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

Hu Ng\textsuperscript{1}, Azmin Alias bin Mohd Azha\textsuperscript{1}, Timothy Tzen Vun Yap\textsuperscript{1}, Vik Tor Goh\textsuperscript{2}

\textsuperscript{1}Faculty of Computer and Informatics, Multimedia University, Cyberjaya, Selangor, 63100, Malaysia
\textsuperscript{2}Faculty of Engineering, Multimedia University, Cyberjaya, Selangor, 63100, Malaysia

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 identifying 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 the 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.
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.

Copyright: © 2022 Ng H et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

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
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.\textsuperscript{12}

The Boruta feature algorithm is a wrapper algorithm that underpins the random forest model. From the results yielded by Tang et al.,\textsuperscript{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.,\textsuperscript{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.\textsuperscript{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 neural network (ANN).\textsuperscript{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)\textsuperscript{14} and Mousinho da Silveira Secondary School (649 students).\textsuperscript{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,\textsuperscript{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\textsuperscript{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.

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

Table 2. Student lifestyle.

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

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>G1</td>
<td>First period grade</td>
<td>0–20</td>
</tr>
<tr>
<td>G2</td>
<td>Second period grade</td>
<td></td>
</tr>
<tr>
<td>G3</td>
<td>Final grade</td>
<td></td>
</tr>
</tbody>
</table>

Table 4. Binary levels classification.

<table>
<thead>
<tr>
<th>Ordinal categorical</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pass</td>
<td>10–20</td>
</tr>
<tr>
<td>Fail</td>
<td>0–9</td>
</tr>
</tbody>
</table>

Table 5. 5 Levels classification.

<table>
<thead>
<tr>
<th>Ordinal categorical</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>15–20</td>
</tr>
<tr>
<td>B</td>
<td>13–14</td>
</tr>
<tr>
<td>C</td>
<td>10–12</td>
</tr>
<tr>
<td>D</td>
<td>8–9</td>
</tr>
<tr>
<td>F</td>
<td>0–7</td>
</tr>
</tbody>
</table>

Table 6. SVM (Binary levels).

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Background 60</th>
<th>Background 50</th>
<th>Lifestyle 60</th>
<th>Lifestyle 50</th>
<th>History 60</th>
<th>History 50</th>
<th>All 60</th>
<th>All 50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.768</td>
<td>0.772</td>
<td>0.789</td>
<td>0.798</td>
<td>0.899</td>
<td>0.908</td>
<td>0.895</td>
<td>0.900</td>
</tr>
<tr>
<td>Precision</td>
<td>0.768</td>
<td>0.772</td>
<td>0.804</td>
<td>0.809</td>
<td>0.928</td>
<td>0.932</td>
<td>0.934</td>
<td>0.931</td>
</tr>
<tr>
<td>Recall</td>
<td>1.000</td>
<td>0.999</td>
<td>0.958</td>
<td>0.966</td>
<td>0.941</td>
<td>0.948</td>
<td>0.929</td>
<td>0.941</td>
</tr>
<tr>
<td>F1 Score</td>
<td>0.869</td>
<td>0.871</td>
<td>0.874</td>
<td>0.880</td>
<td>0.934</td>
<td>0.940</td>
<td>0.931</td>
<td>0.936</td>
</tr>
</tbody>
</table>

Table 7. SVM (5 Levels).

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Background 60</th>
<th>Background 50</th>
<th>Lifestyle 60</th>
<th>Lifestyle 50</th>
<th>History 60</th>
<th>History 50</th>
<th>All 60</th>
<th>All 50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.394</td>
<td>0.389</td>
<td>0.388</td>
<td>0.381</td>
<td>0.742</td>
<td>0.729</td>
<td>0.716</td>
<td>0.708</td>
</tr>
<tr>
<td>Precision</td>
<td>0.195</td>
<td>0.191</td>
<td>0.322</td>
<td>0.329</td>
<td>0.750</td>
<td>0.735</td>
<td>0.715</td>
<td>0.708</td>
</tr>
<tr>
<td>Recall</td>
<td>0.394</td>
<td>0.389</td>
<td>0.388</td>
<td>0.381</td>
<td>0.742</td>
<td>0.729</td>
<td>0.716</td>
<td>0.708</td>
</tr>
<tr>
<td>F1 Score</td>
<td>0.246</td>
<td>0.230</td>
<td>0.286</td>
<td>0.300</td>
<td>0.742</td>
<td>0.711</td>
<td>0.708</td>
<td>0.699</td>
</tr>
</tbody>
</table>
### Table 8. NB (Binary levels).

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

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

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Background</th>
<th>Lifestyle</th>
<th>History</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>60 Train – 40 Test</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.373</td>
<td>0.250</td>
<td>0.748</td>
<td>0.515</td>
</tr>
<tr>
<td>Precision</td>
<td>0.283</td>
<td>0.263</td>
<td>0.752</td>
<td>0.515</td>
</tr>
<tr>
<td>Recall</td>
<td>0.373</td>
<td>0.250</td>
<td>0.748</td>
<td>0.515</td>
</tr>
<tr>
<td>F1 Score</td>
<td>0.306</td>
<td>0.166</td>
<td>0.747</td>
<td>0.497</td>
</tr>
<tr>
<td><strong>50 Train – 50 Test</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.370</td>
<td>0.257</td>
<td>0.719</td>
<td>0.542</td>
</tr>
<tr>
<td>Precision</td>
<td>0.284</td>
<td>0.261</td>
<td>0.747</td>
<td>0.537</td>
</tr>
<tr>
<td>Recall</td>
<td>0.370</td>
<td>0.257</td>
<td>0.739</td>
<td>0.542</td>
</tr>
<tr>
<td>F1 Score</td>
<td>0.310</td>
<td>0.173</td>
<td>0.740</td>
<td>0.523</td>
</tr>
</tbody>
</table>

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

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

### Table 11. MLP (5 Levels).

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

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

<table>
<thead>
<tr>
<th>Model and features</th>
<th>Data</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mathematics (395 students)</td>
<td>Portuguese (649 students)</td>
<td>Mathematics and Portuguese (1044 students)</td>
</tr>
<tr>
<td>SVM on all features [our model]</td>
<td>-</td>
<td>-</td>
<td>0.90</td>
</tr>
<tr>
<td>SVM on history of grades [our model]</td>
<td>-</td>
<td>-</td>
<td>0.91</td>
</tr>
<tr>
<td>SVM on all features $^{+}$</td>
<td>0.89</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Naïve predictor on all features $^{+}$</td>
<td>0.92</td>
<td>0.90</td>
<td>-</td>
</tr>
<tr>
<td>SVM on all features $^{+}$</td>
<td>0.86</td>
<td>0.91</td>
<td>-</td>
</tr>
</tbody>
</table>
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.,\textsuperscript{16} Mega et al.,\textsuperscript{17} and Waheed et al.,\textsuperscript{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


and

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

Data are available under the terms of the Creative Commons Zero “No rights reserved” data waiver (CC0 1.0 Public domain dedication).

**Ethics approval**

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

**References**


Reference Source
Publisher Full Text
Publisher Full Text
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

© 2022 Chen H. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

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, Multimedia University, Cyberjaya, Malaysia

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

© 2022 Hussain S. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Sadiq Hussain

Page 14 of 22
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, Multimedia University, Cyberjaya, Malaysia

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

© 2022 Chen H. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Huiling Chen
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\textsuperscript{1}) and Wang et al. (2022\textsuperscript{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

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**, Multimedia University, Cyberjaya, Malaysia

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.


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.
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, Multimedia University, Cyberjaya, Malaysia

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.


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

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.