sciendo

ICTE Journal

*2021, **10**(1): 32–38*
*DOI: 10.2478/ijicte-2021-0003*

# EFFECT OF ENCODING CATEGORICAL DATA ON STUDENT'S ACADEMIC PERFORMANCE USING DATA MINING METHODS

Moohanad Jawthari[1], Veronika Stoffova[2]

*[1]Department of information systems, Faculty of Informatics , University of Eotvos Lorand, Hungary {mrjibory@gmail.com}*

*[2]Department of computer science, Faculty of education, University of Trnava, Slovakia {email address}*

**ABSTRACT**

Educational data mining (EDM) is the techniques used to discover the knowledge from student's data .it is used to improve the students' performance and teachers' performances as well.

In this paper, we study the effect of encoding some non-ordinal features as one-hot (dummy variables) on the students' performance prediction accuracy. We used techniques form ensemble methods such as Random Forest Trees, Boosting methods specifically namely gradient Boosted trees(GBT), and support vector machines. Also, we compared the performance of Random forest and Gradient boosted trees. We achieve a better result of 81% using random forest classifier. GBT has approximately same performance in all cases. SVM accuracy improved when used dummy variables.

**KEYWORDS**

*E-learning, Educational Data Mining, Student performance, Support Vector machine, Random Forest*

## 1 INTRODUCTION

Educational Data Mining (EDM) is an emerging discipline. It is concerned with developing methods for exploring the unique types of data coming from educational settings and using those methods for better understanding students and the settings which they learn in (Ayesha et al., 2010). It can help the educators to improve teaching methods, to understand learners, to improve learning process, and to improve the learning activities of the learner. Also, EDM helps the administrator to produce good quality outcomes. web-based education, educational repositories and traditional surveys are the resources for collecting educational data, EDM uses various techniques of data mining, machine learning and statistics to analyze and extract the hidden knowledge form educational data context.

The main goal of this article is to study the effect of encoding non ordinal categorical features on the accuracy of predication models. We use the educational dataset of that is collected from learning management system (LMS) called Kalboard 360(Amrieh, Hamtini and Aljarah, 2016). Then we apply data mining techniques, namely, Radom Forest, Gradient Boosted trees, and support vector machines. We do

not apply feature section methods to figure out the accuracy without performing any selection or reduction. Instead, we applied feature encoding methods like dummy variables to make the model understand data and perform efficiently. The results show that there is improvement in the models' performances using GBT and SVM, though GBT slightly improve the prediction accuracy. SVM achieves 7% accuracy improvement over numerical features.

## 2 LITERATURE REVIEW

Much Research has been carried out to demonstrate how important data mining techniques in education, demonstrating that this is a new idea to extract valid and accurate information about the behavior and efficiency in the learning process (Ramaswami and Bhaskaran, 2010) .

Data mining has been utilized to analyze the curriculum and subject of the present research topics in addition to analyzing students' performance. Researchers have investigated in EDM. For instance, (Harwati, Alfiani and Wulandari, 2015) use naïve base algorithm to predict student performance based on 13 variables. Based on the results, a model was built for the purpose of predefining the students who are at risk of failure and thus activating a guidance and counseling program. k-means algorithm was used by Varun and Chadha (2011) to cluster students based on five behavioral features like papers' scores average and seminar notes. According to the results, there was a strong relation between attendance and student performance. Varghese et.al (2010) claim that knowledge through analysis by data mining can improve student performance, organizations management, and the education system in orientation. Another research was conducted on the education system in Portugal (Cortez and Silva, 2008), and the research's results presented a good and precise prediction. This was done by development tools which helped improve the management of education in schools and the effectiveness of learning, which was a significant return.

## 3 DATA SET

The data set was collected by using a learner activity tracker tool, which called experience API (xAPI). The purposed was to monitor the behaviors of students to evaluate the features that may impact the student performance (Amrieh, Hamtini and Aljarah, 2015).

The dataset includes 480 student records and 16 features. The features are classified into three categories: (a) Demographic features such as nationality, gender, place of birth, and relation (parent responsible for student, i.e father or mum). (b) Academic background features such as educational stage, grade Level section id, semester, topic, and student absence days . (c) Behavioral features such as raised hand on class, visited resources, answering survey by parents, and school satisfaction. The dataset features are explained below:

**Table 1** Features descriptions

| Feature | Explanation |
|---|---|
| 1- Gender | student's gender (nominal: 'Male' or 'Female') |
| 2- Nationality | student's nationality (nominal:' Kuwait',' Lebanon',' Egypt',' SaudiArabia',' USA',' Jordan',' Venezuela',' Iran',' Tunis',' Morocco',' Syria',' Palestine',' Iraq',' Lybia') |
| 3- Place of birth | student's Place of birth (nominal:' Kuwait',' Lebanon',' Egypt',' SaudiArabia',' USA',' Jordan',' Venezuela',' Iran',' Tunis',' Morocco',' Syria',' Palestine',' Iraq',' Lybia') |
| 4- Educational Stages | educational level the student belongs (nominal: 'lowerlevel','MiddleSchool','HighSchool') |

| | |
|---|---|
| 5- Grade Levels | grade student belongs (nominal: ‘G-01’, ‘G-02’, ‘G-03’, ‘G-04’, ‘G-05’, ‘G-06’, ‘G-07’, ‘G-08’, ‘G-09’, ‘G-10’, ‘G-11’, ‘G-12 ‘) |
| 6- Section ID | classroom student belongs (nominal:’A’,’B’,’C’) |
| 7- Topic | course topic (nominal:’ English’,’ Spanish’, ‘French’,’ Arabic’,’ IT’,’ Math’,’ Chemistry’, ‘Biology’, ‘Science’,’ History’,’ Quran’,’ Geology’) |
| 8- Semester | school year semester (nominal:’ First’,’ Second’) |
| 9- Relation | Parent responsible for student (nominal:’mom’,’father’) |
| 10- Raised hand | how many times the student raises his/her hand on classroom (numeric:0-100) |
| 11- Visited resources | how many times the student visits a course content(numeric:0-100) |
| 12- Viewing announcements | how many times the student checks the new announcements(numeric:0-100) |
| 13- Discussion groups | how many times the student participate on discussion groups (numeric:0-100) |
| 14- Parent Answering Survey | parent answered the surveys which are provided from school or not (nominal:’Yes’,’No’) |
| 15- Parent School Satisfaction | the Degree of parent satisfaction from school(nominal:’Yes’,’No’) |
| 16- Student Absence Days | the number of absence days for each student (nominal: above-7, under-7) |

The follwoing figures show the topic, nationality, and class distribustions.

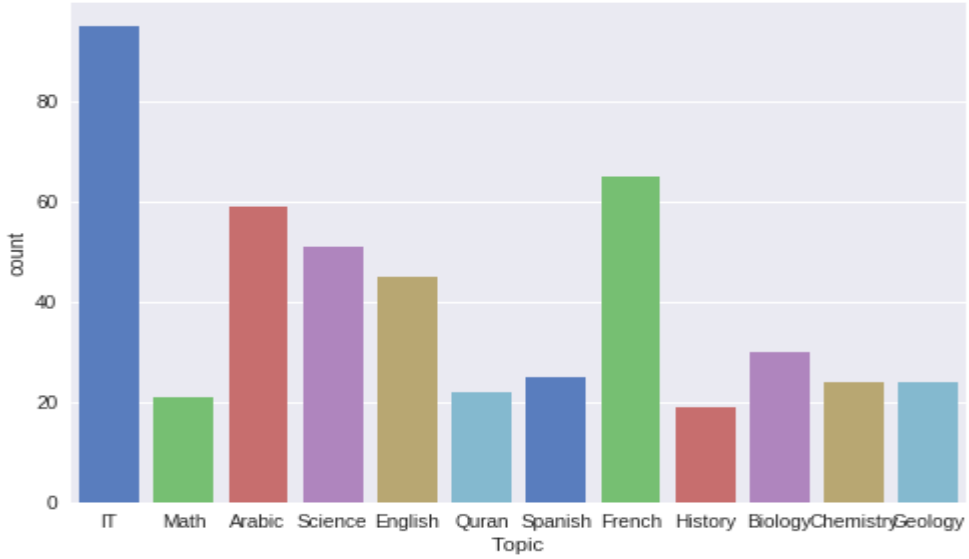


**Figure 1** Topic visualization.

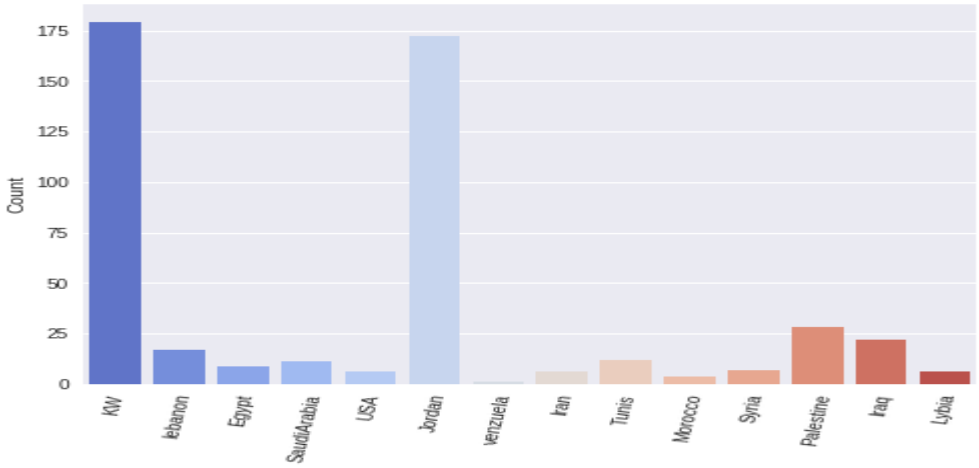


**Figure 1** Nationality visualization.

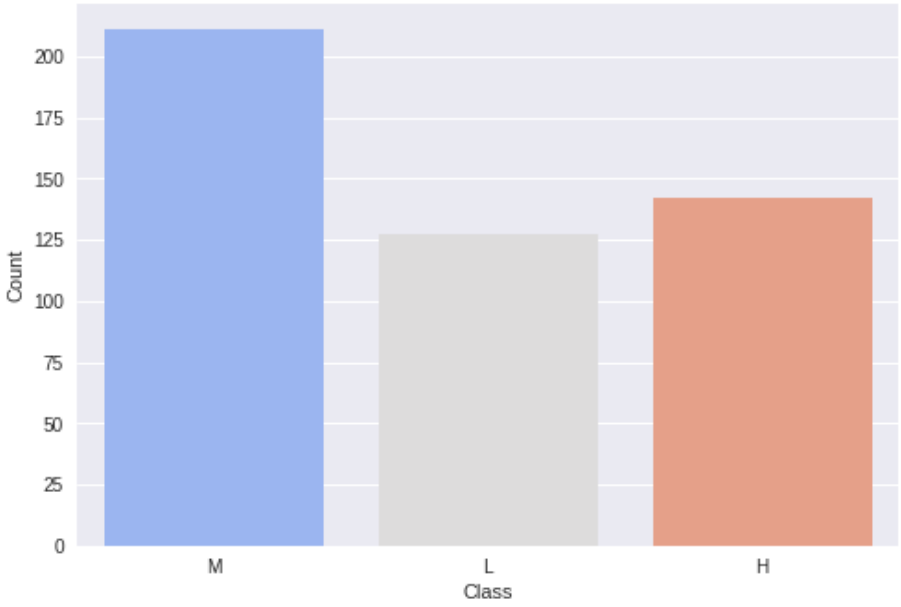


**Figure 2** Class visualization.

## 4 MERTHODOLOGY

The research question was, does using dummy variables instead of numerical improve and affect the performance? since some features, like country, does not have an order, so they will be misunderstood by the model if encoded as numerical. Therefore, we want to explore the encoding effect on the performance. we expect the dummy variable encoding where there is no order in data values improves the prediction accuracy. Hence, we use python get_dummies to obtain dummy variables for categorical variables set. Also, we map the other categorical features that carry order like class to H:2, M:1, and L:0.

We used random forest and gradient boosted trees models for prediction. RF and GBT are ensemble learning methods. They combine the individual trees to obtain predictive results for regression or classification.

Ensemble methods are further categorized as dependent and independent. GBT is an example of dependent methods. RF is an independent ensemble method the combines the output of individual independent learners through voting process.

Also, we use support vector machine (SVM). It performs classification by finding the hyperplane which maximizes the margin between the two classes (binary classification). Multi-class SVMs (MCSVM) are usually implemented by combining several binary SVMs (Chamasemani and Singh, 2011). We have 3 classes, so we utilize multi class prediction models

## 5 EXPERIMENTS AND RESULTS

The study was divided into three parts:

In first part, we encoded all features as numerical as used in the original paper, we further divided and compared the models' performances with using behavioral features and without using the behavioral features. Then, we applied Gradient Random forest and boosted trees using cross validation.

First, we started by comparing the importance of behavioral features (Bf) and impact of the predication accuracy as shown in the following table.

**Table 2** RF vs GBT with and without behavioral features

| Evaluation measure | Random Forest | | Gradient boosted trees | |
|---|---|---|---|---|
| Bf existence | Bf | WBF | Bf | WBF |
| Accuracy | 81.04 | 75.00 | 77.29 | 73.12 |
| Recall | 81.56 | 75.45 | 77.82 | 74.19 |
| Precision | 82.45 | 76.32 | 78.74 | 74.10 |
| F-Measure | 82.00 | 75.88 | 78.28 | 74.14 |

Table 2 shows better results in all cases by using random forest model with behavioral features and without behavioral features. These results are better than results of (Amrieh et al., 2016). Amrieh et al (2016) used decision tree(J48), artificial neural network, and naïve base. They got best accuracy score by ANN WITH 79.1 for BF and 57.0 without behavioral features with only 10 features by using ensemble method. We considered all the features; we did not apply feature selection as they did. Therefore, using behavioral features improved the prediction performance of both models. Now on, we compare between models using behavioral features.

In second part, we used one-hot encoder for the categorical features mentioned above. Random forest (RF) and gradient boosted tree (GBT) are applied. The achieved results are below:

**Table 2** RF and GBT results using demographical features as dummy varaibels.

| Evaluation measure | Demographical features Binary encoded | |
|---|---|---|
| Bf existence | BF-RF | BF-GBT |
| Accuracy | 80.00 | 77.92 |
| Recall | 80.19 | 78.43 |
| Precision | 82.04 | 79.49 |
| F-Measure | 81.10 | 78.96 |

From the table, RF is still the best model in accuracy. we achieved 80% accuracy that is 1% less than encoding all features as numerical. However, GBT achieved a slightly better accuracy with 77.92 compared to 77.29.

In third part, we used one-hot encoder for the categorical features and topic feature. Random forest (RF) and gradient boosted tree (GBT) are applied. The achieved results are below:

| Evaluation measure | Demographical features and topic Binary encoded | |
|---|---|---|
| Bf existence | BF-RF | BF-GBT |
| Accuracy | 78.33 | 77.92 |
| Recall | 77.97 | 78.46 |
| Precision | 80.71 | 79.41 |
| F-Measure | 79.32 | 78.93 |

Table 3 Demographical and Topic encoded as dummy variables

As show in the table, RF scored 78.33 that is less than other previous methods. However, GBT gives same accuracy compared to the previous table (only categorical features).

We also use support vector machine classifier with linear kernel. Here, we arbitrarily used test split for splitting dataset into 80% training set and 20% testing set. We recorded better results with dummy variables as show in the following figure:

| Evaluation measure | SVM | | |
|---|---|---|---|
| Bf existence | All numercal features | Categorical features as dummy variables | Categorical and topic features as dummy variables |
| Accuracy | 0.7 | 0.74 | 0.77 |
| Recall | 0.71 | 0.75 | 0.77 |
| Precision | 0.7 | 0.74 | 0.78 |
| F-Measure | 0.71 | 0.75 | 0.78 |

Table 4 SVM with categorical and topic as dummy variables

Dummy variables encoding has a noticeable impact as shown in table 4. SVM has better accuracy when encoded categorical in addition to topic features. This model supports out hypothesis.

## CONCLUSION

This paper proposes a students' performance prediction model based on trees classifiers (Gradient Boosted trees and random forests). The dataset is provided by (Amrieh, Hamtini and Aljarah, 2016). Also, it is published on Kaggle by the author. This paper achieves better accuracy compared to the original paper by 2% . It compares the performance accuracy when encoding nonordinal features as dummy variables, and when encoding the same features as numerical variables. RF decreased 1% by encoding categorical variables as dummy. GBT has a slight increase in the accuracy by using dummy variables. SVM accuracy increases 7% by encoding nonordinal categorical variables. Therefore, encoding categorical variables affects the prediction accuracy.

## REFERENCES

Al-Barrak, M. and Al-Razgan, M. (2016). Predicting Students Final GPA Using Decision Trees: A Case Study. International Journal of Information and Education Technology, 6(7), pp.528-533.

Amrieh, E., Hamtini, T. & Aljarah, I. (2015). Preprocessing and analyzing educational data set using X-API for improving student's performance. 2015 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT).

Amrieh, E., Hamtini, T. & Aljarah, I. (2016). Mining Educational Data to Predict Student's academic Performance using Ensemble Methods. International Journal of Database Theory and Application, 9(8), pp.119-136.

Ayesha, S., Mustafa, T., Sattar, A. R., & Khan, M. I. (2010). Data mining model for higher education system. European Journal of Scientific Research, 43(1), 24-29.

Bresfelean, V. P. (2007). Analysis and predictions on students' behavior using decision trees in Weka environment. Information Technology Interfaces, 51-56.

Chamasemani, F. & Singh, Y. (2011). Multi-class Support Vector Machine (SVM) Classifiers -- An Application in Hypothyroid Detection and Classification. 2011 Sixth International Conference on Bio-Inspired Computing: Theories and Applications.

Cortez, P., & Silva, A. M. G. (2008). Using data mining to predict secondary school student performance.

Gulati, P. and Sharma, A.(2012). Educational data mining for improving educational quality, Int. J. Comput. Sci. Inf. Technol. Secur., 2(3), pp. 648–650,

Harwati, Alfiani, A. and Wulandari, F. (2015). Mapping Student's Performance Based on Data Mining Approach (A Case Study). Agriculture and Agricultural Science Procedia, 3, pp.173-177.

Kumar, A. D., & Radhika, D. V. (2014). A survey on predicting student performance. International Journal of Computer Science and Information Technologies, 5(5), 6147-6149.

Ramaswami, M., & Bhaskaran, R. (2010). A CHAID based performance prediction model in educational data mining. arXiv preprint arXiv:1002.1144.

Varghese, B. M., Unnikrishnan, A., Sciencist, G., Kochi, N. P. O. L., & Kochi, C. U. S. A. T. (2010). Clustering student data to characterize performance patterns. Int. J. Adv. Comput. Sci. Appl, 2, 138-140.

Varun, D. and Chadha, A. (2011). An Empirical Study of the Applications of Data Mining Techniques in Higher Education. International Journal of Advanced Computer Science and Applications, 2(3).