Categorical Variable Coding for Machine Learning in Engineering Education

Alvin Tran
Computer Science and Engineering
University of Louisville
Louisville, KY, USA
a0tran05@louisville.edu

Christian Zuniga-Navarrete
Industrial Engineering
University of Louisville
Louisville, KY, USA
cszuni01@louisville.edu

Luis Javier Segura
Industrial Engineering
University of Louisville
Louisville, KY, USA
ljsegu01@louisville.edu

Arinan Dourado
Mechanical Engineering
University of Louisville
Louisville, KY, USA
arinan.dourado@louisville.edu

Xiaomei Wang
Industrial Engineering
University of Louisville
Louisville, KY, USA
xiaomei.wang@louisville.edu

Campbell R. Bego*
Engineering Fundamentals
University of Louisville
Louisville, KY, USA
campbell.bego@louisville.edu

***Abstract*—This work-in-progress research paper describes a study of different categorical data coding procedures for machine learning (ML) in engineering education. Often left out of methodology sections, preprocessing steps in data analysis can have important ramifications on project outcomes. In this study, we applied three different coding schemes (i.e., scalar conversion, one-hot encoding, and binary) for the categorical variable of Race across three different ML models (i.e., Neural Network, Random Forest, and Naïve Bayes classifiers) looking at the four standard measures of ML classification models (i.e., accuracy, precision, recall, and F1-score). Results showed that in general, the coding scheme did not affect predictive outcomes as much as ML model type did. However, one-hot encoding – the strategy of transforming a categorical variable with *k* possible values to *k* binary nodes, a common practice in educational research – does not work well with a Naïve Bayes classifier model. Our results indicate that such sensitivity studies at the beginning of ML modeling projects are necessary. Future work includes performing a full range of sensitivity studies on our complete, grant-funded project dataset that has been collected, and publishing our findings.**

***Index Terms*—engineering education, persistence, expectancy-value theory, machine learning**

I. Introduction

Modern machine learning (ML) techniques can process many interrelated factors, which presents interesting opportunities for understanding and predicting complex outcomes in educational research. However, it can be difficult to collect appropriate data and it can take an experienced ML researcher to perform the numerous steps in the analysis pipeline. For example, to use categorical variables such as race, sex, educational level, etc., in ML models, the data must be converted to a numeric value of some kind. As ML technologies become more accessible, it is important to help the education research community understand some preliminary decision-making procedures and their ramifications.

This project is funded by NSF IUSE: EDU, Award #2335725

This paper describes the influence that different categorical variable coding strategies have on ML model performance using a sample engineering persistence dataset. This study begins our work on a funded project that aims to streamline a ML methodology for identifying targeted interventions for students who are predicted to leave engineering. The project goal is to develop a generalized modeling process that can be applied by other institutions such that individualized interventions can be applied on a national scale. Decisions, even in preprocessing, require careful consideration and analysis.

II. Literature Review

A. Machine Learning in Engineering Education Research

The utilization of data-driven methods has facilitated the analysis of student success indicators, including the risk of dropout, attrition risk, and completion risk, which are often correlated with student persistence or retention [1]. In particular, ML models have garnered attention due to their adeptness in handling both quantitative and qualitative/categorical data, yielding better prediction results compared to statistical techniques such as logistic regression and discriminant analysis [2]. The applications of ML techniques mainly focus on two goals: student performance and student retention/attrition prediction.

Several ML techniques have been employed to estimate student performance and identify significant impacting variables. For example, Adejo et al. [3] compared different ML techniques in predicting student performance using data from the student record system, learning management system, and surveys. These techniques included Decision Tree (DT), Neural Network (NN), Support Vector Machine (SVM), and ensemble models, with the ensemble model achieving the highest accuracy of approximately 80%. Slim et al. [4] predicted student success (i.e., GPA score) using a Bayesian belief network model with a margin error of 0.16. Sweeney et al.

[5] utilized regression techniques like Random Forest (RF), *k*-nearest neighbor, and personalized multiple linear regression to predict student grades in upcoming semester courses.

ML techniques have also been employed to predict student retention. For instance, Delen [2] uses NN, DT, SVM, and ensemble model to predict student retention before the sophomore year, achieving an accuracy of approximately 80%. Sensitivity analyses revealed that factors such as fall GPA, loans, and financial aid significantly impact the prediction of student attrition. Raju et al. [6] used ML techniques such as DT and NN to predict student retention. The study identified first-semester GPA, status (full/part-time), earned hours, and high school GPA as factors with a higher impact on prediction.

Furthermore, other studies have developed ML applications focusing on student performance and retention. For instance, Alkhasawneh et al. [7] developed a NN model using demographic, pre-college, and college variables that were emphasized in focus group discussions. They used the model to predict student GPA and retention in Science, Technology, Engineering, and Mathematics (STEM) disciplines. Key factors such as the first math course grade, high school rank, impact of pre-college intervention programs, and SAT math score are found to be useful for predicting both performance and retention.

B. Categorical Variables

Currently, various information sources are collected to evaluate the effect of different factors on student persistence or attrition, such as survey data, ACT scores, etc. [8]. Analyzed factors encompass both numerical attributes (e.g., age, scores, etc.) and categorical attributes (e.g., gender, race, etc.). In particular, categorical variables require encoding into numerical values for integration into ML models. For instance, Aulk et al. [9] employed dummy encoding to map categorical variables in a model predicting student dropout. Similarly, Niyogisubizo et al. [10] utilized one-hot encoding to convert categorical features into binary vectors for an ensemble model predicting student attrition. Pratape et al. [11] applied scalar conversion (i.e., label encoding) for each categorical variable in an educational enrollment status ML model. Although multiple encoding options exist, certain methods may prove inadequate. For instance, scalar conversion can lead to misinterpretation due to assumptions of an ordering relation not universally present in categorical variables [12]. Therefore, it may be necessary to evaluate the impact of encoding methods on ML models' performance to provide guidelines for consistent preprocessing of categorical variables.

C. Current Study

The current study compared three common coding strategies (scalar conversion, one-hot encoding, and binary) of a categorical variable (race) in an engineering education modeling dataset. Accuracy, precision, recall, and F1-scores were calculated for three predictive ML models (NN, RF, and Naïve Bayes, NB, classifiers) using each coding scheme.

Our research questions were as follows:

1) *Does the categorical variable encoding strategy impact the performance of ML classification outcomes?*
2) *Are there any methodology sequences that should be avoided?*

III. Methods

This study was approved by the University of Louisville's Institutional Review Board. The study was retrospective, utilizing data collected from students in past years.

A. Dataset

Participants included in the current study were students who enrolled in the University of Louisville as first-time full-time undergraduate engineering students in Fall 2018 or 2019, took a math course in the fall semester, and completed a survey provided in their principal engineering course (N = 933; 78% Male, 22% Female; 80% White, 6% Asian, 5% Black/African American, 4% Hispanic/Latino, 5% Other). Students with missing data were removed from this exploratory study.

B. Materials

The data in this preprocessing study included the following variables:

1) Demographic Data: gender, race, Pell Grant eligibility
2) Survey Data: individual interest [13], perceived effort, opportunity, and psychological costs [14], perceived academic competence [15], self-efficacy [16],
3) Performance Data: ACT scores (composite, English, math, science reading), term 1 engineering course grades (math, introduction to engineering, and chemistry)
4) Financial Aid: source (federal, state, institutional, private), type (scholarship, loan, grant, work-study), and cause (need, merit)

Variables in this dataset are not described in detail for this Work-In-Progress (WIP) publication, but will be supplied upon request from the corresponding author.

C. Procedures

1) Overview: Different types of variables (i.e., attributes) were first assembled in one data file. The attributes included numerical variables (e.g., ACT scores) and categorical variables (e.g., race and gender). All variables were preprocessed before being fed into the ML models. Numerical attributes were standardized (i.e., converted in variables with mean 0 and standard deviation 1) and the categorical attributes, such as gender, were converted to numerical variables (e.g., binary). Subsequently, the ML models were trained to classify student persistence, where student attrition was labeled as 1 and student persistence was denoted as -1. The classification results were assessed with the ground truth of student persistence using common classification metrics (e.g., accuracy, precision, etc.).

In our analysis, we focused on the race attribute, which included several levels (e.g., White, Asian, etc.). It was transformed by different encoding methods (i.e., scalar conversion, one-hot encoding, and binary encoding). In this study, we

evaluated the effect of the categorical variables' encoding methods on the performance of the ML models.

Our procedures are illustrated in Fig. 1.

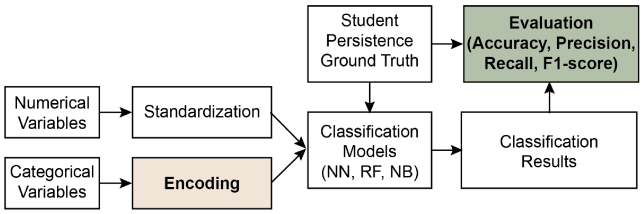

Fig. 1. Scheme of the Proposed Method.

2) Encoding Methods: Three different encoding methods, namely scalar conversion, one-hot encoding, and binary encoding were employed to transform the categorical variable race into numerical values. In the scalar conversion method, an integer is assigned to each race category and the integer values are integrated into the predictors afterwards. The one-hot encoding method produces k binary variables by comparing each level of the categorical variable to a random fixed reference level, being k the number of race levels. The binary encoding method transforms the categories into an ordinal variable, and then this variable is converted into binary code. The binary string is split into separate numerical attributes. Table I shows the corresponding representation of the race categories for the three different encoding methods.

TABLE I
ENCODED REPRESENTATION OF THE ATTRIBUTE 'RACE'

Race	Encoding representation		
	Scalar	Binary	One-hot
White	1	{-1,-1,-1}	{0,0,0,0,0,1}
Two or more	2	{-1,-1,1}	{0,0,0,0,1,0}
Black/African American	3	{-1,1,-1}	{0,0,0,1,0,0}
Asian	4	{-1,1,1}	{0,0,1,0,0,0}
Non-Resident	5	{1,-1,-1}	{0,1,0,0,0,0}
Hispanic/Latino	6	{1,-1,1}	{1,0,0,0,0,0}

3) Classification Models: Three classification models were explored, namely NN, RF, and NB classifiers.

NN is a classification model in which intermediate layers (i.e., hidden layers) between the input (i.e., attributes) and the output (i.e., student persistence) are created. Each layer increases the complexity of the model. Model outcomes are represented as $C = \text{sgn}(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{z})$, where β_0 and $\boldsymbol{\beta}$ are model parameters, and $\boldsymbol{z} = (z_1, z_2, \cdots, z_M)$ is the hidden layer conformed by M hidden units. The hidden units z_m are created from linear combinations of the input variables $\boldsymbol{x}$ via $z_m = \tanh(\alpha_0 + \boldsymbol{\alpha}_m^T \boldsymbol{x}), m = 1, ..., M$, where α_0 and $\boldsymbol{\alpha}_m^T$ are model parameters, and $\tanh$ is the hyperbolic tangent function. The NN is trained by minimizing the error between real outcome values (i.e., ground truth) and predicted values, and performing a back-propagation method to obtain the optimal model parameter values [17].

RF is a classification model that builds a large collection of decorrelated trees and averages them. These trees are created through a random selection of the attributes and samples. Two tuning parameters are involved while building the RF: the number of grouped variables m, and the number of trees B. The RF model is described by $\hat{C}_{RF}^B(\boldsymbol{x}) = majority\ vote\{\hat{C}_b(\boldsymbol{x})\}_1^B$, where $\hat{C}_b(\boldsymbol{x})$ is the class prediction of the bth RF tree and $\boldsymbol{x}$ represent the predictors [17].

NB is a statistical classification technique based on Bayes' theorem. NB classifier assumes that the predictors are independent of each other such that $P(\boldsymbol{x}|C) = \Pi_{i=1}^d P(x_i|C)$, where $\boldsymbol{x} = (x_1, x_2, ..., x_d)$ are the predictors and C the student persistence value. Consequently, the joint probability $P(C, \boldsymbol{x})$ is calculated as $P(C, \boldsymbol{x}) = P(C)\Pi_{i=1}^d P(x_i|C)$. Thus, the NB classified predicts an outcome C for a new predictor $\boldsymbol{x}$ by selecting $\arg\max_C(\hat{P}(C)\Pi_{i=1}^d \hat{P}(x_i|C))$, where $\hat{P}(C)$ and $\hat{P}(x_i|C)$ are the estimation of the outcome probability and the conditional probabilities of the predictors $\boldsymbol{x} = (x_1, x_2, ..., x_d)$ [18].

4) Traning and Evaluation: During the training of ML classification models, the samples were divided into ten randomly generated and equally sized folds for Cross-Validation (CV). Nine out of the ten folds were utilized for model training, while the remaining fold was allocated for model testing. This evaluation process was iterated until all folds had been considered in the testing dataset. The entire procedure was repeated 100 times to analyze the consistency of the models. The classification results during each CV were compared with the ground truth student persistence to evaluate the performance of the models, as shown in Fig 1.

We employed four classification metrics, namely:

1) *accuracy*= $\frac{TP+TN}{TP+TN+FP+FN}$,
2) *precision*= $\frac{TP}{TP+FP}$,
3) *recall*= $\frac{TP}{TP+FN}$, and
4) *F1-score*= $\frac{2TP}{2TP+FP+FN}$,

where TP is true positive (i.e., correct prediction of student attrition), TN is true negative (i.e., correct prediction of student persistence), FP is false positive (i.e., predict student attrition when the student did not leave the program), and FN is false negative (i.e., predict student persistence when the student left the program).

Accuracy represents the percentage of correct predictions. In binary classification with imbalanced classes, such as this study, accuracy measurements can lead to misleading conclusions due to their inherent bias toward favoring the classification of the majority class [19]. On the other hand, precision measures the ratio between the correct predictions of student attrition and all students correctly predicted. Similarly, recall calculates the ratio of the correct predictions of student attrition to the total true number of students who have dropped out. The F1-score represents the harmonic mean between the precision and recall metrics.

IV. RESULTS

The metrics obtained from the 100 iterations of the ten-fold CVs are depicted in Fig. 2. The metrics are represented as a box plot using each combination of encoding methods

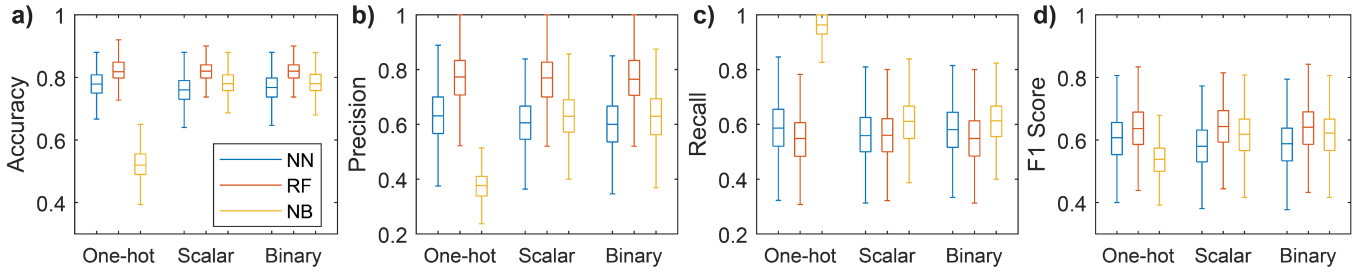

Fig. 2. Evaluation Metrics for the Three Encoding Methods and Three Classification Models: a) Accuracy, b) Precision, c) Recall, and d) F1-score.

and classification models employed in this study. Fig. 2 (a) shows that the highest accuracy is achieved with the RF model, regardless of the encoding method, reaching a mean of ~ 0.81, for all iterations. It is closely followed by the NB and the NN accuracy.

A particular low accuracy (i.e., ~ 0.5) is obtained with the NB model when the race is one-hot encoded (see Fig. 2a). The NB classifier assumes that all predictors (features) are conditionally independent given the class label. This means that once we know the class label, the presence or absence of any particular feature should not influence the presence or absence of any other feature. But, when categorical data is encoded using one-hot encoding strategy, each category within a feature is transformed into its own binary feature. This expansion results in a set of binary features where each feature represents the presence or absence of a specific category within the original feature (e.g., race). Consequently, this encoding strategy introduces dependencies among the binary features derived from the same original categorical feature, as they cannot coexist within the same observation. Due to this encoding process, the assumption of independence among predictors is violated. In other words, the binary features derived from the same categorical variable are not independent of each other, as they are mutually exclusive within each observation. This violation of the independence assumption can help explain the worsened performance of the NB model, as it relies heavily on this assumption for its probabilistic calculations.

Fig. 2b shows that the RF precision is $\sim$ 0.78, which is notably superior compared to the NN and NB classifiers. Notice also that the NB precision is considerably low at around 0.4, which is attributed to the violation of the independence assumption discussed earlier. From Fig. 2 (c) it is observed that all combinations of encoding methods and ML models exhibit similar recall values, around 0.57, except for the one-hot NB case, which achieves recall values close to one. This implies that such a model is more sensitive to student attrition's FN (i.e., predicting student persistence when the student has discontinued his/her studies). The high recall and low precision values using the one-hot NB combination are attributed to a large amount of FP predictions and few FN predictions, as shown in Table II. Furthermore, the mean F1-score of all ML models is approximately 0.6, as depicted in Fig 2 (d). The precision, recall, and F1-score values present more variability across the CV iterations compared to the accuracy variability for all models, as shown in Fig. 2.

TABLE II
CONFUSION MATRIX OF THE 1ST CV ITERATION USING THE ONE-HOT NB MODEL

Ground Truth	**Prediction**	
	Attrition	Persistence
Attrition	279	12
Persistence	469	235

V. DISCUSSION & CONCLUSIONS

In this study, the impact of different coding strategies for categorical data, such as race, on the performance of ML models for engineering persistence prediction was evaluated. Three different coding strategies, namely scalar conversion, one-hot encoding, and binary encoding, were assessed considering three distinct statistical classifiers, i.e., NN, RF, and NB. A significant finding was that the one-hot encoding strategy drastically reduced the performance of the NB model, including, accuracy, precision, and F1-score. Such behavior can be explained by the fact that the one-hot encoding strategy violates the main statistical assumption of the NB model, the assumption of independence between considered features. Therefore, the observed deterioration in performance when using one-hot encoding with the Naïve Bayes Classifier on categorical data can be attributed to the conflict between the encoding strategy and the Naïve Bayes assumption of predictor independence. This combination of encoding type and modeling choice should be avoided in engineering education research. This highlights the importance of considering appropriate encoding strategies that align with the underlying assumptions of the classification model utilized.

Overall, our results showed that though coding strategies for categorical variables did not affect predictive outcomes as much as ML model types, certain coding strategies can lead to violation of model assumptions and impact model performance, thus, it is meaningful to test different preprocessing methodologies rather than arbitrarily selecting one popular method. Correspondingly, our next step is to work on a larger dataset to keep testing different categorical data coding strategies, but we will not continue with the Naive-Bayes and one-hot-encoding modeling combination. We will

instead use a Bayes classifier with a prior distribution, and in addition, work to reduce the number of correlated nodes. We will similarly test continuous variable standardization or normalization methods.

REFERENCES

[1] A. M. Williford and J. Y. Schaller, "All retention all the time: How institutional research can synthesize information and influence retention practices," in *Proceedings of the 45th annual forum of the association for institutional research*. Citeseer, 2005.

[2] D. Delen, "A comparative analysis of machine learning techniques for student retention management," *Decision Support Systems*, vol. 49, no. 4, pp. 498–506, 2010.

[3] O. Adejo and T. Connolly, "Predicting student academic performance using multi-model heterogeneous ensemble approach," *Journal of Applied Research in Higher Education*, vol. 10, pp. 00–00, 12 2017.

[4] A. Slim, G. L. Heileman, J. Kozlick, and C. T. Abdallah, "Predicting student success based on prior performance," in *2014 IEEE symposium on computational intelligence and data mining (CIDM)*. IEEE, 2014, pp. 410–415.

[5] M. Sweeney, H. Rangwala, J. Lester, and A. Johri, "Next-term student performance prediction: A recommender systems approach," *arXiv preprint arXiv:1604.01840*, 2016.

[6] D. Raju and R. Schumacker, "Exploring student characteristics of retention that lead to graduation in higher education using data mining models," *Journal of College Student Retention: Research, Theory & Practice*, vol. 16, no. 4, pp. 563–591, 2015.

[7] R. Alkhasawneh and R. H. Hargraves, "Developing a hybrid model to predict student first year retention in stem disciplines using machine learning techniques," *Journal of STEM Education: Innovations and Research*, vol. 15, no. 3, 2014.

[8] C. Mason, J. Twomey, D. Wright, and L. Whitman, "Predicting engineering student attrition risk using a probabilistic neural network and comparing results with a backpropagation neural network and logistic regression," *Research in Higher Education*, vol. 59, pp. 382–400, 2018.

[9] L. Aulck, N. Velagapudi, J. Blumenstock, and J. West, "Predicting student dropout in higher education," *arXiv preprint arXiv:1606.06364*, 2016.

[10] J. Niyogisubizo, L. Liao, E. Nziyumva, E. Murwanashyaka, and P. C. Nshimyumukiza, "Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization," *Computers and Education: Artificial Intelligence*, vol. 3, p. 100066, 2022.

[11] G. Pratape, K. R. Meesala, S. Panda, and P. Goyal, "Predicting graduation and dropout rates: A machine learning approach," in *2023 International Conference on Advanced Computing & Communication Technologies (ICACCTech)*. IEEE, 2023, pp. 603–609.

[12] A. Von Eye and C. C. Clogg, *Categorical variables in developmental research: Methods of analysis*. Elsevier, 1996.

[13] Linnenbrink-Garcia *et al.*, "Measuring situational interest in academic domains," *Educational and psychological measurement*, vol. 70, no. 4, pp. 647–671, 2010.

[14] T. Perez, J. G. Cromley, and A. Kaplan, "The role of identity development, values, and costs in college stem retention." *Journal of educational psychology*, vol. 106, no. 1, p. 315, 2014.

[15] J. Crocker *et al.*, "Contingencies of Self-Worth in College Students: Theory and Measurement," *Journal of Personality and Social Psychology*, vol. 85, pp. 894–908, 2003.

[16] P. R. Pintrich *et al.*, *A manual for the use of the Motivated Strategies for Learning Questionnaire (MSLQ)*. ERIC, 1991.

[17] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2009, vol. 2.

[18] S. Chen, G. I. Webb, L. Liu, and X. Ma, "A novel selective naïve bayes algorithm," *Knowledge-Based Systems*, vol. 192, p. 105361, 2020.

[19] V. García, R. A. Mollineda, and J. S. Sánchez, "A bias correction function for classification performance assessment in two-class imbalanced problems," *Knowledge-Based Systems*, vol. 59, pp. 66–74, 2014.