

# Categorical encoder based performance comparison in pre-processing imbalanced multiclass classification

Wiyli Yustanti<sup>1,2</sup>, Nur Iriawan<sup>1</sup>, Irhamah<sup>1</sup>

<sup>1</sup>Department of Statistics, Faculty of Science and Data Analytics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

<sup>2</sup>Department of Informatics, Faculty of Engineering, Universitas Negeri Surabaya, Surabaya, Indonesia

## Article Info

### Article history:

Received Nov 29, 2022

Revised May 9, 2023

Accepted May 27, 2023

### Keywords:

Categorical encoding

Classification

Imbalanced

Multiclass

Performance analysis

## ABSTRACT

The contribution of this study is to offer suggestions for coding techniques for categorical predictor variables and comprehensive test scenarios to obtain significant performance results for imbalanced multiclass classification problems. We modify scenarios in the data mining process with the sample, explore, modify, model, and assess (SEMMA) framework coupled with statistical hypothesis testing to generalize the model performance evaluation conclusions as enhanced-SEMMA. We selected four open-source data sets with unequal class distributions and categorical predictors. Ordinal, nominal, dirichlet, frequency, target, leave one, one hot, dummy, binary, and hashing encoder methods are used. We use the grid-search technique to find the best hyperparameters. The F1-Score and area under the curve (AUC) are evaluated to select the optimal model. In all datasets with 10-fold stratified cross-validation and 95% to 99% accuracy for each dataset, the results show that support vector machine (SVM) outperforms the decision tree (DT) K-nearest neighbor (KNN), Naïve Bayes (NB), logistic regression (LR), and random forest (RF) algorithms. Probability-based or binary encodings, such as target, Dirichlet, dummy, one-hot, or binary, are best for situations with less than 3% of minor class proportions. Nominal or ordinal encoders are preferred for data with a minor class proportion of more than 3%.

*This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.*



## Corresponding Author:

Nur Iriawan

Department of Statistics, Faculty of Science and Data Analytics, Institut Teknologi Sepuluh Nopember  
Sukolilo Campus, Surabaya, Indonesia

Email: nur\_i@statistika.its.ac.id

## 1. INTRODUCTION

Classification algorithms are included in the supervised learning method in the context of machine learning. Supervised learning is machine learning with a function that maps input to output based on examples of input-output pair data [1]. This function results from learning a series of labeled training datasets [2]. This training data consists of vector input objects and the desired output value (target). The algorithm used in supervised learning will analyze the training data to produce the classifier function used to map new input data (testing data) to the correct output data. Therefore, the main goal of this machine learning is to get the optimal scenario that can correctly define the class label for the new object. Many factors affect the performance results of object classification, including data types (numeric, categorical, or mixed), data dimensions, sample size, and feature engineering. The feature engineering process could include handling missing values, data transformation, dimension reduction, feature selection, outlier detection, and resampling with under or oversampling if the class distribution is unequal. The classification performance is also influenced by the model selected, the hyperparameter optimization approach, the model validation, and the evaluation techniques. The studies related to the performance comparison of classification algorithms can be found in [3]–[6]. The results

of those studies concluded that support vector machine (SVM) showed better accuracy than other algorithms. Furthermore, related categorical features for prediction models are also found in [7], [8].

The quality of the data and the amount of helpful information are key factors in determining how well machine learning algorithms perform the learning process. Therefore, it is crucial to correctly transform categorical variables before entering such data into machine learning algorithms. It is well known that most classification methods in machine learning require predictor variables of the numeric type. There are several ways that can be used to perform the transformation or encoding from categorical to numeric data. Hancock and Khoshgoftaar [9], there are three data transformation models from categorical to numeric, such as the determined, algorithmic, and automatic techniques. Firstly, the determined approach is a way to transform categorical data by converting categorical data to numeric vectors with low computational complexity. Examples of this technique include ordinal, nominal, target, leave one out, hashing, frequency [10], binary, dummy, one hot [11], [12] and dirichlet [13]–[15].

Second, the algorithmic technique is a categorical encoding method that requires a large computational process, such as latent dirichlet allocation (LDA) for document data [10] and enabling deep learning for generic data classification (EDLT) for tabular data [16]. Lastly, the third is an automatic encoding technique that combines data representation search methods into the machine learning process. Automated methods are more attractive because they are more general-purpose than algorithmic techniques. For example, the algorithm Word2Vec [17]. In this study, ten types of deterministic encoding techniques will be used, namely nominal encoding (NE), ordinal encoding (OE), target encoding (TE), frequency encoding (FE), dirichlet encoding (DRE), leave one out encoding (LE), one-hot encoding (OHE), hashing encoding (HE), binary encoding (BE) and dummy encoding (DE).

Class imbalance is also a common problem in machine learning classification, with a disproportionate ratio in each class. These can be found in medical diagnostics, spam filtering, fraud detection, and emotion classification. For example, the detection of fraud in banking is rare, only 1%. Most machine learning algorithms need to work more effectively with unbalanced data sets. Applying inappropriate evaluation metrics to a model with unbalanced data can be misleading. If a model with exceptional accuracies, such as 99.8% in the majority class and 0% in the minority class, cannot provide valuable information for predicting rare events. What is needed is the ability of the model to predict rare (minority) events. Thus, accuracy could be more appropriate for evaluating unbalanced datasets. In this case, another alternative model evaluation metric that can be used is the F1-Score [18], [19]. Another way is to create a balanced dataset by resampling [20], [21]. Two approaches that can be taken are undersampling and oversampling techniques. Under-sampling is done by reducing the amount of data from the majority class. Meanwhile, oversampling is done by increasing the data from the minority class [22], [23]. Another approach to dealing with the class unbalanced problem is to use stratified k-fold cross-validation (SCV) as an extension of the cross-validation technique. It is usually used for imbalanced classification problems. This method keeps the class ratio in the k-fold as the ratio in the original dataset [24].

Based on the data type factor and class imbalance condition, this research is focused on cases where the input data type is categorical, and the response variable (target) has more than two classes with unequal proportions. The main research problem is determining a categorical data transformation method to improve classification prediction performance using the F1-Scores and area under the curve (AUC) score evaluation measures. The best model selection scenario is developed by modifying the sample, explore, modify, model, and assess (SEMMA) framework [25], which stands for sample, explore, modify, model, and assess experiments, developed by SAS enterprise miner. Next, an experimental design will be used to determine the comparative performance of ten categorical data encoding methods, six classification algorithms, and four public data sets. The research results will provide recommendations on selecting encoding methods for categorical data in cases of unbalanced multiclass data classification and comprehensive test scenarios to obtain accurate results.

## 2. METHOD

This section will explain the structure of the research methodology used. As stated in the previous section, the SEMMA method is one of the most popular data mining techniques [25]. In order to produce valid evaluation results in this study, it is necessary to develop the SEMMA procedure. The SEMMA framework will be coupled with statistical testing procedures to generalize the model performance evaluation conclusions. This methodology is referred to as enhanced-SEMMA. Figure 1 illustrates the flowchart of the enhanced-SEMMA method.

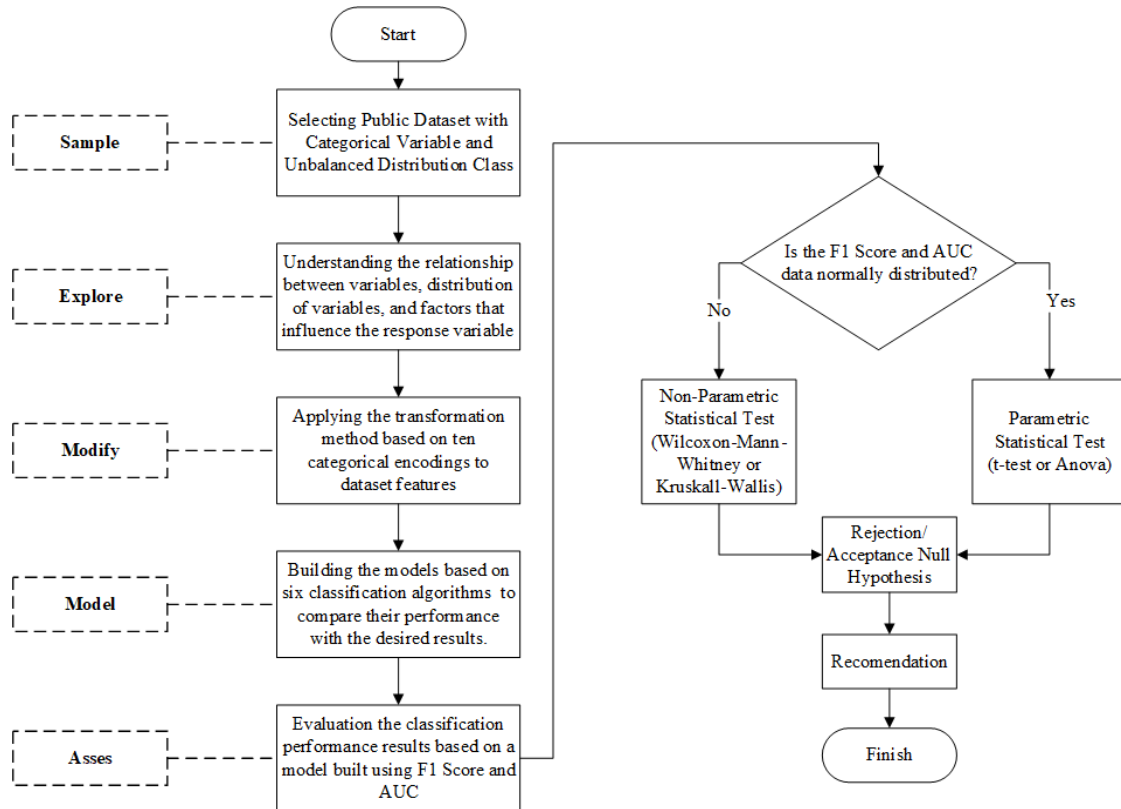


Figure 1. The proposed methodological framework (the enhanced-SEMMA)

**2.1. Sample**

This phase involves selecting the appropriate volume dataset subset from a large dataset provided for model construction. It will help us construct the model more effectively. This phase identifies the independent and dependent variables. The selected subset of data should be representative of the entire dataset initially collected, which means it should contain enough information to retrieve. Additionally, the data is separated for training and validation purposes. The data used in this study consisted of four public datasets taken from the UCI machine learning repository, namely car evaluation, nursery, lymphography, and balance scale data. The structure and characteristics of each dataset can be seen in Table 1.

Table 1. Public dataset with imbalanced class

Dataset	Features type	Features number	Class category	Class proportion (%)					N
				1	2	3	4	5	
Car evaluation	Categorical	6	4	70.02	22.22	3.99	3.76	-	1728
Lymphography	Categorical	18	4	1.35	54.73	41.22	2.70	-	148
Nursery	Categorical	8	5	33.33	0.02	2.53	32.92	31.20	12960
Balance scale	Categorical	4	3	7.84	46.08	46.08	-	-	625

**2.2. Explore**

During this phase, endeavors are undertaken to comprehend the gaps in the data and its interconnections. Univariate and multivariate analysis are two fundamental activities. The univariate analysis involves examining each variable in isolation to comprehend its distribution, while multivariate analysis entails investigating the interrelationships among variables. The utilization of data visualization is prevalent in enhancing the comprehension of data. At this stage, a comprehensive evaluation is conducted on all the variables that impact the final result. The proportion of each class in all datasets can be explained visually through a histogram, as shown in Figure 2. Based on the histogram in Figure 2, the smallest proportions for the car evaluation, lymphography, balance scale, and nursery datasets were 3.76%, 1.35%, 7.84%, and 0.02%, respectively. It can be said that nursery, lymphography, car evaluation and balance scale dataset have severe imbalanced class with different proportion of minority class.

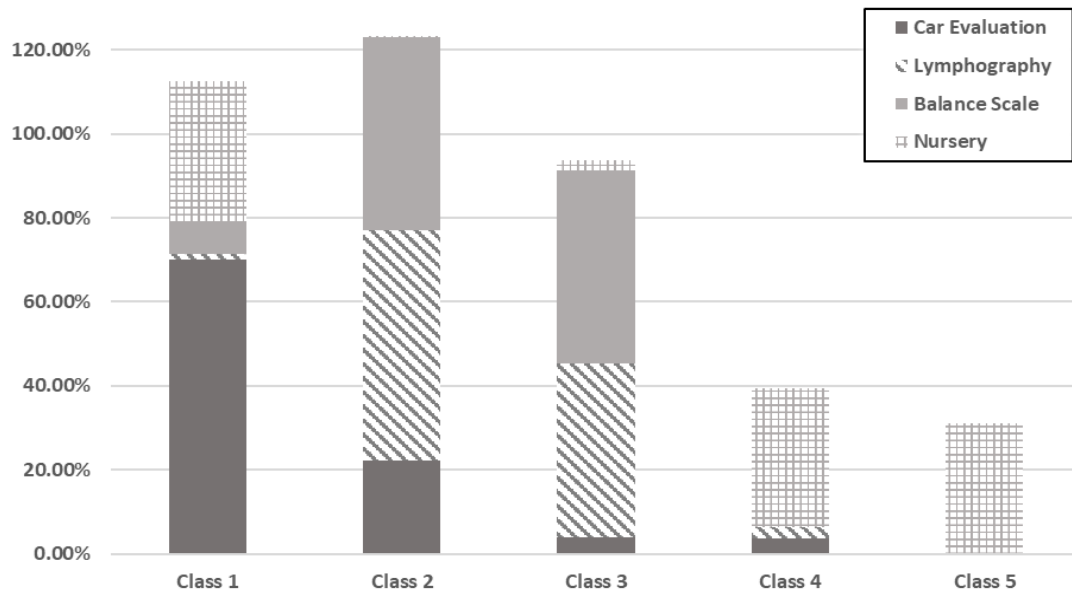


Figure 2. Imbalance distribution of class variable in datasets

### 2.3. Modify

In this phase, variables undergo a cleaning process as deemed necessary. Data cleaning is the process of addressing missing values and noise in a dataset. Following data cleaning, new features are generated by implementing business logic on pre-existing features according to the specified requirements and progressing to the transformation phase. If considered essential, the data shall be transformed into a format required for the analytical technique. The result of this stage is a refined dataset suitable for utilization in the machine learning algorithm for constructing the model. At this stage, an assessment determines how much of the data has been fully transformed. In order to conduct this research, it is imperative to convert categorical data into numerical format. Various categorical encoders were utilized for data conversion. Table 2 delineates the encoding technique implemented during this phase.

Table 2. The number of features before and after encoding

Categorical encoding	Features before encoding				Features after encoding			
	Car	Lym	Nur	Bal	Car	Lym	Nur	Bal
Ordinal	6	18	8	4	6	18	8	4
Nominal	6	18	8	4	6	18	8	4
Target	6	18	8	4	6	18	8	4
Frequency	6	18	8	4	6	18	8	4
Dirichlet	6	18	8	4	24	71	41	12
Leave one out	6	18	8	4	6	18	8	4
Binary	6	18	8	4	18	41	24	16
Hashing	6	18	8	4	8	8	8	8
Dummy	6	18	8	4	15	41	19	16
One Hot	6	18	8	4	21	59	27	20

The explanation of each categorical variable encoding method used is as follows:

- NE: in nominal encoding, an integer is selected for each value of a categorical variable regardless of order.
- OE: ordinal encoding has the same concept as nominal encoding, only in labeling the order of the integer values that are mapped.
- TE: in this encoding, the probability value of each predictors' category level is calculated based on the class (target variable).
- FE: frequency coding is an encoding technique that encodes the value of a categorical feature to its frequency.
- DRE: dirichlet conjugate bayesian method (CBM) has been implemented by Slakey *et al.* [13] by using the prior distribution function is Dirichlet with parameter  $\alpha$ , as in (1), and the likelihood function

of the data is assumed to be multinomial distribution ( $p_1, p_2, \dots, p_k$ ) for multiclass. So that the posterior distribution function produces a Dirichlet distribution with parameters:

$$\alpha^* = \alpha + \sum_{i=1}^n y_i \quad \text{where } \alpha, \alpha^* \in \mathbb{R}^k \quad (1)$$

s.t,

$$p(\theta_{mv}|y) \propto L(\theta_{mv}|y)p(\theta_{mv})$$

$L(\theta_{mv}|y)$  is a function of the likelihood of each value  $v$  in categorical predictor  $m$  against the target  $y$ , and  $\theta_{mv}$  is the distribution parameter for each categorical value, then  $p(\theta_{mv}) \sim \text{dirichlet}(\alpha)$  with  $\alpha = \frac{1}{k}$ ,  $k$  is number of classes.

- f) LE: the idea is the concept of k-fold encoding to compute the target variable's average for all data containing the same value for the categorical feature variable. The average value can be obtained if the target data type is numeric, but if the target data type is categorical, then use the probability value.
- g) OHE: a new variable will be created if the predictor variable is nominal (no order). Each category is mapped with a binary variable containing either 0 or 1.
- h) HE: this approach is suitable for variables that have many categorical levels. Many types of hash functions map random-size data to fixed-size data in a numeric hash.
- i) BE: binary encoding is a combination of hash encoding and one-hot encoding. In this case, the categorical features are first converted to numeric using the ordinal encoder. Then the number is converted to a binary number and divided into different columns.
- j) DE: dummy coding scheme is like one-hot encoding. The dummy encoding is slightly improved over one-hot encoding because its N-1 features represent N labels/categories.

#### 2.4. Model

At this stage, a range of classification modeling algorithms are utilized on pre-processed data to evaluate their efficacy in achieving the intended outcomes. In this phase, six distinct algorithms were employed, specifically decision tree (DT), Naive Bayes (NB), K-nearest neighbor (KNN), random forest (RF), logistic regression (LR), and SVM. A brief explanation of each algorithm is as follows:

- a) DT: a decision tree consists of nodes and branches. Nodes can be divided into root nodes (primary nodes in the tree), decision nodes (conditionally sub-nodes), and leaf nodes (no longer branching nodes). Because the decision tree follows an if-then-else structure, each node uses an independent variable to divide into two or more branches. For categorical variables, the categories are used to determine the segregation of nodes, and for continuous variables, the algorithm generates several thresholds that act as decision rules [5].
- b) NB: Naïve Bayes classifier is a classification method that is rooted in Bayes theorem. The main character is a strong assumption of independence from each event. Olson and Delen [26], each decision class is determined by the probability that the decision class is true. The probabilities involved in producing the final estimate as to the sum of the frequencies from the decision table.
- c) KNN: K-nearest neighbor is a classification algorithm based on k nearest neighbors, and k is the number of nearest neighbors. The most common nearest neighbor search technique is using the distance formula. Distance formulas can use Euclidean, Hamming, Manhattan, or Minkowski.
- d) RF: random forest is a combination of several models of decision trees to make one model. The more DT used the better accuracy. The decision of classification is taken based on the voting results of the formed tree.
- e) LR: logistic regression is a model to classify objects based on probability thresholds. For example, in the case of a binary class, if the probability value is more than 0.5, it will be rounded to 1, which means that the response classification is in the event class. If the probability value is less than or equal to 0.5, it will be rounded to 0, meaning the response classification is in the non-event class. For multiclass classification, the LR model is known as multinomial logistic regression. It formed a separate binary logistic regression model for each response (class) category dummy variable. For example, if it has k class categories, it will produce a k-1 binary LR model. Each model is a probability of the class compared to the reference class.
- f) SVM: the initial concept of SVM was to support binary classification and separate data points into two classes. So, the same principle can be used for multiclass classification by breaking the multiclass problem into several binary classification problems.

Two approaches are often used in handling multiclass cases for SVM, namely the one-vs-one (OVO) and one-vs-all (OVA). In the OVO approach, it takes a hyperplane to separate any two classes ( $r$  and  $s$ ) by ignoring the data points of the other classes. This means the split only takes into account the data points  $i$  of the two classes in each classifier function. The classification function for a new object as in (2):

$$\hat{f}(x_{new}) = \text{sign}(x_{new}^T \hat{w}^{rs} + \hat{b}^{rs} - (1 + \xi^{rs})) \quad (2)$$

where,

$$\hat{w}^{rs} = \sum_{i=1}^n \hat{\alpha}_i y_i x_i, \quad b^{rs} = \frac{1}{n_{sv}} \left( \sum_{i=1}^{n_{sv}} \frac{1}{y_i} - (x_{new}^T \hat{w}^{rs}) \right)$$

by using a kernel trick to map the original predictor variable to a higher dimension then optimization function in (2) can be defined as (3).

$$\min_{w^{rs}, b^{rs}, \xi^{rs}} \frac{1}{2} (w^{rs})^T w^{rs} + C \sum_i \xi_i^{rs} \quad (3)$$

s.t,

$$\begin{aligned} (w^{rs})^T \phi(x_i) + b^{rs} &\geq 1 - \xi_i^{rs}, \quad y_i = r \\ (w^{rs})^T \phi(x_i) + b^{rs} &\leq -(1 - \xi_i^{rs}), \quad y_i = s \end{aligned}$$

The decision of an  $i$ -th object to enter a class uses a voting strategy. In the OVA approach, a strategy is used to create a hyperplane to separate classes and others at once. This means the separation takes into account all data points and then divides them into two groups, namely a group for class data points and a group for all other class data points. A new observation can be classified using (4) as classification measures:

$$\hat{f}(x_{new}) = \text{sign}(x_{new}^T \hat{w}^r + \hat{b}^r - (1 + \xi^r)) \quad (4)$$

where,

$$\hat{w}^r = \sum_{i=1}^n \hat{\alpha}_i y_i x_i, \quad b^r = \frac{1}{n_{sv}} \left( \sum_{i=1}^{n_{sv}} \frac{1}{y_i} - (x_{new}^T \hat{w}^r) \right)$$

where  $x_i$  is the support vector,  $x_{new}$  is the classified data while  $\alpha_i$  is the lagrange multiplier,  $b^r$  is the bias, and  $n_{sv}$  is the number of support vectors. For non-linear separable cases, the kernel trick  $x_i \rightarrow \phi(x_i)$  can be obtained by optimizing function in (5):

$$\min_{w^r, b^r, \xi^r} \frac{1}{2} (w^r)^T w^r + C \sum_i \xi_i^r \quad (5)$$

has the following constraints.

$$\begin{aligned} (w^r)^T \phi(x_i) + b^r &\geq 1 - \xi_i^r, \quad y_i = r \\ (w^r)^T \phi(x_i) + b^r &\leq -(1 - \xi_i^r), \quad y_i \neq r \\ \xi_i^r &\geq 0 \end{aligned}$$

The decision of an  $i$ -th object into a class  $r$  can be predicted by using the highest confidence score (cs) of the input vector  $x_i$  for each of the parameter vectors  $w^r$  and  $b^r$  as in (6):

$$cs = f(x) = w^T \phi(x) + b - (1 + \xi) \quad (6)$$

the grid search method is employed in the hyperparameter optimization procedure for each model to obtain the optimum performance with the parameter settings shown in Table 3.

Table 3. Range value of hyperparameter tuning using grid search

Classification algorithm	Parameters range
Decision tree	Criterion: ['gini', 'entropy'], Min Samples Leaf: [0.1, 1, 10, 100, 500] Max Depth: [1, 10, 100, 500], Min Samples Split: [1, 10, 100, 500]
K nearest neighbor	N- Neighbors: [1, 10, 100, 500], Weights: ['uniform', 'distance'] Metric: ['euclidean', 'manhattan']
Naïve Bayes	Priors: [None], Var Smoothing: np. logspace (0, -9, num=100)
Logistic regression	Penalty: ['l1', 'l2'], C: [1.0, 0.5, 0.1], Solver: ['liblinear']
Random forest	Min Samples Leaf: [1, 10, 100, 500], Max Depth: [1, 10, 100, 500] Min Samples Split: [1, 10, 100, 500]
Support vector machine	Kernel: ['rbf'], C= [0.1, 1, 10, 100], Gamma = [1, 0.1, 0.01] Decision Function Shape: ['ovo', 'ovr']

2.5. Asses

This is the final phase of the SEMMA procedure. Here, the performance of the model is evaluated against test data (not used in model training) to ensure accuracy. The model evaluation metrics can be used is F1-Score [19] with the as in (7) and area under the receiver operating characteristic curve (AUC).

$$F1\text{-Score} = \frac{2 \times AP \times AS}{AP + AS} \tag{7}$$

Where, AP: average precision and AS: average sensitivity (recall).

2.6. Statistical hypothesis testing

The stage of statistical hypothesis testing is an additional step given to complete the SEMMA procedure. Thus, the SEMMA procedure was changed to enhanced-SEMMA. Hypothesis testing is a test of a statement using statistical methods so that the test results can be declared statistically significant. By performing statistical tests on the hypothesis, we can decide whether the hypothesis can be accepted (data does not provide evidence to reject the hypothesis) or rejected (data provides evidence to reject the hypothesis). In the use of parametric statistical methods to test the difference between k-independent samples, there is an assumption that the data is taken from a normally distributed population. If this assumption is met, the one-way analysis of variance F test (Anova) can be used, but if the assumption of normality is not met, then the Kruskall-Wallis test is used [27]. Table 4 describes the design involved four public datasets ( $D_h$ ) with  $h=1,2,3,4$  and ten types of categorical data encoders ( $E_j$ ) with  $j=1,2,\dots,p$  and six classification algorithms ( $A_m$ ) with  $m=1,2,\dots,k$  and validation methods using three types of fold cross-validation, namely 3, 5 and 10 folds. Furthermore, the  $m_{1q3p}$  element means the evaluation measure, i.e., F1-Score or AUC for the first dataset, the k-th classification algorithm, the third cross-validation and the p-th categorical encoder. The experiment was carried out using the Intel Core i7 Gen 7 computer specifications with 64 GB RAM and no observations were made regarding the running time and memory used for each categorical data encoding method in each classification algorithm. This is because the main objective of this research is to select a categorical data encoding method that can contribute to increasing the accuracy (F1-Score) of the classification algorithm performance, especially in cases of multiclass imbalance with categorical features. There are two types of hypotheses that will be tested at this stage, namely the hypothesis regarding the assumption of normality of the data and the hypothesis about the differences in the results of the F1-Score and AUC measurements for each different encoding method, different algorithms, and OVO or OVA strategies in multiclass cases.

Table 4. Experiment design for F1-score and AUC measurement

Dataset	Classification algorithms	q-fold	Categorical encoder			
			E <sub>1</sub>	E <sub>2</sub>	...	E <sub>p</sub>
D <sub>1</sub>	A <sub>1</sub>	3	$m_{1111}$	$m_{1112}$	...	$m_{111p}$
		5	$m_{1121}$	$m_{1122}$	...	$m_{112p}$
		10	$m_{1131}$	$m_{1132}$	...	$m_{113p}$
	A <sub>k</sub>	3	$m_{1k11}$	$m_{1k12}$	...	$m_{1k1p}$
		5	$m_{1k21}$	$m_{1k22}$	...	$m_{1k2p}$
		10	$m_{1k31}$	$m_{1k32}$	...	$m_{1k3p}$
⋮	⋮	⋮	⋮	⋮	⋮	
D <sub>4</sub>	A <sub>1</sub>	3	$m_{4111}$	$m_{4112}$	...	$m_{411p}$
		5	$m_{4121}$	$m_{4122}$	...	$m_{412p}$
		10	$m_{4131}$	$m_{4132}$	...	$m_{413p}$
	A <sub>k</sub>	3	$m_{4k11}$	$m_{4k12}$	...	$m_{4k1p}$
		5	$m_{4k21}$	$m_{4k22}$	...	$m_{4k2p}$
		10	$m_{4k31}$	$m_{4k32}$	...	$m_{4k3p}$

### 3. RESULTS AND DISCUSSION

This section will discuss experimental results based on the design in Table 4. Using the enhanced-SEMMA procedure, data analysis will be carried out to produce recommendations for appropriate transformation methods in case studies of imbalanced multiclass data classification with categorical variables through statistical testing to determine whether there are significant differences. The significant results of each classification performance measurement metric through the F1-Score and AUC based on the type of categorical variable transformation method, the type of classification algorithm, and the type of classification strategy used.

#### 3.1. Performance analysis based classification algorithm

In general, the performance of multiclass classification improved on training-testing validation by 10-fold over the F1-Score measure. To find out which algorithm has superior performance, it can be seen in Table 5. Table 5 shows that SVM has the highest average F1-Score of 0.897 compared to other classification algorithms. To strengthen this conclusion, statistical tests were carried out for more than two independent samples from the calculation results of the average F1-Score in the six classification algorithms. In the analysis of the results of each experiment, it was found that the validation with 10 folds got the highest performance value. The Kruskal Wallis Test approach, which is a nonparametric test as an alternative to the one way anova test, was carried out because the assumption of normality of the data was not met. This Kruskal Wallis test is based on rank which aims to determine whether there are statistically significant differences between two or more groups of independent variables. The test decision shows the significance value is less than 5%. It means that there is a significant difference from the average value of the F1-Score based on the type of classification algorithm used.

Table 5. Descriptive statistics value F1-score classification of multiclass imbalanced dataset based on type of classification algorithm with 10-fold

Classification algorithms	N	Minimum	Maximum	Mean	Std. deviation
DT	40	0.423	0.941	0.769	0.159
KNN	40	0.426	0.947	0.801	0.137
LR	40	0.452	0.978	0.821	0.134
NB	40	0.466	0.934	0.799	0.129
RF	40	0.444	0.944	0.779	0.146
SVM	40	0.392	0.995	0.897	0.143
Valid N (listwise)	40				

#### 3.2. Performance analysis based categorical encoder

Furthermore, an analysis will be carried out on the effect of the type of transformation of categorical variables on classification performance. The data used are four datasets with six types of classification algorithms. Thus, for each type of transformation has a sample of 24 data. Descriptive statistics of the distribution of data can be seen in Table 6. The results of the statistical calculation of the non-parametric test concluded that there were differences in treatment with different types of categorical variable transformation methods for the F1-Score with a significance value of less than 5%. Based on the ranking results of the average F1-Score value, it is found that the three types of categorical variable transformations that rank at the top are ordinal encoding, dirichlet encoding and target encoding. The choice of this type of transformation can be related to the proportion of the level of imbalance in the dataset class used.

Table 6. Descriptive statistics value F1-score classification of multiclass imbalanced dataset based on type of categorical encoder with 10-fold

Categorical encoder	N	Minimum	Maximum	Mean	Std. deviation
OE	24	0.679	0.995	0.879	0.080
NE	24	0.636	0.995	0.857	0.102
FE	24	0.392	0.942	0.659	0.204
DRE	24	0.675	0.987	0.880	0.073
TE	24	0.701	0.995	0.879	0.078
DE	24	0.512	0.989	0.800	0.146
HE	24	0.493	0.868	0.643	0.103
OHE	24	0.609	0.993	0.848	0.103
BE	24	0.656	0.995	0.852	0.100
LE	24	0.523	0.995	0.812	0.161
Valid N (listwise)	24				



To support the selection of the appropriate transformation method, Figure 3 is a comparison chart of the three methods above using different datasets for the SVM algorithm. It can be concluded that for the distribution of class data that has the smallest proportion value of less than 3%, the transformation method with the dirichlet function approach gives the best effect on the results of the F1-Score measurement. Meanwhile, for class data whose smallest proportion is more than 3%, the transformation method with ordinal encoding gives the best prediction performance. The result of the calculation of the non-parametric test statistic is the rejection of the null hypothesis (p-value less than error level), meaning that there are differences in treatment with different types of transformation methods for the F1-Score with a significance level of 5%.

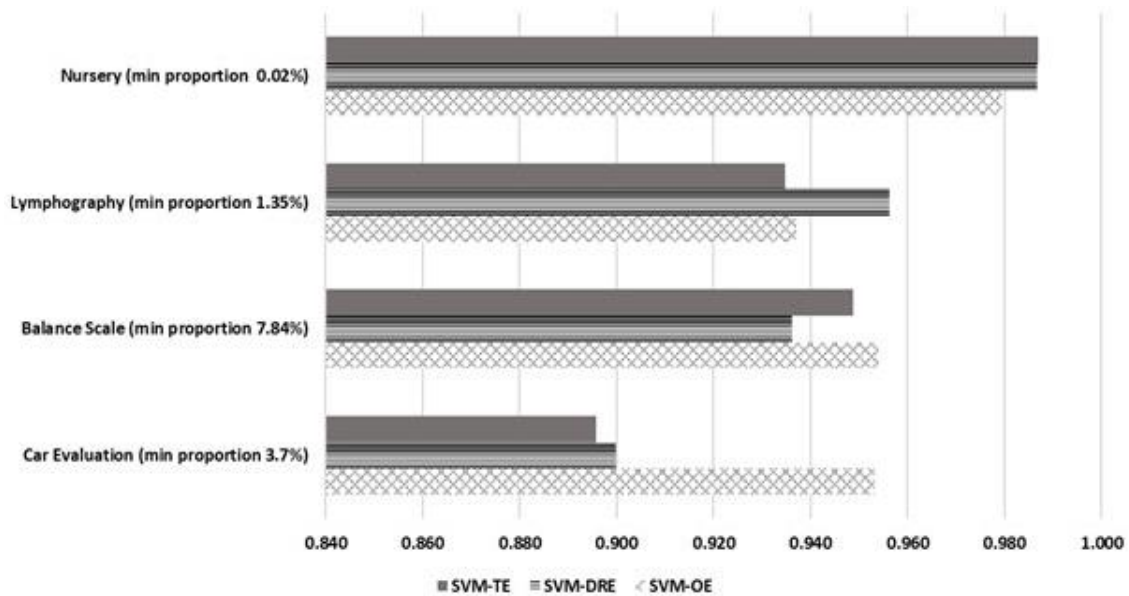


Figure 3. Comparison of average F1-score values for OE, DRE, and TE on the benchmarking dataset

**3.3. Performance analysis based multiclass decision strategies**

In the results of the comparative analysis of classification algorithms, it is concluded that SVM has a higher performance than other algorithms. The transformation method with the dirichlet function or number labels in order (ordinal) can be used based on the analysis of the results of the comparison of the categorical encoding method. Thus, in this section only analyze the AUC value with OVO and OVA approaches for all transformation methods. The decision of statistical testing regarding whether there is a difference between the use of OVO and OVA strategies in the multiclass classification for imbalanced data using the AUC measure, it was found that there was no significant difference for the use of OVO and OVA strategies. However, the results of ranking the AUC performance values between the OVO and OVA strategies in Table 7 show that the OVA ranking is higher than the OVO ranking, thus the OVA strategy is recommended.

Table 7. Calculation of mean rank on AUC data for OVO and OVA strategies with 10-fold

Decision strategies		N	Mean rank	Sum of ranks
AUC	OVO	40	39.23	1569
	OVA	40	41.78	1671
	Total	80		

**4. CONCLUSION**




The conclusion from the results of this study is that three groups of coding methods are recommended as categorical to numeric feature transformation techniques. The first group, called label encoding, consists of ordinal and nominal encoders. Secondly, based on calculating conditional probabilities between class and level categorical variables, namely the target and dirichlet encoders. The third group is the transformation method by increasing the number of features using the binary concept, namely dummy, one-hot, and binary encoder. All three groups yielded very good predictive performance. The enhanced-SEMMA procedure has provided recommendations on which coding methods, algorithms, and decision strategies are suitable for predicting

multiclass classification on unbalanced data. The results of an empirical study with public data show that the SVM algorithm performs better than other algorithms, including DT, NB, KNN, LR, and RF. Statistical tests show that the choice of categorical to numerical transformation method significantly affects the F1-Score value. Different tests show that the ordinal, dirichlet, and target encoding transformation methods occupy the top three recommendations. Probability-based or binary coding, such as target, dirichlet, dummy, one-hot, or binary, is best for situations where the proportion of minor classes is less than 3%. Nominal or ordinal encoders are preferred for data with a minor class proportion of more than 3%. As for the choice of classification determination strategy, both OVO and OVA did not have a significant difference, even though the AUC value in the OVA approach was higher than OVO.




## REFERENCES

- [1] S. J. Russell and P. Norvig, "Artificial intelligence a modern approach," *Inc.*, 2010.
- [2] M. Mohri, A. Rostamizadeh, and A. Talwaker, "Foundation machine learning," in *2nd Editio. London: The MIT Press*, 2018.
- [3] F. Y. Oisanwo *et al.*, "Supervised machine learning algorithms: classification and comparison," *International Journal of Computer Trends and Technology*, vol. 48, no. 3, pp. 128–138, Jun. 2017, doi: 10.14445/22312803/ijctt-v48p126.
- [4] H. Bisgin *et al.*, "Comparing SVM and ANN based machine learning methods for species identification of food contaminating beetles," *Scientific Reports*, vol. 8, no. 1, Apr. 2018, doi: 10.1038/s41598-018-24926-7.
- [5] H. Z. M. Shafri and F. S. H. Ramle, "A comparison of support vector machine and decision tree classifications using satellite data of Langkawi Island," *Information Technology Journal*, vol. 8, no. 1, pp. 64–70, Dec. 2008, doi: 10.3923/itj.2009.64.70.
- [6] Z. Zheng, Y. Cai, Y. Yang, and Y. Li, "Sparse weighted naive bayes classifier for efficient classification of categorical data," in *Proceedings - 2018 IEEE 3rd International Conference on Data Science in Cyberspace, DSC 2018*, Jun. 2018, pp. 691–696, doi: 10.1109/DSC.2018.00110.
- [7] H. C. Rustamaji, O. S. Simanjuntak, S. F. Luhrie, B. Yuwono, and Juwairiah, "Categorical data classification based on fuzzy K-nearest neighbor approach," in *Proceeding - 2019 5th International Conference on Science in Information Technology: Embracing Industry 4.0: Towards Innovation in Cyber Physical System, ICSITech 2019*, Oct. 2019, pp. 171–175, doi: 10.1109/ICSITech46713.2019.8987477.
- [8] M. Azmi, G. C. Runger, and A. Berrado, "Interpretable regularized class association rules algorithm for classification in a categorical data space," *Information Sciences*, vol. 483, pp. 313–331, May 2019, doi: 10.1016/j.ins.2019.01.047.
- [9] J. T. Hancock and T. M. Khoshgoftaar, "Survey on categorical data for neural networks," *Journal of Big Data*, vol. 7, no. 1, Apr. 2020, doi: 10.1186/s40537-020-00305-w.
- [10] Y. Wang and W. Xu, "Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud," *Decision Support Systems*, vol. 105, pp. 87–95, Jan. 2018, doi: 10.1016/j.dss.2017.11.001.
- [11] O. E. Ogundijo, D. He, and L. Parida, "Performance evaluation of different encoding strategies for quantitative genetic trait prediction," in *2015 IEEE 5th International Conference on Computational Advances in Bio and Medical Sciences (ICCBMS)*, Oct. 2015, doi: 10.1109/ICCBMS.2015.7344715.
- [12] F. Pargent, "A benchmark experiment on how to encode categorical features in predictive modeling," *Ludwig-Maximilians-Universität München*, 2019.
- [13] A. Slakey, D. Salas, and Y. Schamroth, "Encoding categorical variables with conjugate bayesian models for wework lead scoring engine," pp. 1–15, 2019, *arXiv:1904.13001*.
- [14] N. Lee and J. M. Kim, "Conversion of categorical variables into numerical variables via Bayesian network classifiers for binary classifications," *Computational Statistics and Data Analysis*, vol. 54, no. 5, pp. 1247–1265, May 2010, doi: 10.1016/j.csda.2009.11.003.
- [15] H. Li, R. Yuan, W. Peng, Y. Liu, and H. Z. Huang, "Bayesian inference of Weibull distribution based on probability encoding method," in *ICQR2MSE 2011 - Proceedings of 2011 International Conference on Quality, Reliability, Risk, Maintenance, and Safety Engineering*, Jun. 2011, pp. 365–369, doi: 10.1109/ICQR2MSE.2011.5976632.
- [16] H. Han, X. Zhu, and Y. Li, "EDLT: enabling deep learning for generic data classification," in *Proceedings - IEEE International Conference on Data Mining, ICDM*, Nov. 2018, pp. 147–156, doi: 10.1109/ICDM.2018.00030.
- [17] Z. Yin and Y. Shen, "On the dimensionality of word embedding," *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*, 2018.
- [18] R. Luo *et al.*, "Feature learning with a divergence-encouraging autoencoder for imbalanced data classification," *IEEE Access*, vol. 6, pp. 70197–70211, 2018, doi: 10.1109/ACCESS.2018.2879221.
- [19] B. S. Arkok and A. M. Zeki, "Classification of Quranic topics based on imbalanced classification," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 22, no. 2, p. 678, May 2021, doi: 10.11591/ijeecs.v22.i2.pp678-687.
- [20] R. Ghorbani and R. Ghousi, "Comparing different resampling methods in predicting students' performance using machine learning techniques," *IEEE Access*, vol. 8, pp. 67899–67911, 2020, doi: 10.1109/ACCESS.2020.2986809.
- [21] T. Al-Shehari and R. A. Alsowail, "An insider data leakage detection using one-hot encoding, synthetic minority oversampling and machine learning techniques," *Entropy*, vol. 23, no. 10, p. 1258, Sep. 2021, doi: 10.3390/e23101258.
- [22] A. K. Hamoud *et al.*, "A prediction model based machine learning algorithms with feature selection approaches over imbalanced dataset," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 28, no. 2, pp. 1105–1116, Nov. 2022, doi: 10.11591/ijeecs.v28.i2.pp1105-1116.
- [23] H. Hartono, E. Ongko, and Y. Risyani, "Combining feature selection and hybrid approach redefinition in handling class imbalance and overlapping for multi-class imbalanced," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 21, no. 3, pp. 1513–1522, Mar. 2021, doi: 10.11591/ijeecs.v21.i3.pp1513-1522.
- [24] M. Bhagat and B. Bakariya, "Implementation of logistic regression on diabetic dataset using train-test-split, K-fold and stratified K-fold approach," *National Academy Science Letters*, vol. 45, no. 5, pp. 401–404, Jul. 2022, doi: 10.1007/s40009-022-01131-9.
- [25] A. Azevedo and M. F. Santos, "KDD, SEMMA and CRISP-DM: a parallel overview," in *Proceedings of the IADIS European Conference on Data Mining*, 2008, pp. 183–185.
- [26] D. L. Olson and D. Delen, "Advanced data mining techniques," *Springer Science & Business Media*, 2008.
- [27] S. Sawilowsky and G. Fahoome, "Kruskal-wallis test: basic," *Wiley StatsRef: Statistics Reference Online*, Sep. 2014, doi: 10.1002/9781118445112.stat06567.




**BIOGRAPHIES OF AUTHORS**

**Wiyli Yustanti**    is a doctoral candidate in the Department of Statistics, Faculty of Science and Data Analytics, Institut Teknologi Sepuluh Nopember (ITS) Surabaya. She received a bachelor's degree in statistics, and her master's degree is in informatics at the same university (ITS). She is an associate professor at the department of informatics, faculty of engineering, Universitas Negeri Surabaya. Her research interests include soft computing, machine learning, data mining, data science, business intelligence, and decision support systems. She can be contacted by email at: [wilylyustanti@unesa.ac.id](mailto:wilylyustanti@unesa.ac.id).



**Nur Iriawan**    received a bachelor's degree in statistics from the Sepuluh Nopember Institute of Technology (ITS) Surabaya, a master's degree in computer science from the University of Maryland USA, and a Ph.D. in statistics from Curtin University of Technology, Australia. Currently a professor at the department of statistics, faculty of science and data analysis, Sepuluh Nopember Institute of Technology (ITS), Surabaya. He also serves as head of the computational statistics and data science laboratory. He has supervised and co-supervised over 20 master and 10 Ph.D. students. He has authored or co-authored more than 60 Scopus articles, with 12 H-indexes and over 1,000 citations. His research interests include stochastic processes, statistical computations, and Bayesian models. He can be contacted via email at: [nur\\_i@statistika.its.ac.id](mailto:nur_i@statistika.its.ac.id).



**Irhamah**    completed his doctoral studies in the department of mathematics, Universiti Teknologi Malaysia (UTM), Malaysia. Her undergraduate and master's degrees were earned in statistics. Currently, his research area at that time was operations research, genetic algorithms, time series analysis and statistical computing. She is also a member of the research group in the computational statistics and data science laboratory at the department of statistics, faculty of science and data analysis, Sepuluh Nopember Institute of Technology (ITS), Surabaya. She has been involved in more than 50 Scopus-indexed publications. She can be contacted via email at: [irhamah@statistika.its.ac.id](mailto:irhamah@statistika.its.ac.id).