Data Fusion for Prediction of Variations in Students Grades

Renata Teixeira[1][0009−0007−5904−9596], Francisco S. Marcondes[1][0000−0002−2221−2261], Henrique Lima[2], Dalila Durães[1][0000−0002−8313−7023], and Paulo Novais[1][0000−0002−3549−0754]

[1] LASI/ALGORITMI Centre, University of Minho, Guimarães, Portugal
[2] Codevision, S.A., Braga, Portugal
pg47603@alunos.uminho.pt, francisco.marcondes@algoritmi.uminho.pt, henrique.lima@e-schooling.com, dad@di.uminho.pt, pjon@di.uminho.pt

Abstract. Considering the undeniable relevance of education in today's society, it is of great interest to be able to predict the academic performance of students in order to change teaching methods and create new strategies taking into account the situation of the students and their needs. This study aims to apply data fusion to merge information about several students and predict variations in their Portuguese Language or Math grades from one trimester to another, that is, whether the students improve, worsen or maintain their grade. The possibility to predict changes in a student's grades brings great opportunities for teachers, because they can get an idea, from the predictions, of possible drops in grades, and can adapt their teaching and try to prevent such drops from happening. After the creation of the models, it is possible to suggest that they are not overfitting, and the metrics indicate that the models are performing well and appear to have high level of performance. For the Portuguese Language prediction, we were able to reach an accuracy of 97.3%, and for the Mathematics prediction we reached 95.8% of accuracy.

Keywords: Data fusion · Academic performance · Education · Computer science · Machine Learning.

1 Introduction

The relevance of education in our lives is remarkable. From a very young age we enter school and start learning not only about writing and reading, but also about history and science. Education is an essential aspect that plays a huge role in the modern and industrially driven society. People need a good education to be able to survive in this competitive world [1]. Educated people are better able to form opinions about various aspects of life, and they also have better job opportunities. Education helps us grow and develop.

Despite the importance of education, it is undeniable that there are still students who fail, and there aren't there many ways to predict whether or not a

student is at risk of failing, one can only conclude this when they get negative grades or even fail the year. One way to try to prevent this problem may come from early prediction of school failure, since the accurate detection of students at risk of failing is of vital importance for educational institutions, it can provide feedback to support educators in making decisions about how to improve student's learning and enable them to apply intervention measures and learning strategies aimed at improving the academic performance of students [2].

The present world is marked by the abundance of diverse sources of information, which makes it difficult to ignore the presence of multiple possibly related datasets, since they may contain valuable information that will be lost if these relationships are ignored [3]. Data Fusion can take advantage of the large amounts of data to help create more complete and consistent datasets. This method is also used in the area of education, namely the area of student performance prediction, due to the amount of different data that can affect the performance of students, like academic grades, parents' education level, interest in school, prospects for the future, *etc.*

In this paper we propose to predict variations in student's Portuguese language and Mathematics' classes' grades from one trimester to another, in order to be able to know beforehand the possibility of a student's grade getting worse, which, in more worrisome cases, can be very useful.

The rest of this paper is organized as follows. In Section II, related work is reviewed using the PRISMA statement and checklist. The proposal is presented in Section III, providing in-depth details about how the dataset was created and processed, about the building of the predictive models, and an analysis of the experimental results through relevant metrics. Finally, the conclusions and further considerations are summarized in Section IV.

2 Related Work

2.1 Methodology

This review of articles on Data Fusion in the field of education, namely those that use Data Fusion to predict student's academic performance, was based on the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) [4].This choice is justified by the fact that PRISMA is widely accepted by the scientific community in engineering and computing.

The literature search was conducted on March 2023 in the popular database for computer science: SCOPUS. Considering the field of the study was based on school failure and student performance, the keywords *Student Performance*, *Academic Performance*, *School Failure*, *School Dropout* and *Academic Failure* were added to the query to specify the fields of this work. And since this work focus mainly on data fusion, the keyword *Data Fusion* was added using a conjunction, while all the other keywords were aggregated using disjunctions.

The query shown below was used for the search in the SCOPUS repository applied to the title, keywords and abstract of the documents.

```
TITLE-ABS-KEY ( "Data Fusion" AND ( "Student Performance"
OR "Academic Performance" OR "School Failure"
OR "School Dropout" OR "Academic Failure" ))
```

In order to eliminate unwanted articles among the articles found, some exclusion criteria were defined. Thus, the documents are excluded if they fall in one of these:

EC1 Do not come from the field of computer science or engineering;

EC2 Not freely accessible;

EC3 Do not focus on the variables studied or is out of context.

EC4 Were not written in English or Portuguese, as these are the languages the authors understand.

EC5 Does not show results.

The final query resulted from adding some authors and article titles:

AC1 Articles already studied and considered important for this analysis.

2.2 Data Search Results

The search in the SCOPUS repository identified 15 articles to which the inclusion and exclusion criteria were subsequently applied.

The inclusion criteria **AC1** was introduced because these articles were already studied and considered essential for this analysis. As a result, two articles were introduced, which led us to 17 articles in total.

On the search page of the database, the documents that met the first exclusion criteria **EC1** were filtered out, leaving us with 16 articles that came from the field of computer science or engineering. Of these documents received, the title and abstract were read and it was found that even though all were written in English (**EC4**), 2 documents were not freely available(**EC2**). The remaining 14 articles were read in full, and the third exclusion criteria **EC3** was responsible for eliminating 5 more articles that were considered to be out of the context of this study.

In Fig. 1 is the PRISMA flow diagram related to this study, which helps in understanding the whole process described above.

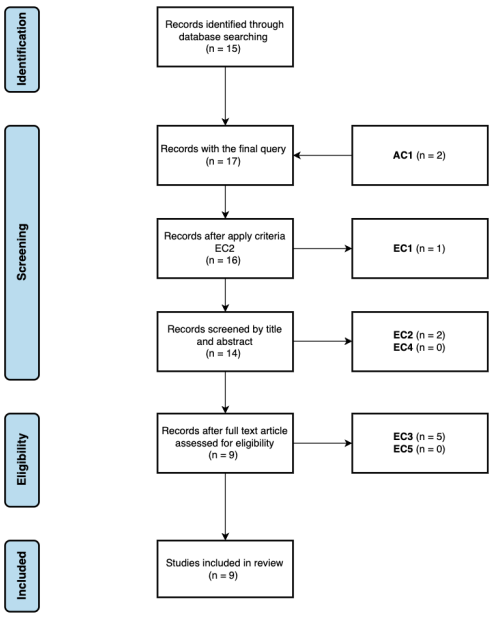

Fig. 1. PRISMA Flow Diagram

Below, an overview of the remaining 9 articles is given.

The article [5] merged collected and manually aligned student behavior data with textual data of the course comments to predict student performance, using a designed multimodal data fusion approach. The empirical research indicated that the proposed method could fuse two different types of data and achieves the best classification performance compared to the base methods. The study outcomes show that the classification method can achieve better classification results in terms of RECALL, F1-measure (F1) and the area under the receiver operating characteristic curve (AUC).

The study conducted in [6], intended to scientifically evaluate the effect of video teaching mode, find out its advantages and find common rules. For this purpose, two classes with 30 students were selected, and different teaching methods were adopted for the two classes: one class adopted the traditional teaching mode, and the other class used the video teaching method. It was concluded that in video image teaching, the students could acquire knowledge faster, and their understanding had been greatly improved. According to the article, adding data fusion also helps teachers and students to improve teaching methods, to provide more targeted assistance, so that teachers' teaching efficiency is continuously improved, and students' learning outcomes are constantly increasing.

The article [7] focuses on engineering students using real-life data and focuses on creating a predictive model, based solely on academic data. The authors of

this paper believe that applying Feature Engineering techniques has a great impact on learning predictive systems and discuss numerous processes on data, such as dealing with redundancy, correlation, missing values, feature creation or deletion, and data fusion, for example. For part of the study, they worked with the three datasets in parallel, applying different data processing and feature engineering techniques, however, at one point they considered it essential to merge the datasets.

The paper [8] was also accepted in this review as it does not fall in any exclusion criteria but it is written by the same authors as [9], and it is considered an initial approaching to solve the proposed problem, so, a review of the article [9] is presented below.

Data Fusion is also an important role in [9], as this paper proposes to use a data fusion and mining methodology for predicting students' final performance starting from multi- source and multimodal data. It gathers data from several sources: theory classes, practical sessions, online sessions with Moodle, and the course final exam. It also applies some pre-process tasks for generating datasets in two formats: numerical and categorical. Secondly, it uses different data fusion approaches like merging all attributes, selecting the best attributes, using ensembles, and using ensembles and selecting the best attributes, and several white-box classification algorithms with the datasets. Finally, after comparing the predictions produced by the models, the authors conclude that the best result was produced using the fourth approach of using ensembles and selection of the best attributes.

The study [13] aimed to develop, train, and test classification models to predict whether students would persist into the second semester beyond traditional measures of performance. According to the authors of this article, data that has been aggregated over time can provide insight into which students are most likely to fail and may need closer attention, while individual student-level performance data can be used to flag students who may be diverging from success to failure. Multiple characteristics such as each student's academic performance, engagement, and demographic background were made available for this project from various sources, which were then compared at the student level and merged into a single dataset.

The article [14] does not aim to predict student performance or failure, instead it proposes a detection framework for detecting students' mental health. It was not eliminated from this review because it was still considered relevant, since it uses data fusion as a main and important part of the development of the work, it is within the context of education and because we believe that mental health and student performance are very closely related. The first step of this work was using representation learning for the fusion of students' multimodal information, like social life, academic performance, and physical appearance.

Similarly to the case above, the paper [15] does not predict school failure, it actually builds and tests prediction models to track middle-school student distress states during educational gameplay. It was considered relevant due to the fact that it is within the context of education and uses data fusion as a main

part of the prediction. In this study multiple types of data from 31 students was collected during a gameplay session. With the collected multimodal data, a multimodal data fusion was implemented in order to predict changes in the outcome variable, which was the state of distress. As a result of this experiment, it was concluded that the classifier with multimodal data fusion outperformed the prediction by each of those classifiers with unimodal data sources. These results led the authors to affirm that the improved performance of the fused classifier in this study corroborates the usefulness of multimodal data fusion when building a learning analytics system.

Finally, the article [16] aimed to predict university students' learning performance using different sources of performance and multimodal data from an Intelligent Tutoring System. In this paper multimodal data was collected and preprocessed, and in total three different data fusion approaches and six white-box classification algorithms were used.

3 Proposal

The importance of being able to predict students' performance in school is well established and, as it was possible to confirm in the review above, data fusion is used in such predictions and allows to reach good results and conclusions.

Figure 2 provides visual representation of the entire system for predict student failure. It includes fusion data, the pre-processing steps and the model used for prediction. The system architecture comprises three main components, namely fusion, pre-processing, and ML model.

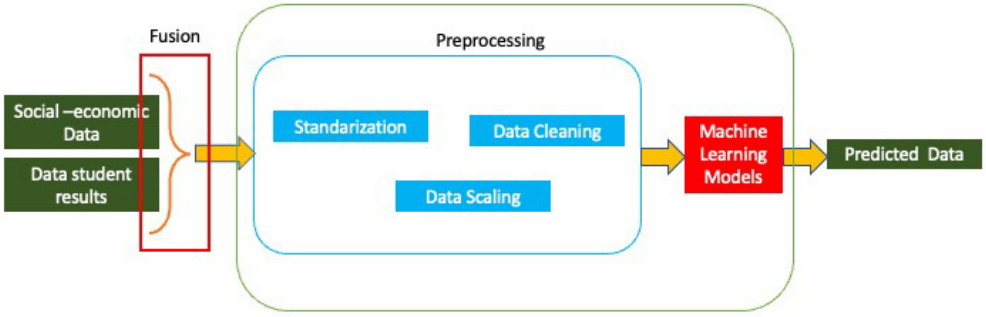

Fig. 2. Visual representation for predicted failure.

3.1 Data Fusion

We used information from 845 middle school students from a school of Northern Portugal. We collected student-related data provided by the school, with data

from a questionnaire to which the students responded, that had questions about school motivation, long and short-term self-regulation, mastery and performance motivation, school value, family, friends and teachers' support and about school engagement.

The resulting dataset is a fused dataset consisted of 845 instances and 123 attributes. These attributes included: sex, age, school year, parent's education level, parent's marriage status, Portuguese and Mathematics classes' grades from the first and second trimester, failure history, satisfaction with grades, tutoring, number of times of tutoring per week, familiar support, friends' support, teachers' support and more.

3.2 Preprocessing of data

We took advantage of the pandas library to handle data as when performing Data Engineering operations.

The first step of the preprocessing was the standarization. So, after fusing the data related to each student into the same dataset, we started treating missing values. We noticed the attributes *"Q9-tipo de explicação"*, type of tutoring, and *"Q10-quantos dias semana"*, how many days of tutoring per week the student has, had 568 and 567 null values each, which meant that 568 students did not get tutoring. For that matter we decided to replace the null values with the value 0. We also noticed that some grades from the first trimester were missing, in this case those rows were dropped. The attributes *"bullying vítima"*, victim of bullying, and *"bullying bully"* also had missing values, we decided to replace those values with the value -1.

Following the process, the second step was data cleaning of handling missing values. We began dropping columns that we didn't consider to be relevant to the context and purpose of this prediction. We noticed the attribute *"Autorização questionário"*, authorization to participate in the questionnaire, had the same value for every instance, so it was dropped. The same happened with the attribute *"Autorização notas"*, authorization to provide grades. The attribute *"Código participação"*, code of participation, was different for each student but considered irrelevant to this study. The attribute *"Escola"*, school that the student attends, was also dropped, as it was considered insignificant.

With the previous task finished, and since most machine learning algorithms only work with numeric values, the next step the data scaling. It was convert categorical features into numerical. The features called *"Q24 Sexo"*, gender of the student, *"Q29-existência de reprovação"*, failure occurrence, *"Q31-escolaridade mãe"*, mom's education level and *"Q35-escolaridade pai"*, dad's education level, were converted to numeric values.

As the resulting dataset had information regarding grades of two different subjects, and the purpose of the study was to attempt to predict the variation of the grades of each subject from the first to the second trimester, the team decided it would be wise to create two identical datasets which would be similar to the original dataset but each would only have information about either Math

or Portuguese Language second trimester grades, let's call them, *Port_ dataset* and *Math_ dataset*.

Since the dataset didn't have a target feature, we created one. For each dataset we subtracted the first trimester grade of one subject from the second trimester grade of that subject and stored the resulting value in a variable. Then we checked whether the variable was greater than, less than, or equal to 0, which would mean the student had improved, worsened or maintained the subject's grade from one trimester to the other, and the target value would be 0,1 or 2 for each possibility described.

Then, in each dataset, we dropped the column *"Q20-nota port. 2.º periodo"*, second trimester grades of the Portuguese Language subject, and the column *"Q26-nota mat 2.º periodo"*, second trimester grades of the Mathematics subject, respectively.

After studying the amount of each target value, we noted that the dataset *Port_ dataset* was made of 671 instances with target value 1, 134 with target value 2 and just 39 with the value 0, meanwhile the dataset *Math_ dataset* had 635 instances with target value 1, 134 with target value 2 and 75 with the value 0. Having balanced data for model training is very important, it gives us the same amount of information to help predict each class and therefore gives a better idea of how to respond to test data, therefore, our data needed to be balanced. We decided to apply data augmentation to solve the problem using random oversampling, resulting in two datasets with 2013 and 1905 instances each.

Finally, as part of the data pre-processing, we performed feature selection on both datasets, in order to compare the results given by the resulting datasets with the results with the bigger datasets. The best features of each dataset were selected and two new datasets were created with them. For this, a function with the target name and the number of features desired as input was created, where the correlation matrix was calculated, and the correlation values of the target variable with all other features extracted. Then, the top n features with the highest absolute correlation coefficients were selected and the resulting dataframe returned. We decided to select the best 20 features for both datasets.

The features resulting from the feature selection for predicting Portuguese grades were: the risk of failing, first trimester Portuguese grade, first and second trimester math grade, satisfaction with grades, the three different total values attributed to the support given by friends, the answer to the first, second, fifth, sixth and seventh questions about friend's support, the answer to the first and fourth questions about short term self-regulation, the total value attributed to the short-term self-regulation, the answer to the sixth question about family support, sex, and the answer to the second question about the student's relationship with peers.

In turn, the features resulting from the feature selection for predicting Mathematics grades were: first trimester Math grade, the answer to the seventh question about long-term self-regulation, personal perspective about grades, the answer to the fifth question about short-term self-regulation, the risk of failure, the

answer to the fourth question about motivation for mastery, satisfaction with grades, age, how many days a week they have tutoring, school year, the two different total values assigned to school valuation, second trimester Portuguese grade, total value attributed to the mastery motivation, the answer to the first and third questions about school valuation, the total value attributed to the long-term self-regulation, the answer to the seventh question about teacher´s support and the answer to the fifth question about school engagement.

3.3 Building of Predictive Models

Two different experiments were conducted in this study, using several classification algorithms. These being using the classification models with hyperparameter optimization, Gridsearch , on the *Port_ dataset* and *Math_ dataset*, or using the same classification algorithms for the datasets resulting from feature selection. After those experiments, the results were compared.

Regarding the large number of available classification algorithms and the enormous possible values to assign to the algorithm hyperparameters, considerations that clearly affect the performance of the models, we decided on some of the algorithms which have been proved successful in related work, which were *Random Forest* [10], *XGBoost*, [11] and *Decision Tree* [12].

First we started by creating the test and training datasets for each case of study. This process consisted of splitting the datasets into 2 parts: the training dataset containing 80% of the instances of the total dataset, and the test dataset making up the remaining 20% of the instances.

As already mentioned above, a feature selection function was created, which had as output the resulting dataframe with the number of features that were given as input. After several experiments and results comparisons, eventually the team decided to choose the best 20 features of both datasets, and for that matter, we finally ended up with a total of 4 datasets, the *Port_ dataset* and *Math_ dataset* and the two dataframes made up of only the 20 selected features, were split into training and test datasets, as already stated.

For each classification algorithm, in order to find the best hyperparameters, we decided to apply grid search instead of random search, since grid search looks at every possible combination of hyperparameters to find the best model and random search only selects and tests a random combination of hyperparameters, instead of conducting an exhaustive search. Therefore, for every model, we defined the parameters to be tuned, created an instance of GridSearchCV and fit the model, and with the best parameters found, we then made the predictions for test data and evaluated the results.

In short, we applied 3 different classification algorithms to the four datasets. The results are shown in the tables below.

In Table 1 we show the results for the *Port_ dataset* and for the one with the selected 20 features of this dataset, which we well call *PFS_ dataset*.

Table 1. Results for the Portuguese Language Prediction

Datasets	Models	Accuracy	F1 Score	Recall	Precision
Port_ Dataset	**XGBoost**	**0.888**	0.867	0.888	0.895
	Decision Tree	0.834	0.834	0.834	0.841
	Random Forest	0.817	0.740	0.817	0.809
PFS_Dataset	**XGBoost**	**0.973**	0.973	0.973	0.974
	Decision Tree	0.948	0.947	0.948	0.952
	Random Forest	0.970	0.970	0.970	0.971

In Table 2 the results of the Mathematics predictions are shown, for both datasets that contain information about this specific subject. The dataset resulting from the feature selection we call *MFS_ dataset*.

Table 2. Results for the Mathematics Prediction

Datasets	Models	Accuracy	F1 Score	Recall	Precision
Math_ Dataset	**XGBoost**	**0.775**	0.712	0.775	0.699
	Decision Tree	0.763	0.671	0.763	0.599
	Random Forest	0.775	0.677	0.775	0.601
MFS_Dataset	**XGBoost**	**0.958**	0.957	0.958	0.960
	Decision Tree	0.908	0.904	0.908	0.916
	Random Forest	0.942	0.941	0.942	0.943

For each algorithm that obtained the best results for each dataset we then applied cross-validation, in order to evaluate the performance of the models, in terms of overfitting. This way we can assure that the models would also perform well on new, unseen data. The results are presented in Table 3.

Table 3. Cross-validation

Datasets	Mean Accuracy	Standart Deviation
Port_Dataset	0.849	0.034
PFS_Dataset	0.966	0.009
Math_Dataset	0.745	0.030
MFS_Dataset	0.940	0.006

Considering the mean accuracy in cross-validation for all four models, the scores indicate that the models are performing well across multiple folds, with high values of mean accuracy and relatively low standard deviations. This can suggest that the models are generalizing well to new data, and are not overfitting to the training data.

Considering that the models are not overfitting, looking at the values in the tables 1 and 2, it is possible to state that the models appear to have a high

level of performance in terms of accuracy, F1 score, recall, and precision. It is also possible to notice that the models performed better with the datasets that consisted only of the 20 selected features instead of the original 117.

4 Conclusions and Future Work

It is of high importance for educational institutions to predict students' grades in order to create innovative strategies taking into account the students and their specific cases, since education is key for the formation of opinions and for the world of employment itself.

In this paper we use data fusion to fuse information about students, their relatives and a questionnaire to which they participated and responded, and predict variations in middle school student's grades in Mathematics and Portuguese Language. If it becomes possible to predict these variations, teachers and educators gain an advantage that allows them to shape their teaching according to the predictions, and instead of being surprised by decreases in grades, they can avoid them.

After the creation of the models, we can suggest that they are not overfitting, and the metrics indicate that the models are performing well and appear to have high level of performance. For the Portuguese Language prediction, we were able to reach an accuracy of 97.3%, and for the Mathematics prediction we reached 95.8% of accuracy. Even though the best results were achieved using the smaller datasets, with the 20 selected features, these features were not only answers given to the questionnaire, but were also grades and characteristics of the students, information that was initially merged, which proves the importance and relevance of data fusion in this study.

As a prospect for future work, we believe it would be interesting to take this to a real-life experience to actually verify whether the model's predictions would really help teachers to accurately predict school failure of their students and whether they would be successful in trying to avoid this failure when shaping their teaching according to individual student needs. One drawback of this study is that the predictions were made manually. However, the plan for future research is to develop a decision support system that can use the prediction results automatically to guide or alert teachers, parents, and other school personnel. Also it is possible to address some issues like discrimination issues.

Acknowledgements

This work is supported by: FCT - Fundação para a Ciência e Tecnologia within the RD Units Project Scope: UIDB/00319/2020 and the Northern Regional Operational Programme (NORTE 2020), under Portugal 2020 within the scope of the project "Hello: Plataforma inteligente para o combate ao insucesso escolar", Ref. NORTE-01-0247-FEDER-047004.

References

1. Dr. Chetlal Prasad, Mrs. Pushpa Gupta, Educational Impact on the society. International Journal of Novel Research in Education and Learning, ISSN 2394-9686, 2020.
2. Romero, Cristóbal & Ventura, Sebastian. (2013). Data Mining in Education. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery.
3. D. Lahat, T. Adali and C. Jutten, "Multimodal Data Fusion: An Overview of Methods, Challenges, and Prospects," in Proceedings of the IEEE, vol. 103, no. 9, pp. 1449-1477, Sept. 2015.
4. Asar S, Jalalpour S, Ayoubi F, Rahmani M, Rezaeian M. PRISMA; Preferred Reporting Items for Systematic Reviews and Meta-Analyses. JRUMS (2016)
5. Y. Qu, F. Li, L. Li, X. Dou and H. Wang, "Can We Predict Student Performance Based on Tabular and Textual Data?," in IEEE Access, vol. 10, pp. 86008-86019, 2022.
6. Zou, Wei & Li, Yanlong & Shan, Xinru & Wu, Xinge. (2022). Application of Data Fusion and Image Video Teaching Mode in Physical Education Course Teaching and Evaluation of Teaching Effect. Security and Communication Networks. 2022.
7. A. J. Fernández-García, J. C. Preciado, F. Melchor, R. Rodriguez-Echeverria, J. M. Conejero and F. Sánchez-Figueroa, "A Real-Life Machine Learning Experience for Predicting University Dropout at Different Stages Using Academic Data," in IEEE Access, vol. 9, pp. 133076-133090, 2021.
8. Chango, Gustavo & Cerezo, Rebeca & Romero, Cristóbal. (2019). Predicting academic performance of university students from multi-sources data in blended learning. DATA '19: Proceedings of the Second International Conference on Data Science, E-Learning and Information Systems. 1-5. 10.1145/3368691.3368694.
9. Wilson Chango, Rebeca Cerezo, Cristóbal Romero, Multi-source and multimodal data fusion for predicting academic performance in blended learning university courses, Computers & Electrical Engineering, Volume 89, 2021, 106908, ISSN 0045-7906.
10. Diaz, P., Salas, J. C., Cipriano, A., & Nunez, F. (2021). Random forest model predictive control for paste thickening. Minerals Engineering, 163, 106760.
11. Ogunleye, A., & Wang, Q. G. (2019). XGBoost model for chronic kidney disease diagnosis. IEEE/ACM transactions on computational biology and bioinformatics, 17(6), 2131-2140.
12. Tong, W., Hong, H., Fang, H., Xie, Q., & Perkins, R. (2003). Decision forest: combining the predictions of multiple independent decision tree models. Journal of Chemical Information and Computer Sciences, 43(2), 525-531.
13. Aguiar, E., Ambrose, G. A. A., Chawla, N. V., Goodrich, V., & Brockman, J. (2014). Engagement vs Performance: Using Electronic Portfolios to Predict First Semester Engineering Student Persistence. Journal of Learning Analytics.
14. T. Guo, W. Zhao, M. Alrashoud, A. Tolba, S. Firmin and F. Xia, "Multimodal Educational Data Fusion for Students' Mental Health Detection," in IEEE Access, vol. 10, pp. 70370-70382, 2022.
15. Moon, J., Ke, F., Sokolikj, Z., & Dahlstrom-Hakki, . I. (2022). Multimodal Data Fusion to Track Students' Distress during Educational Gameplay. Journal of Learning Analytics, 9(3), 75-87.
16.