommendations.

4. Conclusion

With a prediction accuracy of 85%, the random forest classification prediction model is more appropriate for this online class dataset on student epidemics. Students' online class time and sleep time account for a significant component of the importance, as determined by an analysis of the random forest classification prediction model. Schools should reduce online class time and assign appropriate homework, while parents should encourage their children not to remain up late. These techniques effectively reduce the incidence of physical health issues. However, the sample size is still too small for the classification predictions to be accurate, and the data preprocessing could be more precise and comprehensive.The BMI dataset can also be used to predict future student health levels through logistic regression, similar to "Exploring the Impact of COVID-19 Epidemic on BMI of 6-12 Year Old Children in Suzhou City Based on Student Health Monitoring from 2016 to 2020" [9].

To make the method more applicable to the circumstances, the author will divide students into four age categories in future studies: primary school students aged 6 to 12, secondary school students aged

DOI: 10.54254/2755-2721/8/20230266

13 to 18, college students aged 18 to 22, graduate students aged 22 and older, and others. All of these will be subjected to an importance analysis based on a random forest model, and then the model will be refined for each grade level. Classifying data by grade level can improve the precision of recommendations and results. According to Li Fenglin's paper [8], the factors that contribute to different effects of online courses can be categorised as either internal or external. Internal and external factors contribute to the condition of physical health.

References

[1] Wu Zhipeng, Zhang Min, Liu Zhibiao, Xu Zhiyuan. A Study on the Psychological Health Problems and Countermeasures of College Students in the Context of Epidemic Prevention and Control [J]. Modern Business Trade Industry,2021, (24),30.

[2] COVID-19, School Closures and Learning Situation | Kaggle, 2021. https://data.unicef.org/resources/education-disrupted/

[3] COVID-19 and its Impact on Students | Kaggle, 2021. https://data.unicef.org/resources/education-disrupted/

[4] Correlation matrix, 2018. https://blog.csdn.net/zzw000000/article/details/81205027

[5] Zhen Tan. Data Analysis - Detailed Explanation of Missing Value Processing, 2020. https://zhuanlan.zhihu.com/p/137175585

[6] Concept of confusion matrix, 2021. https://blog.csdn.net/qq_38436431/article/details/120538673

[7] Yu Mingquan, Pan Xiao, Sui Xiuzhi. Research on the Impact and Countermeasures of the Epidemic on the Physical Fitness of College Students [J]. Pedagogy Digest, 2021, 40-45.

[8] The negative impact of COVID-19 on students' academic performance, 2022. https://zhuanlan.zhihu.com/p/541285903

[9] Hu Jia, Han Di, Ding Ziyao, Hai Bo, Yin Jieyun, Yang Haibing, Shen Hui. Exploring the Impact of COVID-19 Epidemic on BMI of 6-12 Year Old Children in Suzhou City Based on Student Health Monitoring from 2016 to 2020 [J].Intended for healthcare professionals, 2021, 33-35.https://doi.org/10.1136/bmj.n953

[10] Li Fenglin. A Study on the Factors Influencing the Adaptability of Vocational College Students to Online Learning during the Epidemic [D]. Guangdong Technical Normal University, 2022, 60-62. DOI：10.27729/d.cnki.ggdjs.2022.000073