Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 44 No. 6 (2023)

Evaluating the Effectiveness of Categorical Encoding Methods on Higher Secondary Student's Data for Multi-Class Classification

[1]P. Amutha ,[2]Dr. R. Priya

[1]Research Scholar, Dept. of Computer Science

VISTAS, India

[2]Professor, Dept. of Computer Science

VISTAS, India

***Abs*tract**

The multi-class classification is a prominent research area focused by researchers and academicians to classify data labels in various fields such as crop yield, health care, accounting, finance, agriculture, bioinformatics, cyber security, cloud computing, simulation, education, etc., to solve problems, alleviate risks and to possess new opportunities. The encoding of categorical data into the numeric values trigger off the field of data mining-machine learning. Because, algorithms in these domains are difficult to understand the string values in data set and produce poor performance in classification. This research study focused on converting categorical data into numeric values using encoding methods. Various classifiers were considered for performance comparison of the encoding methods. The experimental result was compared in terms of accuracy, precision, F1 score, and recall. It was revealed that the combination of Random Forest and label encoding outperformed other classification methods for multiclass classification. .

Keywords: *multi-class, data mining, machine learning, encoding, random forest, classification.*

I. Introduction

Data-mining is a method of consolidating knowledge from huge data sources and predicting the upcoming trends and behaviors so as to take decisions proactively. Data mining approaches have been applied in various fields of accounting, finance, agriculture, health care, bioinformatics, cloud computing, simulation, education, etc., to redress issues and risks to foresee emerging opportunities. In this context, educational data mining (EDM) facilitates bringing out the current scenarios in the field of education [6].

As a sub-field of data mining, EDM is employed to analyse educational data to unravel hidden patterns that could assist in promoting decision-making with better accuracy. Data mining-machine learning techniques play a significant role to improve academic and administrative qualities in education. EDM is applied to predict students' performance-based admission requirements [8], risk of academic failure [9], predict students' performance [6], create a feedback system [7], to draft a framework for higher course selection [2], to predict students' abilities [11] as well as recommendation systems to the students that could assist them to select suitable programmes in their graduation in university [10][15].

In EDM, techniques of prediction, sequential pattern, clustering, regression, classification and association rule analyses are involved to detect the problems in the education sector [13]. One of the most popular techniques is classification which is supervised learning used to classify and predict data labels effectively and efficiently. Generally, there are four kinds of classification techniques in data mining which include machine learning namely, Binary Classification, Multi-Class Classification, Multi-Label Classification and Imbalanced Classification. This current research work focused on preprocessing methods such as encoding categorical data into the numeric values for multi-class classification to determine the best encoding method for purely categorical data sets.

II. Related Work

A novel framework was developed to find out the final test (FT) yield at the water fabrication stage for semiconductor manufacturing Final test was performed using machine learning approaches [3]. They introduced Gaussian mixture models, one hot encoder and another label encoder for handling both numerical and categorical data in preprocessing stage. Further, model selection and model ensemble was adopted along with F1-Macro to enhance the classification performance in binary and multi-class classification. The findings concluded that the developed novel framework was engaged in various water technologies and manufacturing flows of mass production. Jawthari, Moohanad, and VeronikaStoffova [4] evaluated the effectiveness of categorical encoding on students' data to improve academic performance. They applied dummy variables encoding one-hot on non-ordinal attributes. Random forest, Gradient Boosted tree and Super Vector Machine classifiers experimented on encoded data. A better outcome was determined by Random Forest; Gradient Boosted trees in all cases yielded similar findings and SVM performance was improved in the dummy variables model. Do Thi Thu Hien et al., [5] focused to design a model using deep learning and encoded categorical values in predicting students' academic performance. They investigated three encoding methods along with deep-learning techniques namely Deep- Dense Neural and Long-Short memory neural networks in academic performance prediction for the students in Vietnam. The popular encoding techniques Label, One Hot and Learned Embedding encoding were adopted for the experiment. It was shown that the learned embedding encoding and long Short-term memory combination performed well compared to other combinations to predict students' academic performance. A comparative analysis was conducted to find out the solution for forecasting the real estate objects [12]. The One-Hot encoding was considered for encoding categorical attributes and regression algorithms adopted in the comparison process. It focused on real estate type, area, and building structure quality as objects' internal things in comparison. The analysis brought out that the real estate objects' prices were modeled using regression analysis. Additionally, the finding suggested that external factors also had a significant impact to determine the real estate value to be a future research goal. A comparative analysis was conducted to study the encoding methods for categorical variables using Artificial Neural Networks for the classification of the Car data set [14]. Seven categorical techniques namely One Hot Coding, Ordinal Coding, Sum Coding, Helmert Coding, Polynomial Coding, Backward Difference Coding, and Binary Coding were applied in preprocessing the Car data set. It was found that the Sum and Backward difference encoding methods produced high accuracy than the other five techniques and was concluded that the two techniques (Sum and Backward) were most the preferred in encoding categorical data for prediction processes.

Iii. Research Methodology

The key processes of the research study focused on handling categorical data in students' data set and classifying the higher education courses using Data mining-Machine learning techniques. The sequential stages of the present study are shown in Figure 3.1.

Data collection

The post-secondary students' data collected using a well-defined questionnaire was involved in the assessment. The questionnaire of the study was prepared based on inputs from previous research work as well as suggestions from academicians. The data set consisted of 33 attributes with no target variables and 925 instances initially. Table 3.1 presents the students' data set after data preprocessing.

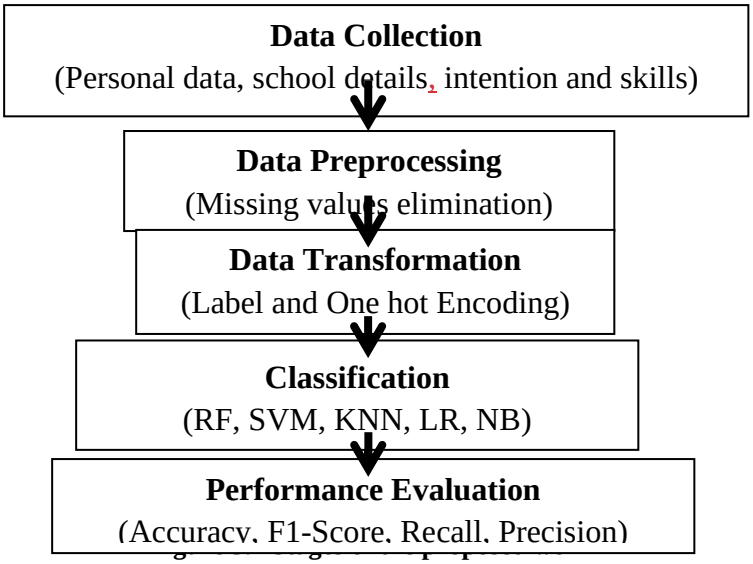

Table 3.1 Description of Student's data set after Pre-processing

S.NO	Description	Type	No of Values
1	20 Attributes	Object	More than two options
2	8 Attributes	Boolean	Two options
3	Target variable ABBHEC	Object	54 class labels

Data Preprocessing

Data preprocessing is one of the most significant tasks of Data mining–Machine learning approaches. Missing values in Students' data were detected as well as removed for smoothening classification using machine learning. A prediction algorithm was used to identify the target variable as the target variable is important for supervised classification.

Data Transformation

The machine learning performance could be affected by many factors; the type of data used to feed into the machine learning algorithms could be one of the factors. As such, encoding categorical data is a crucial process because it converts data into categorical variables which can be understood by models in machine learning. One hot and Label encoding are the most common practices to deal with categorical data.

In this study, all the features in the data set were categorical data with more than two options. The most popular categorical encoding methods one-hot encoding and Label encoding applied for converting categorical attributes of the student data set.

One-hot encoding created a high dimensionality compared to label encoding. Besides, label encoding produced high performance in classification than one-hot encoding. Label encoding was applied as it does not create high dimensionality of data instead only assigns a numerical value from the 0-maximum option in a particular column.

Classification

Classification is a supervised method used to classify the class label. In this analysis, Random Forest, Support vector machine, K- nearest neighbor, Naïve Bayes, and Logistics Regression were considered as base classifiers to classify higher education programmes in students' data set. In classification, 70% of student's data was considered for training-data and 30% of student's data was considered for test data.

Performance Evaluation

Performances of classifiers are evaluated in terms of accuracy, f1-score, recall, and precision.

Accuracy rate (AC): the correct predictions in percentage

AC= TN + TP/ TP + FP + FN + TN, (1)

Precision (P): It returns fraction of the positive observations which are predicted correctly from the total number of predicted positive observations.

P = TP/TP + FP, (2)

Recall (R): It returns fraction of the positive observations which are predicted correctly from whole observations in the class.

R =TP/ TP + FN, (3)

F-measure: 'Precision and Recall criteria can be interpreted together rather than individually. F provides both the level of accuracy of the classification and its robustness (less data loss)'.

F − measure =2 × P × R P + R, (4)

IV. Result And Discussion

In this study, preprocessing, encoding data and data mining approaches were carried out using Python 3.3 SciKit Learn packages under Spider IDE 4.1.5. The system specifications were Dell Inspiron, 12 GB RAM, Intel(R) Core i5- 5200U and 64 bit Windows 10. For converting the categorical into numeric values, both one-hot and label encoding methods were applied on the data set. Converted data were sent through the various classifiers and comparison was performed with respect to the accuracy, F1-score, recall and precision with less training time consumption. Tables 4.1, 4.2, 4.3 and 4.4 have shown the result of classification using various metrics respectively.

Table 4.1 Accuracy

CLASSIFIER	LABEL	OHEHOT
RF	90	88
SVM	38	88
KNN	56	61
LR	66	85
NB	82	66

Table 4.2 F1-Score

CLASSIFIER	LABEL	OHEHOT
RF	88	87
SVM	38	88
KNN	56	60
LR	67	85
NB	81	67

Table 4.3 Recall

CLASSIFIER	LABEL	OHEHOT

Table 4.4 Precision

CLASSIFIER	LABEL	OHEHOT

RF	89	88
SVM	38	88
KNN	56	61
LR	66	85
NB	82	66

RF	88	88
SVM	41	92
KNN	56	67
LR	73	89
NB	88	77

Figures 4.1, 4.2, 4.3 and 4.4 have shown the findings of various classifies applied in encoded data. The experiment determined that the Random Forest with Label encoding outperformed than other classifier in the classifying multiple class label in students' data. Table 4.5 has shown time utilized by classifiers during training.

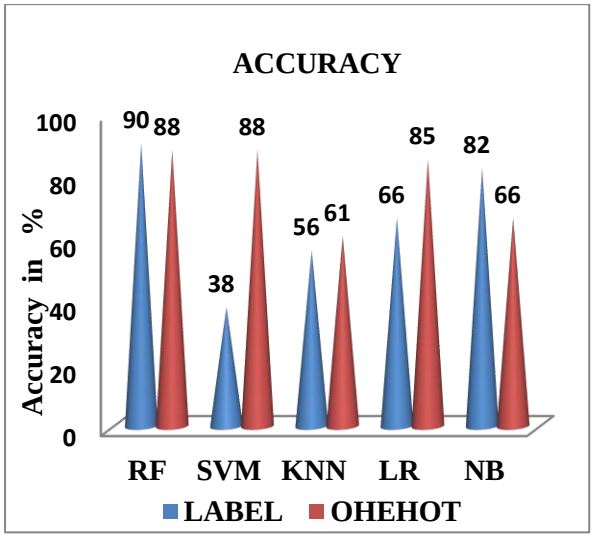

Figure 4.1 Accuracy Comparison

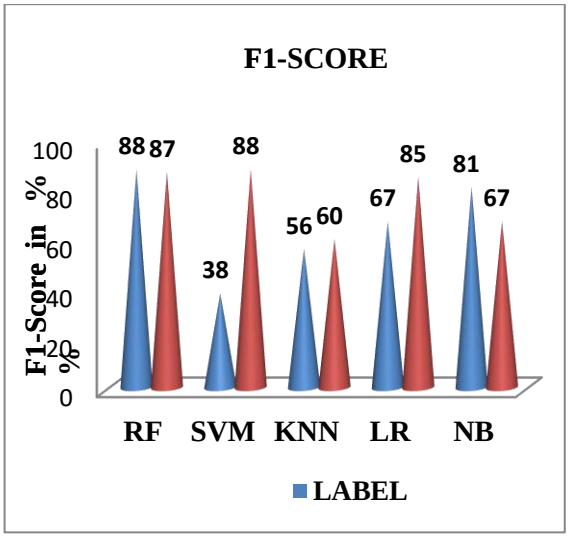

Figure 4.2 F1- Score Comparison

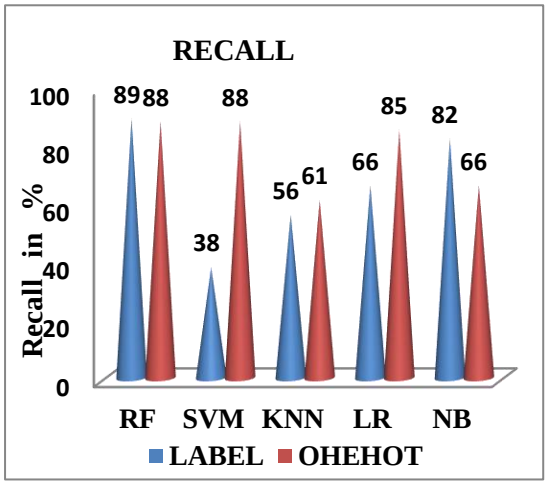

Figure 4.3 Recall Comparison

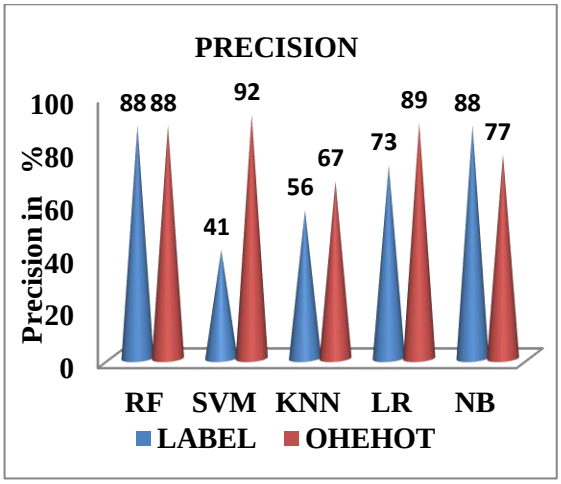

Figure 4.4 Precision Comparison

Table 4.5 Training Time in msec

CLASSIFIER	LABEL	OHEHOT
RF	0.56	0.88

SVM	107.02	3.23
KNN	0.075	0.6
LR	2.85	5.7
NB	0.078	0.078

V. Conclusion

The assessment explored a comparison of classification accuracy using various classification methods in encoded data of the categorical attributes. The key aim of the evaluation was to identify the preference of encoding method for students' data set which includes purely categorical features. The data set applied in the study had 53 class labels in target features. Two encoding techniques were adopted in the experiment for a finding preferable encoding method for multi-class classification. The classification metrics F1-score, recall, precision, and accuracy were evaluated among one hot and Label encoding. It was shown conclusively that Label encoding with Random Forest produced higher in the considered metrics and suggested that Label encoding with Random Forest could be the effective method for multiclass classification.

References

[1] Alshehri, Eman, et al. "A Comparison of EDM Tools and Techniques." International Journal of Advanced Computer Science and Applications, Vol. 11, No. 12, 2020

[2] P. Amutha, R. Priya,, Conceptual Course Selection Framework for Post-Secondary Students' Enrolment in Indian Universities and Colleges, Journal of Advanced Research in Dynamical & Control Systems, Vol. 12, 03-Special Issue, 2020

[3] D. Jiang et al., Novel Framework for Semiconductor Manufacturing FT Yield Classification Using Machine Learning Techniques, IEEE open Access, Volume 8, 2020. DOI: 10.1109/ACCESS.2020.3034680

[4] Jawthari, Moohanad, and VeronikaStoffova, Effect Of Encoding Categorical Data On Student's Academic Performance Using Data Mining Methods, The 16th International Scientific Conference eLearning and Software for Education Bucharest, April 23-24, 2020 521-526.

[5] Do Thi Thu Hien et al., Optimize the Combination of Categorical Variable Encoding and Deep Learning Technique for the Problem of Prediction of Vietnamese Student Academic Performance, International Journal of Advanced Computer Science and Applications, Vol. 11, No. 11, 2020.

[6] Sultana, Jabeen, M. Usha, and M. A. H. Farquad. "Student's performance prediction using deep learning and data mining methods." Int. J. Recent Technol. Eng.(IJRTE)(1S4) (2019): 1018-1021.

[7] Hassan, Muhammad Awais, et al. "An adaptive feedback system to improve student performance based on collaborative behavior." Ieee Access 7 (2019): 107171-107178.

[8] Adekitan, Aderibigbe Israel, and EtinosaNoma-Osaghae. "Data mining approach to predicting the performance of first year student in a university using the admission requirements." Education and Information Technologies 24.2 (2019): 1527-1543.

[9] Sarra, Annalina, Lara Fontanella, and Simone Di Zio. "Identifying students at risk of academic failure within the educational data mining framework." Social Indicators Research 146.1 (2019): 41-60.

[10 Kurniadi, D., et al. "A proposed framework in an intelligent recommender system for the college student." Journal of Physics: Conference Series. Vol. 1402.No. 6.IOP Publishing, 2019.

[11] Yang, Fan, and Frederick WB Li. "Study on student performance estimation, student progress analysis, and student potential prediction based on data mining." Computers & Education 123 (2018): 97-108.

[12] Parygin, D. S., et al. "Categorical data processing for real estate objects valuation using statistical analysis." Journal of Physics: Conference Series. Vol. 1015.No. 3.IOP Publishing, 2018.

[13] P. Amutha, R. Priya, A survey on educational data mining techniques in predicting student's academic performance, International Journal of Engineering & Technology, 7 (2.33) (2018) 634-636

[14] Potdar, Kedar, Taher S. Pardawala, and Chinmay D. Pai. "A comparative study of categorical variable encoding techniques for neural network classifiers." International journal of computer applications 175.4 (2017): 7-9.

[15] R. Sumitha, E. Vinothkumar, and P. Scholar, "Prediction of students outcome using data mining techniques," International Journal of Scientific Engineering and Applied Science (IJSEAS)–Volume-2, Issue-6, 2016.