A Machine Learning Approach to Predictive Modelling of Student Performance

RESEARCH ARTICLE

REVISED A Machine Learning Approach to Predictive Modelling of Student Performance [version 2; peer review: 2 approved]

Hu Ng[1], Azmin Alias bin Mohd Azha[1], Timothy Tzen Vun Yap[1], Vik Tor Goh[2]

[1]Faculty of Computer and Informatics, Multimedia University, Cyberjaya, Selangor, 63100, Malaysia
[2]Faculty of Engineering, Multimedia University, Cyberjaya, Selangor, 63100, Malaysia

v2 **First published:** 11 Nov 2021, **10**:1144
https://doi.org/10.12688/f1000research.73180.1
Latest published: 23 May 2022, **10**:1144
https://doi.org/10.12688/f1000research.73180.2

Abstract

Background - Many factors affect student performance such as the individual's background, habits, absenteeism and social activities. Using these factors, corrective actions can be determined to improve their performance. This study looks into the effects of these factors in predicting student performance from a data mining approach. This study presents a data mining approach in identify significant factors and predict student performance, based on two datasets collected from two secondary schools in Portugal.

Methods – In this study, two datasets are augmented to increase the sample size by merging them. Following that, data pre-processing is performed and the features are normalized with linear scaling to avoid bias on heavy weighted attributes. The selected features are then assigned into four groups comprising of student background, lifestyle, history of grades and all features. Next, Boruta feature selection is performed to remove irrelevant features. Finally, the classification models of Support Vector Machine (SVM) , Naïve Bayes (NB) , and Multilayer Perceptron (MLP) origins are designed and their performances evaluated.

Results - The models were trained and evaluated on an integrated dataset comprising 1044 student records with 33 features, after feature selection. The classification was performed with SVM, NB and MLP with 60-40 and 50-50 train-test splits and 10-fold cross validation. GridSearchCV was applied to perform hyperparameter tuning. The performance metrics were accuracy, precision, recall and F1-Score. SVM obtained the highest accuracy with scores of 77%, 80%, 91% and 90% on background, lifestyle, history of grades and all features respectively in 50-50 train-test splits for binary levels classification . SVM also obtained highest accuracy for five levels classification with 39%, 38%, 73% and 71% for the four categories respectively. The results show that the history of grades form significant influence on the student performance.

Open Peer Review

Approval Status ✓ ✓

	1	2
version 2 (revision) 23 May 2022	✓ view	✓ view
version 1 11 Nov 2021	? view	? view

1. **Sadiq Hussain**, Dibrugarh University, Dibrugarh, India

2. **Huiling Chen**, Wenzhou University, Wenzhou, China

Any reports and responses or comments on the article can be found at the end of the article.

Keywords
Student performance, data mining, support vector machine, naïve bayes, multilayer perceptron

This article is included in the Research Synergy Foundation gateway.

Corresponding author: Hu Ng (nghu@mmu.edu.my)

Author roles: Ng H: Conceptualization, Project Administration, Supervision, Writing – Review & Editing; **bin Mohd Azha AA**: Conceptualization, Methodology, Software, Writing – Original Draft Preparation; **Yap TTV**: Conceptualization, Methodology, Supervision, Writing – Review & Editing; **Goh VT**: Validation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: The author(s) declared that no grants were involved in supporting this work.

How to cite this article: Ng H, bin Mohd Azha AA, Yap TTV and Goh VT. ***A Machine Learning Approach to Predictive Modelling of Student Performance*** **[version 2; peer review: 2 approved]** F1000Research 2022, **10**:1144 https://doi.org/10.12688/f1000research.73180.2

First published: 11 Nov 2021, **10**:1144 https://doi.org/10.12688/f1000research.73180.1

REVISED **Amendments from Version 1**

Referring to comments from the reviewers, we have made the following changes: 1) We have added papers into the Literature Review. 2) We have also amended the Introduction to better show the contributions, significant findings as well as the structure of the paper. 3) We have also added statements that better present the research problem in the Introduction. 4) To improve the flow of the paper, we have also changed the title of the Section 'Methodology' to 'Methodology and Results'. 5) We have also emphasized the significant findings in the discussions and the Conclusions. 6) Updates have been made to Figure 2 to show missing values. 7) The Conclusions have been revised to include the significant results, contribution, and the weakness, which will be addressed in future work.

Any further responses from the reviewers can be found at the end of the article

Introduction

The definition of a student is a person who attends school or any education institution level to achieve a certain level of knowledge or skill set in a course under the supervision of an educator. Almost everyone was once a student with responsibilities to acquire proper education. Acquiring knowledge by means of getting the right education is of utmost importance and each person should have basic equality in receiving education.

When discussing education at the secondary level, a vital aspect to consider is student performance. Student performance can be assessed in a variety of dimensions, either through exam-based assessment or participation-based assessment. Exam-based assessment includes quizzes, midterms, and final exams, while participation-based assessment is a two-way communication during learning and group activities.

Apart from the obvious, there are so many factors that can affect student performance, such as individual habits, absenteeism, social activities after school and others. This gives way to having machines to learn patterns from data so that they can predict how well a student performs; by acknowledging these factors and subsequently detecting and improve their performance as early as possible.

The contributions of this paper are the identification of significant features that influence student assessment, which in turn can be used to develop various predictive models to ascertain student performance. This will assist educators to form corrective or remedial actions can help to improve student performance. In addition, this may also assist in formulating curriculums that may direct students to career pathways that are most suitable for them.

The paper is structured as follows: Related Works, Methodology and Results, followed by Conclusions.

Related Works

Student performance is an essential part in a secondary-level education as it will show where the student stands when continuing to higher education. Daud *et al.*[1] noted that the ability to predict the success of a student is essential and seems to be a fascinating area to dig into.

Sokkhey *et al.*[2] found out that mathematics is one of the subjects that has scientific progression on students. All aspects of human life at various levels are influenced by mathematics and there are no instances in life where mathematics is not used.

Akhtar *et al.*[3] discovered that social status is correlated with family's social and monetary wealth. They managed to find the effect of monetary wealth on students' grade in Pakistan.

Amazona *et al.*[4] as well as Hussain *et al.*[5] have adopted educational data mining (EDM) methods to perform gathering, achieving, and studying of information concerning student's assessment and learning.

On another note, researchers have also looked at student dropout,[6] interpersonal influences[7] as well as career decisions after graduation albeit at the tertiary level.[8–10]

Exploratory data analysis (EDA)[11] is a method of analyzing dataset to summarize the important features *via* visualization. EDA helps:

- to find errors.
- to check assumptions.
- to determine the tentative choice of suitable models and tools.
- to determine the relationship between the dependent and independent variables.
- to detect the directions and size of the relationship between variables.

Feature selection is a component of dimensionality reduction where it reduces the number of features to maximize the performance of a machine learning model. Too many features in a dataset can overwhelm a machine learning classifier and potentially reduce the efficacy.[12]

The Boruta feature algorithm is a wrapper algorithm that underpins the random forest model. From the results yielded by Tang *et al.,*[12] feature selection is able to effectively recognize and improve overall evaluation metrics on their medical dataset research.

Support Vector Machine (SVM) is able to build the best possible boundary of a line called hyperplanes, which can segregate dimensional spaces into classes. In the work of Sekeroglu *et al*.,[13] they achieved good results with SVM on Mathematics and Portuguese subjects from two secondary schools.

Naïve Bayes (NB) is based on Bayes rule of conditional probability and has high capabilities in dealing big datasets.[4] The method is used to estimate the probability of a property given set of data as proof and Bayes' theorem. The posterior is calculated from the product of likelihood and prior and divisible by its evidence.

Multilayer perceptron (MLP) underpins the artificial neutral network (ANN).[4] It has an interconnection of perceptron in which it flows from the input to the output in a single direction with multiple routes.

Methodology and Results

In this research work, the approach consists of seven stages, namely data acquisition, data processing, data integration, data discretization, data transformation, feature selection and classification. The flow of the research is shown in Figure 1.

a) Data acquisition

The dataset of student performance is taken from a population of two Portuguese secondary schools namely Gabriel Pereira Secondary School (395 students)[14] and Mousinho da Silveira Secondary School (649 students).[15] In the survey, the students were taking the subjects, Mathematics and Portuguese. The two datasets were combined and consisted of 1044 students' personal data and scores for the two subjects. The datasets are visualizations and shown in Figures 2 to 6.

b) Data processing

This process helps to validate the two datasets by making sure there is no missing term in any feature.

c) Data integration

The two datasets were combined and consisted of 1044 students' records with 33 features. By adopting EDA,[11] the selected features are then assigned into four groups comprising of student background (12 features), lifestyle (18 features), history of grades (three features) and all features. Tables 1 to 3 shown the features in student background, lifestyle, history of grades respectively. The category 'all' consists of the entire 33 features.

d) Data discretization

Tables 4 and 5 show the binary levels and 5 levels[5] after discretization, representing the grades of the students.

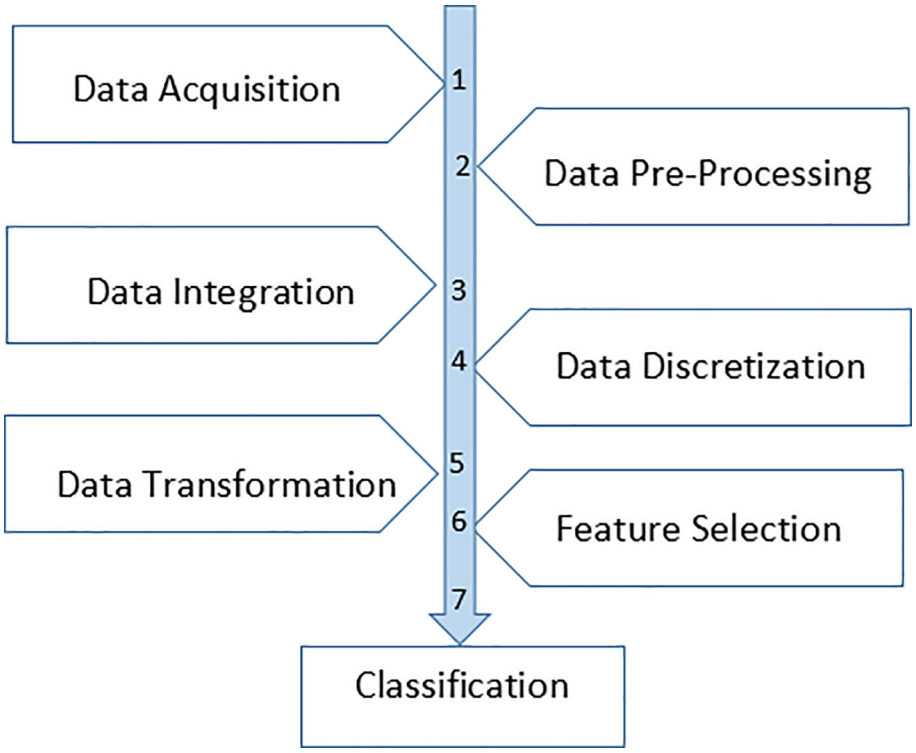

Figure 1. **Flow of the processes.**

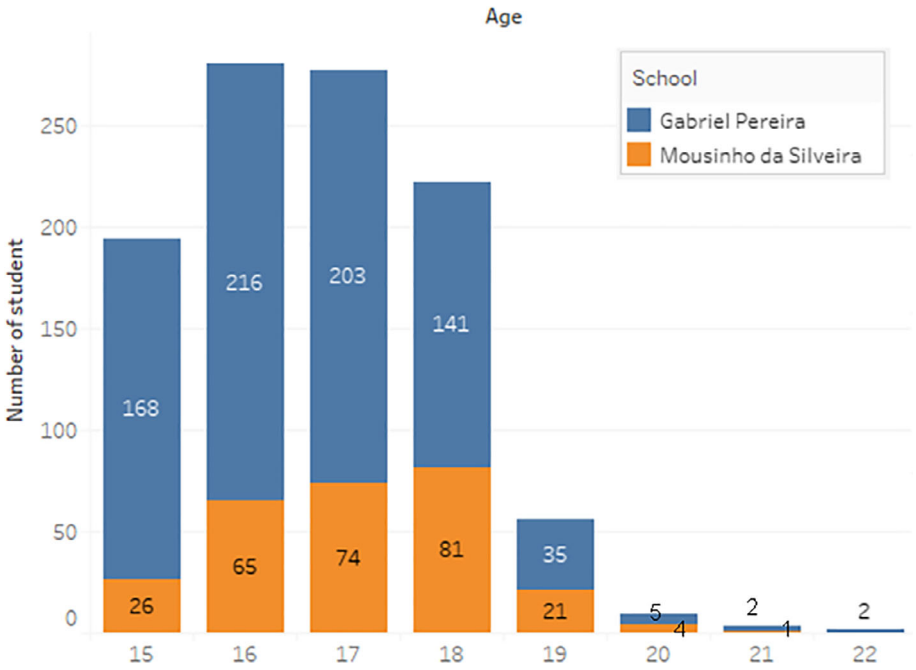

Figure 2. **Distribution of age.**

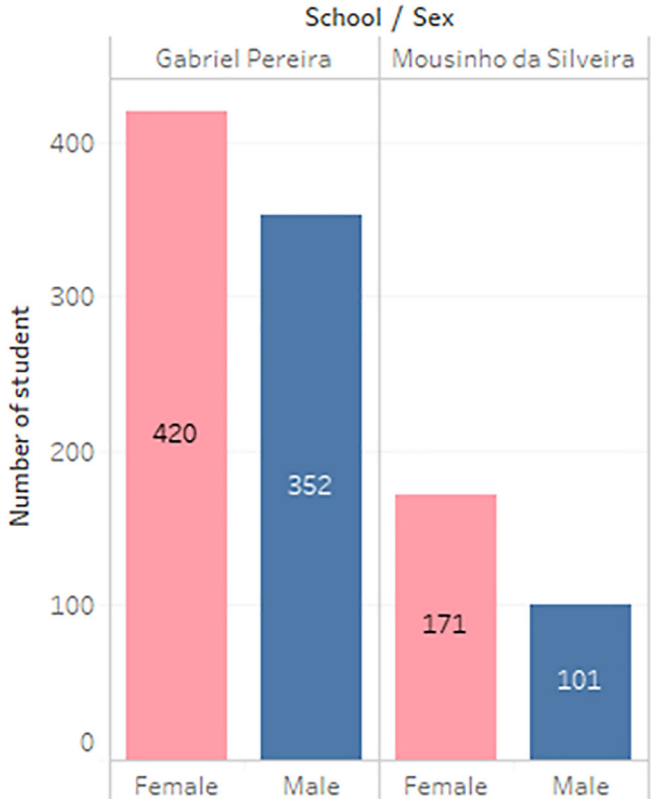

Figure 3. **Distribution of gender.**

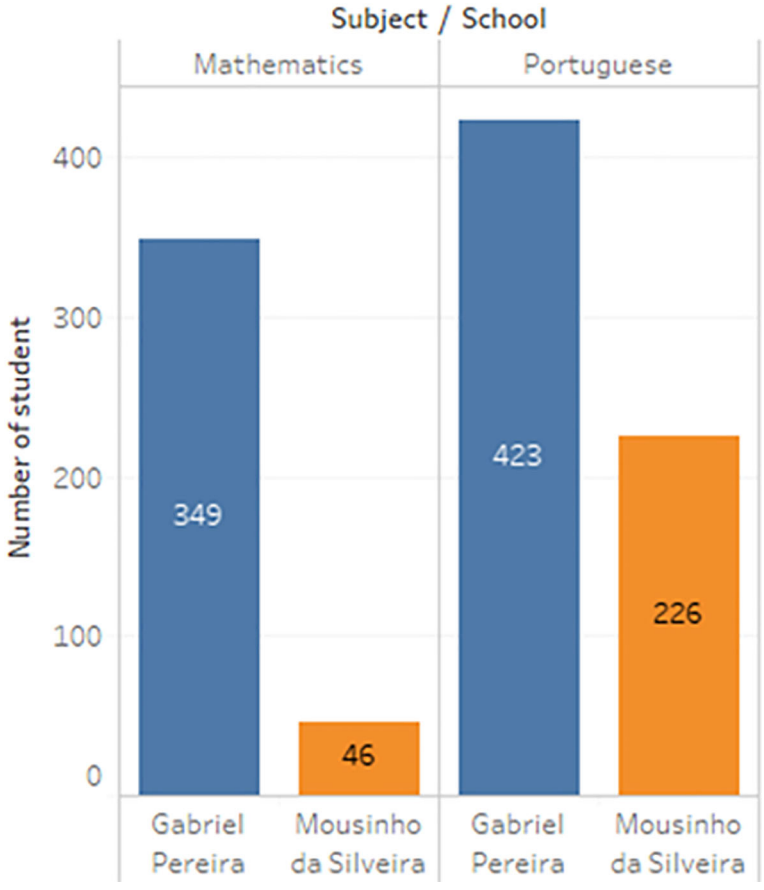

Figure 4. **Distribution of subject.**

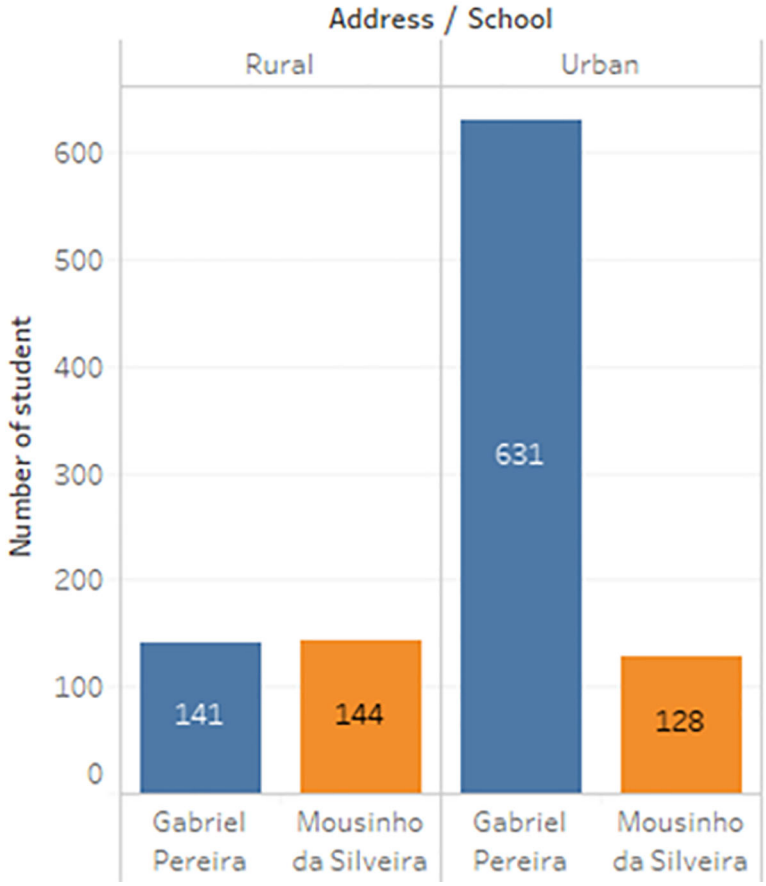

Figure 5. **Distribution of student accommodation.**

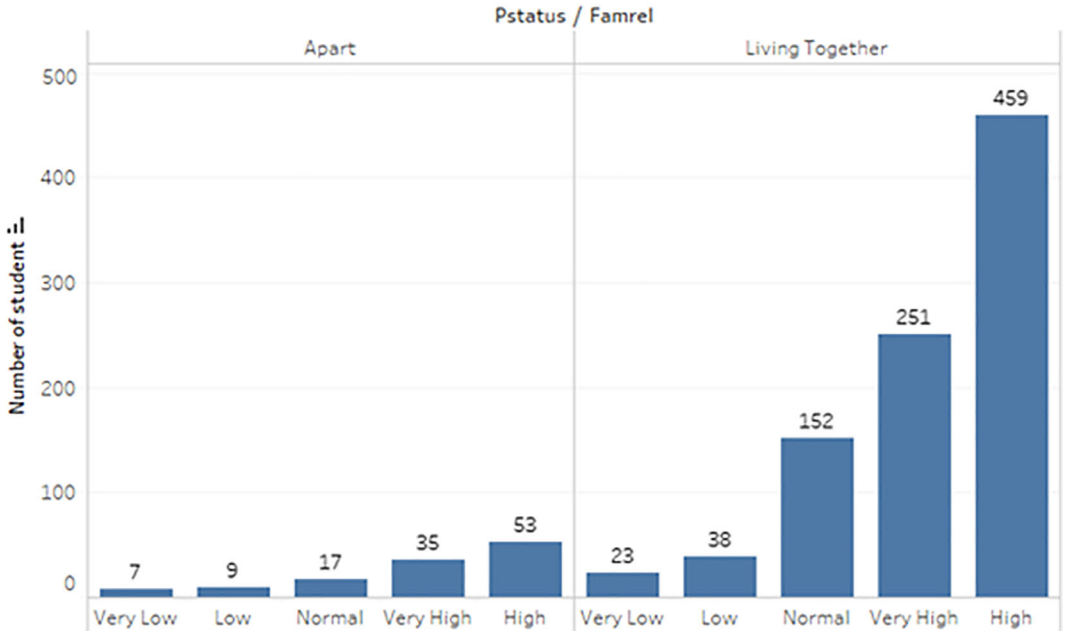

Figure 6. **Distribution of relationship with parents.**

Table 1. Student background.

Feature	Description	Value
sex	Gender of student	Male or Female
age	Age of student	15–22
school	School of student	Gabriel Pereira or Mousinho da Silveira
address	Type of student's home address	Urban or Rural
famsize	Size of family	≤3 or >3
Pstatus	Parent's cohabitation status	Living together or apart
Medu	Education of parents	None, Primary education, 5th to 9th grade, Secondary education, Higher education
Fedu		
Mjob	Job of parents	At home, Civil services, Teacher, Healthcare related, Other
Fjob		
reason	Reason to choose the school	Close to home, School reputation, Course preference, Other
guardian	Guardian of student	Father or mother, Other

Table 2. Student lifestyle.

Feature	Description	Value
traveltime	Travel time from home to school	<15 minutes 15 to 30 minutes 30 minutes to 1 hour >1 hour
studytime	Weekly study time	<2 hours 2 to 5 hours 5 to 10 hours >10 hours
failures	Number of past class failures	n if 1 ≤ n < 3, else 4
schoolsup	Extra educational school support	Yes or no
famsup	Educational support from family	
paid	Extra paid classes within the course subject	
activites	Extra-curricular activities	
nursery	Attended nursery school	
higher	Plans for higher education	
internet	Have internet access at home	
romantic	In a romantic relationship	
famrel	Quality relationship with family	Very low (1) to very high (5)
freetime	Free time after school	
goout	Going out with friends	
Dalc	Weekday alcohol consumption	
Walc	Weekend alcohol consumption	
health	Current health status	
absences	Number of school absences	0–93

Table 3. **Student history of grades.**

Feature	Description	Value
G1	First period grade	0–20
G2	Second period grade	
G3	Final grade	

Table 4. **Binary levels classification.**

Ordinal categorical	Value
Pass	10–20
Fail	0–9

Table 5. **5 Levels classification.**

Ordinal categorical	Value
A	15–20
B	13–14
C	10–12
D	8–9
F	0–7

Table 6. **SVM (Binary levels).**

Metrics	Background	Lifestyle	History	All
60 Train – 40 Test				
Accuracy	0.768	0.789	0.899	0.895
Precision	0.768	0.804	0.928	0.934
Recall	1.000	0.958	0.941	0.929
F1 Score	0.869	0.874	0.934	0.931
50 Train – 50 Test				
Accuracy	0.772	0.798	0.908	0.900
Precision	0.772	0.809	0.932	0.931
Recall	0.999	0.966	0.948	0.941
F1 Score	0.871	0.880	0.940	0.936

Table 7. **SVM (5 Levels).**

Metrics	Background	Lifestyle	History	All
60 Train – 40 Test				
Accuracy	0.394	0.388	0.742	0.716
Precision	0.195	0.322	0.750	0.715
Recall	0.394	0.388	0.742	0.716
F1 Score	0.246	0.286	0.742	0.708
50 Train – 50 Test				
Accuracy	0.389	0.381	0.729	0.708
Precision	0.191	0.329	0.735	0.708
Recall	0.389	0.381	0.729	0.708
F1 Score	0.230	0.300	0.711	0.699

Table 8. **NB (Binary levels).**

Metrics	Background	Lifestyle	History	All
60 Train – 40 Test				
Accuracy	0.760	0.785	0.903	0.891
Precision	0.775	0.826	0.954	0.948
Recall	0.968	0.913	0.918	0.907
F1 Score	0.861	0.867	0.935	0.927
50 Train – 50 Test				
Accuracy	0.761	0.787	0.907	0.894
Precision	0.779	0.830	0.958	0.948
Recall	0.964	0.911	0.922	0.912
F1 Score	0.861	0.868	0.939	0.930

Table 9. **NB (5 Levels).**

Metrics	Background	Lifestyle	History	All
60 Train – 40 Test				
Accuracy	0.373	0.250	0.748	0.515
Precision	0.283	0.263	0.752	0.515
Recall	0.373	0.250	0.748	0.515
F1 Score	0.306	0.166	0.747	0.497
50 Train – 50 Test				
Accuracy	0.370	0.257	0.719	0.542
Precision	0.284	0.261	0.747	0.537
Recall	0.370	0.257	0.739	0.542
F1 Score	0.310	0.173	0.740	0.523

Table 10. **MLP (Binary levels).**

Metrics	Background	Lifestyle	History	All
60 Train – 40 Test				
Accuracy	0.767	0.787	0.899	0.886
Precision	0.769	0.803	0.928	0.921
Recall	0.995	0.957	0.941	0.934
F1 Score	0.868	0.873	0.934	0.927
50 Train – 50 Test				
Accuracy	0.767	0.785	0.906	0.886
Precision	0.773	0.791	0.932	0.919
Recall	0.987	0.982	0.948	0.936
F1 Score	0.867	0.786	0.940	0.927

Table 11. MLP (5 Levels).

Metrics	Background	Lifestyle	History	All
60 Train – 40 Test				
Accuracy	0.386	0.383	0.744	0.715
Precision	0.236	0.305	0.751	0.707
Recall	0.386	0.383	0.744	0.715
F1 Score	0.264	0.301	0.735	0.700
50 Train – 50 Test				
Accuracy	0.371	0.375	0.720	0.705
Precision	0.213	0.361	0.721	0.708
Recall	0.391	0.385	0.720	0.715
F1 Score	0.239	0.326	0.692	0.706

Table 12. Comparison of our models with others research work on two Portuguese secondary schools.

Model and features	Data		
	Mathematics (395 students)	Portuguese (649 students)	Mathematics and Portuguese (1044 students)
SVM on all features [our model]	-	-	0.90
SVM on history of grades [our model]	-	-	0.91
SVM on all features[4]	0.89	-	-
Naive predictor on all features[17]	0.92	0.90	-
SVM on all features[17]	0.86	0.91	-

e) Data transformation

The features are normalized with linear scaling to avoid bias on heavy weighted attributes.

f) Feature selection

Next, Boruta feature selection was performed to remove irrelevant features.

g) Classification

Three supervised machine learning techniques were implemented which are support vector machine, naïve Bayes, and multilayer perceptron 60–40 and 50–50 train-test splits and 10-fold cross validation. Four categories that comprise of student background, student lifestyle, student history of grades (history) and all features. Experiments are carried out on binary levels and five level classification. Binary levels classification will indicate fail or pass, meanwhile for the five levels classification is for student scores F, D, C, B and A.

GridSearchCV is applied to perform hyperparameter tuning. The performance metrics are accuracy, precision, recall and F1-Score. The experiments results are shown from Tables 6 to 11.

SVM obtained the highest accuracy, with scores of 77%, 80%, 91% and 90% on background, lifestyle, history of grades and all features respectively in 50–50 train–test splits for binary classification (pass or fail). SVM also obtained highest accuracy for the five-class classification (grade A, B, C, D and F) with 39%, 38%, 73% and 71% for the four categories respectively. Based on the results, history of student grades shows significant contribution to a good student performance,

where the classification rates obtained are the highest among the four respective categories in each respective classifier. This finding is consistent with the observations from Hwang *et al.*,[16] Mega *et al.*[17] and Waheed *et al.*,[18] that the students' performance is highly related to the history of grades.

Table 12 shows the comparison of our models with other research work in 50–50 train–test splits for binary classification (pass or fail) on the dataset with population of two Portuguese secondary schools.

Conclusions

The paper presented predictive modelling of student performance based on four categories. Based on the results, history of student grades shows significant contribution to a good student performance. SVM obtained the highest accuracy with scores of 77%, 80%, 91% and 90% on background, lifestyle, history of grades and all features respectively in 50-50 train-test splits for binary classification (pass or fail). SVM also obtained highest accuracy for five class classification (grade A, B, C, D and F) with 39%, 38%, 73% and 71% for the four categories respectively. The results show that the history of grades form significant influence on the student performance. The study looks at data only from Portugal and may not reflect a general view of the case. Future work will include more datasets from different countries. Also, other classifiers will be explored and investigated.

Data availability

Underlying data

Kaggle: A machine learning approach to predictive modelling of student performance

https://www.kaggle.com/larsen0966/student-performance-data-set

and

https://archive.ics.uci.edu/ml/datasets/Student+Performance

Ethics approval

Ethical Approval Number: EA1612021 (From Technology Transfer Office (TTO), Multimedia University).

References

1. Daud A, Aljohani NR, Abbasi RA, *et al.*: **Predicting student performance using advanced learning analytics.** *Proceedings of the 26th international conference on world wide web companion.* 2017, April; (pp. 415–421). Publisher Full Text
2. Sokkhey P, Okazaki T: **Comparative Study of Prediction Models on High School Student Performance in Mathematics.** *34th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC).* 2019, June; (pp. 1–4). IEEE. Publisher Full Text
3. Akhtar Z: **Socio-economic status factors effecting the students achievement: a predictive study.** *Int. J. Soc. Sci. Educ.*. 2012; **2**(1): 281–287.
4. Amazona MV, Hernandez AA: **Modelling student performance using data mining techniques: Inputs for academic program development.** *Proceedings of the 2019 5th International Conference on Computing and Data Engineering.* 2019, May; (pp. 36–40). Publisher Full Text
5. Hussain S, Dahan NA, Ba-Alwib FM, *et al.*: **Educational data mining and analysis of students' academic performance using WEKA.** *Indones. J. Electr. Eng. Comput. Sci.* 2018; **9**(2): 447–459.
6. Chung JY, Lee S: **Dropout early warning systems for high school students using machine learning.** *Child. Youth Serv. Rev.* 2019; **96**: 346–353.
7. Nauta MM, Saucier AM, Woodard LE: **Interpersonal influences on students' academic and career decisions: The impact of sexual orientation.** *Career Dev. Q.* 2001; **49**: 352–362.
8. Lee PC, Lee MJ, Dopson LR: **Who influences college students' career choices? An empirical study of hospitality management students.** *J. Hosp. Tour. Educ.* 2019; **31**: 74–86.
9. Kim SY, Ahn T, Fouad N: **Family influence on Korean students' career decisions: A social cognitive perspective.** *J. Career Assess.* 2016; **24**: 513–526. Publisher Full Text
10. Wang Z, Liang G, Chen H: **Tool for Predicting College Student Career Decisions: An Enhanced Support Vector Machine Framework.** *Appl. Sci.* 2022; **12**(9): 4776.
11. Komorowski M, Marshall DC, Salciccioli JD, *et al.*: **Exploratory data analysis.** *Secondary analysis of electronic health records.* 2016; 185–203. Publisher Full Text
12. Tang R, Zhang X: **CART Decision Tree Combined with Boruta Feature Selection for Medical Data Classification.** *2020 5th IEEE International Conference on Big Data Analytics (ICBDA).* 2020, May; (pp. 80–84). IEEE. Publisher Full Text
13. Sekeroglu B, Dimililer K, Tuncal K: **Student performance prediction and classification using machine learning algorithms.** *Proceedings of the 2019 8th International Conference on*

Educational and Information Technology. 2019, March; (pp. 7–11).
Publisher Full Text

14. Cortez P, Silva A: *Student Performance Data Set.* 2014. cited as 2 October.
Reference Source

15. Cortez P, Silva A: **Using data mining to predict secondary school student performance.** *15th European Concurrent Engineering Conference 2008, ECEC 2008-5th Future Business Technology Conference, FUBUTEC 2008.* 2008; **2003**(2000): 5–12

16. Hwang A, Kessler EH, Francesco AM: **Student networking behavior, culture, and grade performance: An empirical study and pedagogical recommendations.** *Acad. Manag. Learn. Edu.* 2004; **3**(2): 139–150.
Publisher Full Text

17. Mega C, Ronconi L, De Beni R: **What makes a good student? How emotions, self-regulated learning, and motivation contribute to academic achievement.** *J. Educ. Psychol.* 2014; **106**(1): 121–131.
Publisher Full Text

18. Waheed H, Hassan SU, Aljohani NR, *et al.*: **Predicting academic performance of students from VLE big data using deep learning models.** *Comput. Hum. Behav.* 2020; **104**: 106189.
Publisher Full Text

Open Peer Review

Current Peer Review Status: ✔ ✔

Version 2

Reviewer Report 13 June 2022

https://doi.org/10.5256/f1000research.134172.r138617

✔ **Huiling Chen**
College of Computer Science and Artificial Intelligence, Wenzhou University, Wenzhou, China

The authors have solved all the comments, now the paper is ready for indexing.

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 14 Jun 2022

Hu NG

Thank you for the approval.

Cheers,
nghu

Competing Interests: No competing interests were disclosed.

Reviewer Report 24 May 2022

https://doi.org/10.5256/f1000research.134172.r138618

✔ **Sadiq Hussain**

System Administrator, Dibrugarh University, Dibrugarh, Assam, India

All my queries and concerns were handled and the article may be accepted for indexing.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Educational Data Mining, Medical Analytics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 24 May 2022

Hu NG

Dearest Reviewer,

Thank you for the comments. Thank you for everything.

Competing Interests: No competing interests were disclosed.

Version 1

Reviewer Report 12 May 2022

https://doi.org/10.5256/f1000research.76815.r136540

?

Huiling Chen
[1] College of Computer Science and Artificial Intelligence, Wenzhou University, Wenzhou, China
[2] College of Computer Science and Artificial Intelligence, Wenzhou University, Wenzhou, China

The authors propose a data mining approach to identify significant factors and predict student performance, based on two datasets collected from two secondary schools in Portugal. The overall structure of the article is also reasonable. The interpretation and description of the experimental results are also detailed. The experimental results show that the proposed method is excellent in terms of accuracy, precision, recall, and F1 score. In my opinion, this method has some practical value.

However, this manuscript suffers from a number of weak points, it should be further improved before being considered for indexing. Let's elaborate on some of them:

1. In the results of the Abstract, the authors summarize the classification results of SVM on the dataset. However, we do not see the impact and contribution of the proposed method on the experimental results.

2. Introduction - include the description of the innovations, contributions, and the structure of the article.

3. In the introduction, I also suggest the authors make a comprehensive investigation on the machine learning method such as the works by Wang *et al.* (2022[1]) and Wang *et al.* (2022[2]), in the literature in the introduction part and give analysis of the existing works to make the whole work more in-depth.

4. The title of the fourth part of the paper, “Methodology”, should be changed to “Methodology and Experimental Results”.

5. The number of students aged 20 and 21 is not given in Figure 2, is it a problem with the data set ?

6. With the experimental results in Table 6-11, there are differences in the results obtained by different classifiers. What is the theoretical basis of the paper for the choice of classifier?

7. In the conclusion, the contributions and flaws of the proposed method are not discussed.

References

1. Wang Z, Liang G, Chen H: Tool for Predicting College Student Career Decisions: An Enhanced Support Vector Machine Framework. *Applied Sciences*. 2022; **12** (9). Publisher Full Text
2. Wang M, Liang Y, Hu Z, Chen S, et al.: Lupus nephritis diagnosis using enhanced moth flame algorithm with support vector machines.*Comput Biol Med*. 2022; **145**: 105435 PubMed Abstract | Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

Yes

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: data mining

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 17 May 2022

Hu NG

Dear Prof Huiling Chen,
We are greatly appreciative of the insightful comments and helpful suggestions that you have provided.
The following are our response on the issues that you have highlighted:

Your comment:

1. In the results of the Abstract, the authors summarize the classification results of SVM on the dataset. However, we do not see the impact and contribution of the proposed method on the experimental results.

Our response:
Thank you for the comments. From this research work, we found out that history of grades forms significant influence on the student performance. This is the main impact and contribution.

Your comment:
2. Introduction - include the description of the innovations, contributions, and the structure of the article.

Our response:

The last 2 paragraphs of the Introduction have been rewritten to reflect this.

The contributions of this paper are the identification of significant features that influence student assessment, which in turn can be used to develop various predictive models to ascertain student performance. This will assist educators to form corrective or remedial actions can help to improve student performance. In addition, this may also assist in formulating curriculums that may direct students to career pathways that are most suitable for them.
The paper is structured as follows: Related Works, Methodology and Results, followed by Conclusions.

Your comment:

3. In the introduction, I also suggest the authors make a comprehensive investigation on

the machine learning method such as the works by Wang *et al.* (2022[1]) and Wang *et al.* (2022[2]), in the literature in the introduction part and give analysis of the existing works to make the whole work more in-depth.

Our response:
We have added one of the papers recommended into the Related Works. Thank you for the recommendation.

Wang, Z., Liang, G., & Chen, H. (2022). Tool for Predicting College Student Career Decisions: An Enhanced Support Vector Machine Framework. Applied Sciences, 12(9), 4776.

Your comment:
4. The title of the fourth part of the paper, "Methodology", should be changed to "Methodology and Experimental Results".

Our response:
The title has been rephrased to 'Methodology and Results'.

Your comment:
5. The number of students aged 20 and 21 is not given in Figure 2, is it a problem with the data set ?

Our response:
Figure 2 had been edited to show the number of students aged 20 and aged 21.

Your comment:
6. With the experimental results in Table 6-11, there are differences in the results obtained by different classifiers. What is the theoretical basis of the paper for the choice of classifier?

Our response:
Due to previous work, we found out that these classifiers work well for our use cases, that is why in this work we have only applied these. We will compare other classifiers in future work.

Your comment:
7. In the conclusion, the contributions and flaws of the proposed method are not discussed.

Our response:
The conclusions have been revised to include the significant results, contribution and the weakness, which will be addressed in future work.

Competing Interests: No competing interests were disclosed.

Reviewer Report 18 November 2021

https://doi.org/10.5256/f1000research.76815.r99883

?

Sadiq Hussain
[1] System Administrator, Dibrugarh University, Dibrugarh, Assam, India
[2] System Administrator, Dibrugarh University, Dibrugarh, Assam, India

The paper presented a predictive model on student performance based on the data from schools in Portugal. Student grades are the most important feature observed in the study. The study is complete from an experimental perspective, but it needs improvement in related works and the introduction section. After these modifications, the study will be approvable.

1. There are grammatical errors in this paper, which should be revised.

2. The literature review should be more in detail and add at least five more papers.

3. The conclusion should also be elaborated a little more. The major findings in their study should be discussed. For example, out of the classifiers applied, which classifier demonstrated the best accuracy? An evaluation of the methodology that the authors deployed would be welcome.

4. The introduction section should also focus on the research problem. Why this kind of research is beneficial, and to whom? How can management take advantage of it and how can companies evaluate these results to find the best students/universities for job placements?

5. Was the dataset balanced? What was the imbalance ratio?

6. Kindly provide justification for using the applied classifiers only: why were other classifiers not considered?

7. What were the most influential features in student performance? Was there any unnecessary feature that was taken into account?

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Educational Data Mining, Medical Analytics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 17 May 2022

Hu NG

Dear Prof Sadiq Hussain,
We are greatly appreciative of the insightful comments and helpful suggestions that you have provided.
The following are our response on the issues that you have highlighted:

Your comment:
a) There are grammatical errors in this paper, which should be revised.

Our response:
The paper has been proofread for grammatical errors. Thank you for pointing this out.

Your comment:
b) The literature review should be more in detail and add at least five more papers.

Our response:
Six papers on student dropout, interpersonal influences, as well as career decisions have been added to the Literature Review.

1. Hussain, S., Dahan, N. A., Ba-Alwib, F. M., & Ribata, N. (2018). Educational data mining and analysis of students' academic performance using WEKA. Indonesian Journal of Electrical Engineering and Computer Science, 9(2), 447-459.
2. Chung, J.Y.; Lee, S. Dropout early warning systems for high school students using machine learning. Child. Youth Serv. Rev. 2019, 96, 346–353.
3. Nauta, M.M.; Saucier, A.M.; Woodard, L.E. Interpersonal influences on students' academic and career decisions: The impact of sexual orientation. Career Dev. Q. 2001, 49, 352–362.
4. Lee, P.C.; Lee, M.J.; Dopson, L.R. Who influences college students' career choices? An empirical study of hospitality management students. J. Hosp. Tour. Educ. 2019, 31, 74–86.
5. Kim, S.-Y.; Ahn, T.; Fouad, N. Family influence on Korean students' career decisions: A social cognitive perspective. J. Career Assess. 2016, 24, 513–526.

6. Wang, Z., Liang, G., & Chen, H. (2022). Tool for Predicting College Student Career Decisions: An Enhanced Support Vector Machine Framework. Applied Sciences, 12(9), 4776.

Your comment:
c) The conclusion should also be elaborated a little more. The major findings in their study should be discussed. For example, out of the classifiers applied, which classifier demonstrated the best accuracy? An evaluation of the methodology that the authors deployed would be welcome.

Our response:
The conclusion has been elaborated with the results. The following has been added:

SVM obtained the highest accuracy with scores of 77%, 80%, 91% and 90% on background, lifestyle, history of grades and all features respectively in 50-50 train-test splits for binary classification (pass or fail). SVM also obtained highest accuracy for five class classification (grade A, B, C, D and F) with 39%, 38%, 73% and 71% for the four categories respectively.

Your comment:
d) The introduction section should also focus on the research problem. Why this kind of research is beneficial, and to whom? How can management take advantage of it and how can companies evaluate these results to find the best students/universities for job placements?

Our response:
The following statements have been added to the Introduction to address this:

This will assist educators to form corrective or remedial actions can help to improve student performance. In addition, this may also assist in formulating curriculums that may direct students to career pathways that are most suitable for them.

For management and companies, this is not in the scope of this research and will be explored in future work.

Your comment:
e) Was the dataset balanced? What was the imbalance ratio?

Our response:
The original datasets are not balanced due to the distribution of grades. In order to overcome this, in this work we performed discretization, to form 2 groups – binary levels and 5 levels. As per referred to the work from

1. Hussain, S., Dahan, N. A., Ba-Alwib, F. M., & Ribata, N. (2018). Educational data mining and analysis of students' academic performance using WEKA. Indonesian Journal of Electrical Engineering and Computer Science, 9(2), 447-459.

Your comment:
f) Kindly provide justification for using the applied classifiers only: why were other classifiers not considered?

Our response:
Due to previous work, we found out that these classifiers work well for our use cases, that is why in this work we have only applied these. We will compare other classifiers in future work.

Your comment:
g) What were the most influential features in student performance? Was there any unnecessary feature that was taken into account?

Our response:
The most influential features come from the history of student grades (Table 3). The following statement can be found in the discussions of the classification result:

Based on the results, history of student grades shows significant contribution to a good student performance, where the classification rates obtained are the highest among the four respective categories in each respective classifier.

No unnecessary features were taken into account.

Competing Interests: No competing interests were disclosed.