PLOS ONE

RESEARCH ARTICLE

Combination of unsupervised discretization methods for credit risk

José G. Fuentes Cabrera[1,2], **Hugo A. Pérez Vicente**[1], **Sebastián Maldonado**[3,4], **Jonás Velasco**[5]*

1 Departamento de Ingeniería Química, Industrial y de Alimentos, Universidad Iberoamericana Ciudad de Mexico, Mexico City, México, **2** Facultad de Estudios Superiores Acatlán, UNAM, Naucalpan de Juárez, México, **3** Department of Management Control and Information Systems, School of Economics and Business, University of Chile, Santiago, Chile, **4** Instituto Sistemas Complejos de Ingeniería (ISCI), Santiago, Chile, **5** CONAHCYT-Centro de Investigación en Matemáticas (CIMAT), Guanajuato, México

* jvelasco@cimat.mx

Citation: Fuentes Cabrera JG, Pérez Vicente HA, Maldonado S, Velasco J (2023) Combination of unsupervised discretization methods for credit risk. PLoS ONE 18(11): e0289130. https://doi.org/10.1371/journal.pone.0289130

Editor: Haris Khurram, National University of Computer and Emerging Sciences, PAKISTAN

Received: December 2, 2022

Accepted: July 11, 2023

Published: November 27, 2023

Data Availability Statement: Australian: https://archive.ics.uci.edu/ml/datasets/statlog+(australian+credit+approval) Microfinance Loan Credit Scoring: https://www.kaggle.com/shahrukhkhan/microfinance-loan-credit-scoring German: https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data) Give me some credit: https://www.kaggle.com/c/GiveMeSomeCredit HMEQ: https://www.kaggle.com/ajay1735/hmeq-data Japanese: https://archive.ics.uci.edu/ml/datasets/Japanese+Credit+Screening Lending club: https://www.kaggle.com/wordsforthewise/lending-club Mortgage: https://github.com/JLZml/Credit-

Abstract

Creating robust and explainable statistical learning models is essential in credit risk management. For this purpose, equally spaced or frequent discretization is the de facto choice when building predictive models. The methods above have limitations, given that when the discretization procedure is constrained, the underlying patterns are lost. This study introduces an innovative approach by combining traditional discretization techniques with clustering-based discretization, specifically *k* means and Gaussian mixture models. The study proposes two combinations: Discrete Competitive Combination (DCC) and Discrete Exhaustive Combination (DEC). Discrete Competitive Combination selects features based on the discretization method that performs better on each feature, whereas Discrete Exhaustive Combination includes every discretization method to complement the information not captured by each technique. The proposed combinations were tested on 11 different credit risk datasets by fitting a logistic regression model using the weight of evidence transformation over the training partition and contrasted over the validation partition. The experimental findings showed that both combinations similarly outperform individual methods for the logistic regression without compromising the computational efficiency. More importantly, the proposed method is a feasible and competitive alternative to conventional methods without reducing explainability.

1 Introduction

Credit risk is one of the most popular fields of application of machine learning techniques due to the abundance of datasets and the number of developments deployed in the financial sector and in particular in retail credit [1, 2].

When a machine learning classification problem is faced, it can be mistakenly assumed that the problem is solved due to the large number of efforts made by the scientific community to produce multiple theoretical solutions. However, the complexity of the problem increases as applications demand more sophisticated methods to meet the specific requirements of the niche where they are encountered [3–5].

Scoring-Data-Sets/tree/master/6.%20Credit%20risk%20analysis/mortgage Polish: https://archive.ics.uci.edu/ml/datasets/Polish+companies+bankruptcy+data Taiwan: https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients The dataset referred to as "Mexico" cannot be shared for proprietary and confidential reasons.

Funding: This study was financially supported by Universidad Iberoamericana Ciudad de México in the form of a graduate scholarship received by JF. This study was also supported by Universidad Iberoamericana Ciudad de México in the form of salary for HP. The specific role of this author is articulated in the 'author contributions' section. This study was also financially supported by ANID PIA BASAL in the form of an award (AFB180003) received by SM. This study was also financially supported by FONDECYT Chile in the form of a grant (1200221) received by SM. This study was also financially supported by Chairs Program of the National Council of Humanities, Science and Technology (CONAHCYT) project (2193) award received by JV. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

To this end, the methodology for knowledge discovery in large data volumes CRISPDM (Cross Industry Standard Process for Data Mining) [6] with the subsequent steps to follow: business understanding, data understanding, data preparation, modeling, assessment, and deployment, is widely used in the field. As mentioned earlier, each step features large samples of research work, seeking efficiencies in these crucial stages necessary to resolve the application problem satisfactorily [7–9].

Despite the previous, a lack of focus on the explainability methods through generalized linear models is present in practice. Also, simple discretization methods based on equal frequency or distance are still used as the first option.

It is essential to notice that logistic regression, in combination with discretization methods, is the current gold standard for credit scoring, which is the main application for this study. This is due to regulatory constraints (Basel II/III), which prohibit the use of black box models such as neural networks [10].

Although decision trees (DTs) are white box models and regulatory restrictions do not apply, they do not benefit from a discretization step since DTs aim is to generate an optimal binning to define if/then rules that maximize the discriminative power of the variables. Therefore, a previous discretization step would be redundant and negatively affect the method's performance.

On the other hand, artificial neural networks (ANNs) are designed to capture complex relationships between (numerical) variables. Using binning as a preprocessing step limits the ability of ANNs to construct non-linear functions, affecting the predictive performance negatively. Additionally, improving the interpretability of an ANN with a discretization step is not possible.

Discretization of continuous predictors is a widely used technique for data preprocessing [11–13], which is a crucial stage prior to entering data in the learning process in multiple machine learning classifiers. The above is based on the fact that such preprocessing captures patterns and minimizes statistical noise by grouping similar elements and facilitating the mapping between the feature space $\mathcal{X}$ and the binary response vector $\vec{y}$.

Several approaches of discretization methods are used to pursue different objectives, such as monotonic risk clusters [13], optimal clusters regarding the response feature [11, 14], or clustering to improve a specific classifier [15, 16]. However, little attention is given to combination methods.

Combining different discretization approaches can overcome the information intrinsically lost during the clustering process without lack of explainability of the resulting models, which is of utmost importance in the financial and healthcare industry [17–20].

The proposal of this paper outlines two innovative approaches. Firstly, to assemble the unsupervised discretization methods most commonly used in the industry: equal distance and equal frequency methods to improve the predictions in a classification problem using a logistic regression model. Secondly, to add two clustering algorithms (k means and Gaussian mixture models) to the combination to produce more similar and homogeneous segments. Multiple supplementary segments can be incorporated into the model by using this approach, improving the classification quality of the classifier without losing interpretability.

The paper is articulated as follows: in Section 2, the necessary theoretical background to read this paper is briefly reviewed. In Section 3, the available related papers in the literature are discussed. Section 4 discusses the details on the availability of the software, data description, data processing and experimentation derived from this paper. In Section 5, the proposed combination algorithm approach is described, as well as the two implementation variations raised: Discrete Competitive Combination (DCC) and Discrete Exhaustive Combination (DEC). Additionally, the experimentation methodology is underlined. Experimental results and their corresponding discussion are presented in Section 6. Finally, the conclusions and possible future lines of work are addressed in Section 7.

2 Methodological background

This section will briefly describe the critical and necessary theoretical concepts used in this research paper.

2.1 Discretization methods

2.1.1 Equal distance method. A simple discretization method in which the continuous feature will be divided into equal intervals [13]. The number of intervals must be determined *a priori*. With this understanding, let A be a feature with domain $D_A \in (a, b)$ for $a > b, \ a, b \in \mathbb{R}$ and k the number of intervals of the discretization. Then, the boundaries of each interval will be:

$$\left\{a + i\frac{(b-a)}{k} \mid i = 0, \ldots, k\right\}$$

2.1.2 Equal frequency method. This method also requires establishing the number of intervals in which feature A will be divided. The equality of intervals will not be set through distances but through the frequency associated with each interval [13]. For this, the data domain's percentile calculation (denoted p). Therefore, given $D_A \in (a, b)$ for $a > b, \ a, b \in \mathbb{R}$ and k the number of desired intervals, then the interval boundaries will be provided by:

$$\{a\} \cup \left\{p_{\frac{100i}{k}} \mid i = 1, \ldots, k-1\right\} \cup \{b\}$$

2.1.3 *k* means discretization method. In the case of k means [21], the discrete intervals will be established by unsupervised learning in which the space will only be formed by feature A. The method works through the minimization of the cost of the clusters formed. Then, given a set of points: $C_j = \{i \in \{1, \ldots, n\}|z^{(j)}\}$ represents $x^{(j)}$, then:

$$\text{Cost}\left(z^{(1)}, \ldots, z^{(k)}\right) = \sum_{i=1}^{n} \min_{j=1,\ldots,k} \|x^{(i)} - z^{(j)}\|^2$$

For this purpose, the k means algorithm finds the best clusters for the centroids and the best centroids for the clusters. Iterations are performed according to the following:

1. Initialize the centroids $z^{(1)}, \ldots, z^{(k)}$
2. Repeat until there is no significant change in the Cost function:

(a). For each $j = 1, \ldots, k$:
$C_j = \{i|x^{(i)}$ is the closest to $z^{(j)}\}$

(b). For each $j = 1, \ldots, k$:
$z^{(j)} = \frac{1}{|C_j|}\sum_{i \in C_j} x^{(i)}$

2.1.4 Gaussian mixture model discretization method. Gaussian mixture models [22] is an unsupervised learning method that produces k clusters. The assignment of each cluster is produced according to the probabilities produced by a Gaussian distribution. The one with the highest probability will be selected as the representative cluster. The Gaussian mixture model is a weighted sum of multivariate normal distributions as the following:

$$\sum_{j=1}^{k} p_j N(x; \mu^{(j)}, \sigma_j^2 I)$$

where p_j represents the weights of the mixture, $\mu^{(j)}$ the mean vector, and $\sigma_j^2 I$ the covariance matrix.

2.2 Classification metrics

2.2.1 Area under the ROC curve (AUC). As indicated in [23], the procedure for obtaining AUC is based on the confusion matrix resulting from training with the peculiarity that the probability cutoff to determine the class is measured in different *k* points, thus producing the two-dimensional coordinates for the formation of a curve called receiver operating characteristic or ROC curve. Each coordinate corresponds to the ordered pair of sensitivity and 1 minus specificity. Subsequently, the indicator will be obtained from the calculation of the area under the mentioned curve, where the reference value 0.5 indicates an identical classification to the random effect and 1.0 indicates a perfect classification. Therefore, values closer to 1.0 indicate a model with better classification.

2.2.2 Kolmogorov Smirnov. The statistic of the Kolmogorov Smirnov (KS) test is based on the maximum difference between the cumulative distributions of an empirical and a theoretical distribution [24] (The positive and negative classification will take such roles indistinctively but not simultaneously). Higher values of KS value indicate better classification, being zero, its minimum value in the presence of identical distributions.

2.3 Logistic regression

Logistic regression allows modeling a dichotomous outcome in a probabilistic manner where p represents the probability of occurrence of an event in particular. The complement of p is denoted by $q = 1 - p$. The modeled probability approach will be based on the odds ratio, calculated using the expression:

$$\frac{p}{q} = \frac{p}{1-p}$$

If each independent variable x_i proportionally contributes to the response feature, the weighted sum of the predictors in linear correspondence with the logarithm of the odds ratio is proposed, hence:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \sum_{i=1}^{n} \beta_i x_i$$

Generating an expression for p:

$$p = \frac{1}{1 + \exp(-\beta_0 - \sum_{i=1}^{n} \beta_i x_i)} \tag{1}$$

The previous expression is a logistic function (sigmoid), which returns values between 0 and 1. These will be interpreted as the probability of occurrence of the event for each vector presented. β_i parameters are usually estimated using the maximum likelihood method [25].

2.4 Weight of Evidence and information value

The *Weight of Evidence* (WoE) encoding converts the risk associated with the probability of occurrence of an outcome in a linear scale which is easily interpreted [10]. To generate such transformation, the following is taken into consideration:

Let P be the number of occurrences of a certain outcome in particular and N the number of non-occurrences. Then, the WoE encoding for the i-th attribute associated with a particular feature is given by:

$$\mathrm{WoE}_i = \log\left(\frac{N_i}{\sum_{i=1}^{n} N_i} \Big/ \frac{P_i}{\sum_{i=1}^{n} P_i}\right)$$

WoE encoding is suitable for logistic regression because a linear correspondence with the sigmoid function is present in its structure. A consequence of the WoE transformation consists in the capability to measure the classification quality of each explanatory variable through the statistic known as *Information Value* (IV), which is calculated using the equation:

$$\mathrm{IV} = \sum_{i=1}^{n}\left[\left(\frac{N_i}{\sum_{i=1}^{n} N_i} - \frac{P_i}{\sum_{i=1}^{n} P_i}\right)\mathrm{WoE}_i\right]$$

Information Value (IV) corresponds to the Kullback-Leibler divergence between the event, and non-event distributions [26].

3 Related work

Credit risk is one of the field applications in which the aim is to propose a simple and interpretable model to find the pattern that will determine if a borrower will or will not repay the granted loan [27]. Ai et al.'s [28] work presents an approach through the order of categorical data to cluster information in the face of the potential for fraud in claims under the insurance policies industry, the method used by the author is called RIDIT (Relative to an Identified Distribution Integral Transformation), in which an ordinal measurement is assigned to discrete variables, in this regard, the idea is compatible with the proposal of this paper. On the other hand, there are several applications, such as those shown by Nehrebecka [29] Melchiori [30], and Bunker [31] related to the default risk of companies, change of category of regulatory classification and improvement of the classification quality by adding information, respectively. In particular, both in [29, 30], there are regulatory implications for having explainable methods, such as the one raised in this paper, which is essential. Moreover, regarding the effects of the systemic impact of the COVID-19 pandemic for measuring the effect in the probability of default for non-financial companies using a stress-testing framework is proposed by Nehrebecka [32].

Additionally, the applications shown by Wang [33] and Zhang [34] outline approaches based on deep learning for peer-to-peer lending. In Wang's case, an approach based on long-short term memory (LSTM) architectures is used, whereas, in Zhang's paper, deep networks and gradient boosting models are assembled to produce the online integrated credit scoring model (OICSM), adding online update capabilities. Furthermore, Guo's paper [35] proposes a self-adaptive method to assemble classifiers applied to credit risk. The method involves creating base classifiers and selecting the optimal combination through Bayesian optimization. Finally, some of the datasets used in this paper [36–38] were also selected by Guo for experimentation.

Many techniques with different approaches, such as segment restrictions and desirable engineering characteristics regarding discretization methods, are widely present in the literature. An interesting summary of the taxonomy of such methods can be found in [12]. However, in this review, no references were found regarding a discretization proposal based on Gaussian mixture models as proposed in this paper. Moreover, it should be noted that discretization based on clustering is implemented in the scikit learn [39] machine learning library through the *k* means method, as well as in different CRAN repository packages in the R programming language [40] due to which there is a history of widespread use of a similar approach.

A broad range of approaches with similar aims to the proposal of this paper can be found in connection with discretization methods. For example, Adeodato [11] proposes a method to maximize the KS (Kolmogorov Smirnov) statistic through clustering in the discretization process. In addition, in Cano's [41] paper, a method called LAIM (Label Attribute Interdependence Maximization) is proposed to have a specific discretization for multinomial classification problems.

A comparison between discretization methods is discussed in Lund's [15] paper regarding the added value of discretization through the WoE (Weight of Evidence) transformation versus an approximation based on the binarization of the segments of the discrete predictors. In Saia's [42] paper, discretization alternatives are proposed to improve the classification quality applied to credit scoring through the method called Discretized Enriched Data (DED).

The method mentioned above consists of two steps: the first is to reduce the patterns of predictors through discretization, and the second is to enrich data through meta predictors. Finally, Navas [14] proposes the formation of optimal discretization groups through mathematical programming.

Following the aim of this paper's research (to discretize and combine to improve classification quality applied to credit risk), alternative studies that supplement this effort are present. For instance, Wojciak [43] compares four discretization methods to assemble them into a Bayesian network. On the other hand, in Yang's [16] research, discretization methods are combined into a Naive Bayes classifier. In Hsieh's [44] work, an assembly of several methods is generated to boost the classifier applied to credit scoring. Finally, Vejkanchana [45] proposes an innovative way of generating groups in the discretization process through supervised modeling using evolutionary computation, allowing the user to determine the form and number of desired clusters through a constrained optimization.

As far as Raymaeker's [46] investigation is concerned, a fascinating approach can be found based on the creation of non linear segments through splines, seeking to ensure the capture of complex patterns within data without detriment to interpretability. The latter has similarities with this paper by aligning in vision and differing through the approach of discrete segment formation. No less important is to stress that papers such as Silva's [17], Raymaekers's [47] and Iftikhar's [48, 49] highlight the importance of the resulting classifiers' explainability in multiple industries and cases of use, which is one of the purposes of this paper.

The main contributions of this research are focused on supplementing the notorious efforts of the scientific community to provide suitable tools for their practical application; for this reason, the proposal of this paper focuses on supplementing and joining the efforts discussed regarding discretization techniques, improvement of classification quality in a classifier and its direct applications in credit risk.

4 Data processing and software availability

The software created during this research is available in the GitHub repository https://github.com/JGFuentesC/creditrisk_discretization_combination. The repository contains everything necessary to process raw data, replicate the computational experiment and delve into the programming logic used to develop the proposed methodology.

4.1 Data description

A computational experiment is proposed based on proving the competitiveness of the combination in various datasets. The chosen datasets have helpful features to achieve the aim of this study, namely: credit risk classification for several retail and business banking products, sufficient data volume, presence of continuous variables, and presence in different related works. Following is a list of the used datasets in detail:

1. Australian credit—Australian [37]
2. Microfinance loan credit scoring—Farmers [50]
3. German Credit—German [36]
4. Give me some credit—Give-me-some-credit [51]
5. HMEQ—Hmeq [52]
6. Japan credit—Japan [38]
7. Lending club—Lending-club [53]
8. Mexican credit (used in a previous paper of the authors)—Mexico [27]
9. Mortgage—Mortage [52]
10. Polish companies—Polish [54]
11. Taiwan credit—Taiwan [55]

The basic metadata of each dataset is shown in Table 1. Note that when a dataset surpasses 10,000 observations, a simple random sampling of 5,000 (for maintaining the original event distribution) is taken for conducting the experiment.

4.2 Dataset topology

To illustrate the variety of the chosen data, two dimensions data topology in each one of the datasets that will be studied is shown in Fig 1. The objective of the display is to graphically

Table 1. Dataset description.

Dataset	Num. Obs.	Predictors	Pos. Cases	Neg. Cases
Australian	690	6	307	383
Farmers	5000	12	2164	2836
German	959	7	275	684
Give-me-some-credit	5000	10	327	4673
Hmeq	5960	10	1189	4771
Japan	690	4	383	307
Lending-club	5000	77	693	4307
Mexico	5000	14	579	4421
Mortgage	5000	15	112	4888
Polish	5000	64	239	4761
Taiwan	5000	6	534	4466

Dataset basic description.

https://doi.org/10.1371/journal.pone.0289130.t001

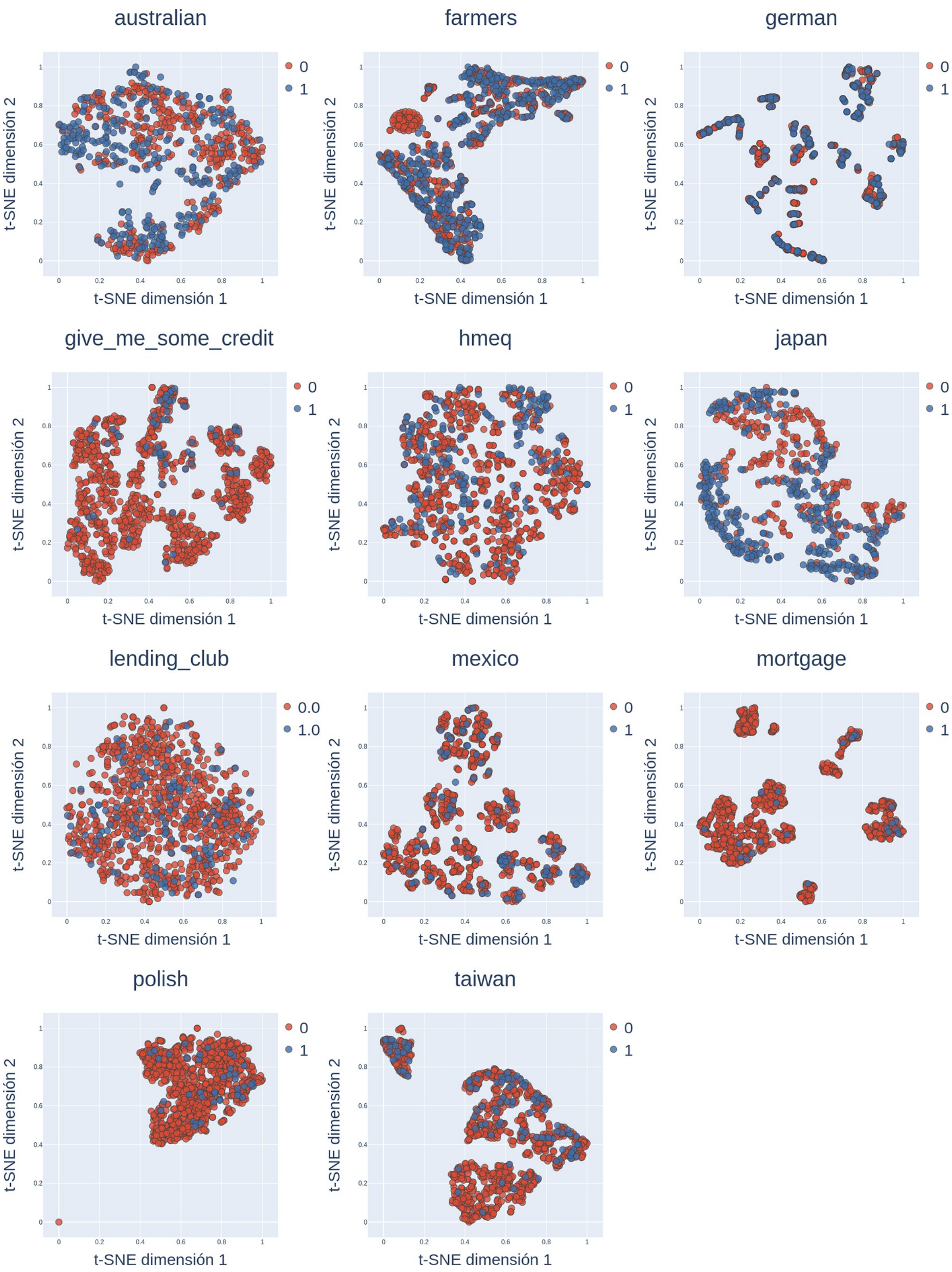

Fig 1. Datasets topology. Data topology, t-SNE with event based hue.

https://doi.org/10.1371/journal.pone.0289130.g001

show the variability in the datasets through a standard of comparison. The information presented is obtained using the technique known as *t* Distributed Stochastic Neighbor Embedding (t-SNE) [56], which represents the predictor hyperspace in its most likely form through probability distribution fitting in pairs of objects in the high dimensional space in order to ensure that similar objects are assigned to higher probabilities. Subsequently, the method carries out a similar process in a low dimensional space to minimize the Kullback-Leibler divergence between both distributions.

With respect to data topology shown in Fig 1, the following characteristics are highlighted:

- Australian, Farmers, and Japan are the only datasets showing class balance.
- Complex patterns are likely to emerge in datasets closest to real-world conditions such as lending club, mexico, polish, and mortgage and give me some credit.
- In general, no linear separation is shown.

As can be observed, the presented datasets have a broad range of topologies. Therefore, the preceding indicates the soundness of this paper's proposal, whereby the generalization capability for multiple configurations and not only through classic or linearly separable examples are tested.

5 Methodology

5.1 Combination method definition

This paper proposes a combination method that considers the information contributed by each discretization method to generate the most accurate classifier without loss of interpretability. Moreover, the information captured by a particular discretization method will not necessarily consider the benefits of its counterparts. This principle is used similarly in the algorithms based on decision trees, as seen in Chen's [57]. In that work, the memory capacity is suppressed under the further inclusion of each previously chosen predictor, except that a sub-partition will improve the classification capacity. The proposal of this paper uses this approach analogously; however, the way to capture additional information will not be through successive partitions of the same variable but through the compliment that a partition determined by a different rule may provide.

This approximation is represented intuitively in Fig 2:

Fig 1 shows how different techniques will produce distinct segments in each predictor feature; in addition, the proportion (purity) of the partition between the elements of the positive and negative class of the binary classifier will also be different. Consequently, the selection of the technique that generates a purer partition will be chosen; hence, in contrast to the individual selection, any combination is expected to deliver better results experimentally.

In order to carry out the task, each one of the continuous predictors is discretized in $k = 2, \ldots, 10$ segments according to each method. The choice of the *k* range is based on the fact that being interpretability a critical element of the proposed method, having more than 10 segments would make interpretation difficult.

In this manner, if *n* is the number of predictors, a total of $\sum_{k=2}^{10} 4nk = 216n$ univariate segments in the predictors' space will be generated. Therefore, the two variations that may derive from this combination are described below:

- Discrete Competitive Combination (DCC): let *m* be the discretization method used and *k* the number of segments in which each predictive feature *x* will be discretized, *m* and *k* are

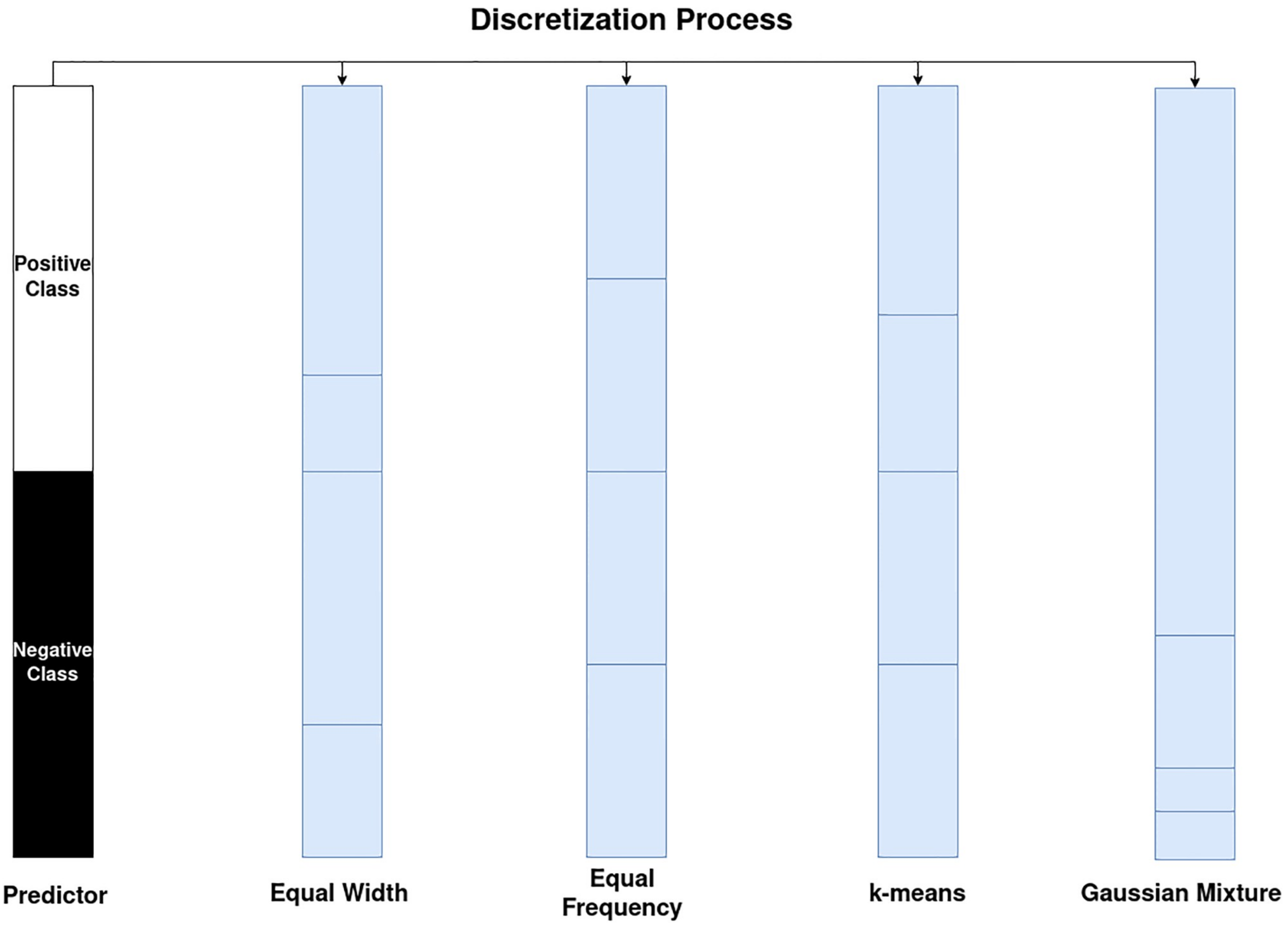

Fig 2. Intuitive discretization process. Intuitive differentiation between segments produced by different discretization techniques.

https://doi.org/10.1371/journal.pone.0289130.g002

chosen in such a manner that:

$$\underset{m,k}{\arg\max} \quad \mathrm{IV}(x, m, k), \quad \forall\, x \in \mathcal{X}$$

In this manner, the discretization methods *compete* with each other to determine the method/number combination of segments that maximizes each predictor's classification quality.

- Discrete Exhaustive Combination (DEC): in this variation, the four discretization methods chosen are included in a comprehensive manner. Let $\mathcal{M}$ be the set of discretization methods, k is chosen for each predictor $x \in \mathcal{X}$ such as:

$$\underset{k}{\arg\max} \quad \mathrm{IV}(x, m, k), \quad \forall\, m \in \mathcal{M}$$

The above forces the inclusion of each discretization method to provide the additional information that each method contributes individually to the combination.

After this selection, the generated segments enter as predictive features of the logistic regression (after WoE transformation), thus combining the additional information in each predictor generating a more powerful classifier instead of only using individual methods.

Finally, the logistic regression parameters are adjusted to measure the complete model's classification quality and enable comparison between methods. The algorithms Discrete Competitive Combination (DCC) and Discrete Exhaustive Combination (DEC) are shown as Algorithm 1 and 2, respectively. Each element of the set of discretization methods $\mathcal{M} = \{GMM, KME, EQF, EQD\}$, corresponds to the Gaussian mixture models, *k* means, equal frequency and equal distance, respectively. $\mathcal{X}$ refers to the matrix for continuous predictors. θ is the vector of logistic regression parameters. Vector $\vec{y}$ contains the positive and negative labels of the training set.

Algorithm 1 Discrete Competitive Combination (DCC)

```
Procedure DCC($\mathcal{X}$)
  $k \leftarrow 2$
  while $k \leq 10$ do
    for $m$ in $\mathcal{M}$ do
      Discretize per method $m$ in $k$ segments
    $k \leftarrow k + 1$
  Select best $k$ for each $m$
  Select best $m$
  Transform to WoE encoding using the training partition of set
  $(\mathcal{X}, \vec{y})$
  Estimate $\hat{\theta}|(\mathcal{X}, \vec{y})$
```

Algorithm 2 Discrete Exhaustive Combination (DEC)

```
procedure DEC($\mathcal{X}$)
  $k \leftarrow 2$
  while $k \leq 10$ do
    for $m$ in $\mathcal{M}$ do
      Discretize per method $m$ in $k$ segments
    $k \leftarrow k + 1$
  Select best $k$ for each $m$
  Transform to WoE encoding using the training partition of set
  $(\mathcal{X}, \vec{y})$
  Estimate $\hat{\theta}|(\mathcal{X}, \vec{y})$
```

5.2 Proposed methodology

The methodology used in this paper is illustrated in Fig 3. The proposed methodology allows comparison of the proposed combination's competitiveness through the systematic application of a standard machine learning process with the addition of pre-processing (discretization) and the innovative predictor selection (combination) algorithm. The steps of this methodology are listed below.

1. Selection of datasets related to credit risk.
2. Data cleansing and engineering: removal of inconsistencies, standardize names of predictors, eliminate categorical variables.
3. Discretization in *k* segments, *k* = 2, 3, . . ., 10 for each method (Gaussian mixture models, *k* means, equal frequency, and equal distance).
4. Predictor selection.

(a). Selection of the best method and number of segments for each variable (DCC, Algorithm 1).

(b). Selection of the best number of segments including a discretization for each method (DEC, Algorithm 2).

(c). Selection of the best number of segments only using one method at a time.

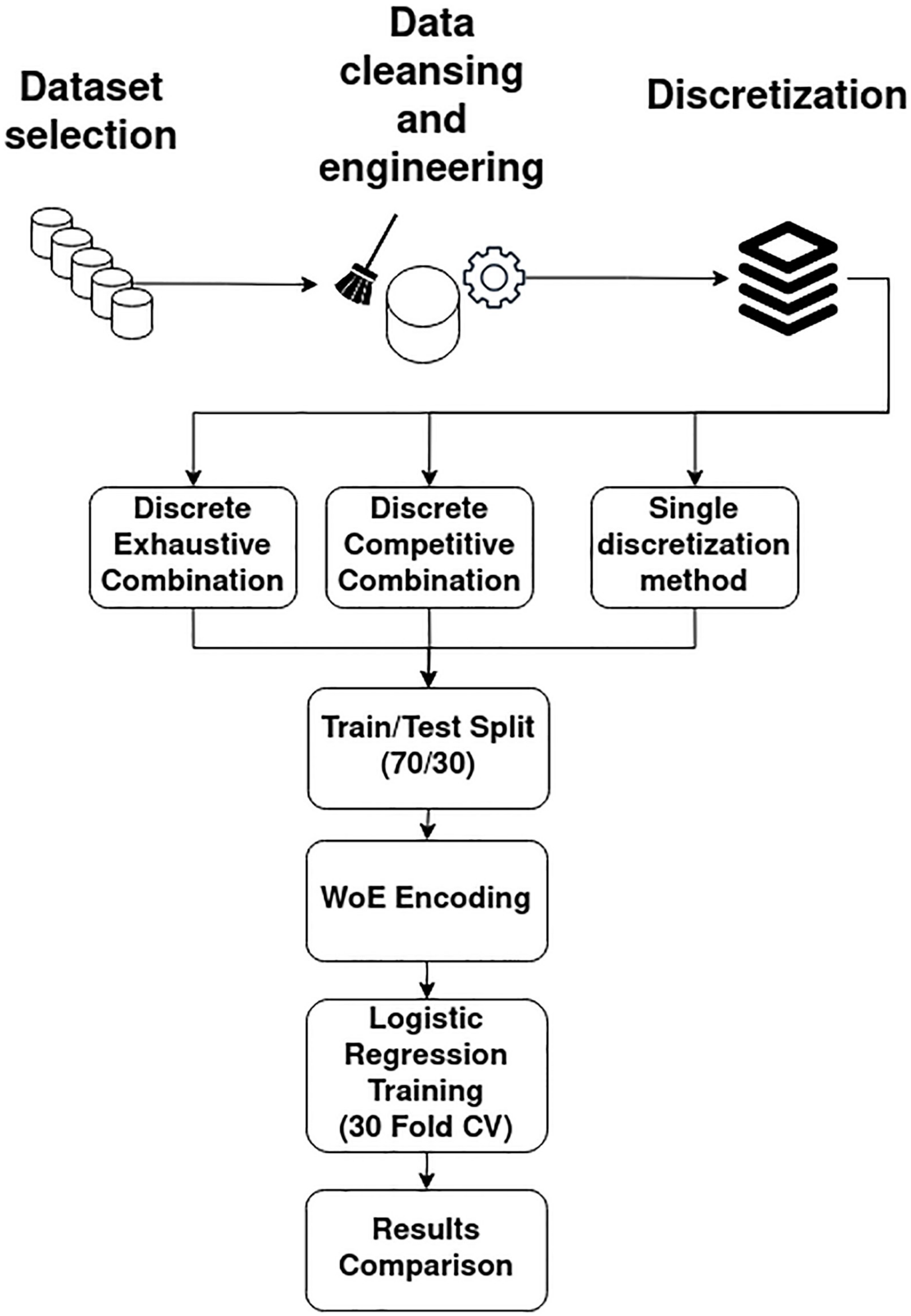
Dataset selection
Data cleansing and engineering
Discretization
Discrete Exhaustive Combination
Discrete Competitive Combination
Single discretization method
Train/Test Split (70/30)
WoE Encoding
Logistic Regression Training (30 Fold CV)
Results Comparison

Fig 3. Proposed methodology. Proposed methodology steps illustration diagram.

https://doi.org/10.1371/journal.pone.0289130.g003

5. Data partition (70% training and 30% validation). A commonly used ratio in practice [58].
6. WoE encoding.
7. 30-fold cross-validation logistic regression training. Literature suggests that ten folds usually suffice [59]. In order to strengthen the robustness of the experiment, 30 folds were used.
8. Comparison of results.

6 Results and discussion

In this section, the steps of the experimental framework used to generate the necessary data for comparing the single discretization methods against the proposed discretization combinations are described below.

1. Combination selecting the best number of segments and method for each variable, the methods compete with each other (DCC, Algorithm 1).
2. Combination selecting the best number of segments for each method and predictor, the inclusion of each discretization method in the combination (DEC, Algorithm 2) is forced.
3. Training using the best number of segments for each variable by the Gaussian mixture model (GMM) discretization method.
4. Training using the best number of segments for each variable by the *k* means (KME) discretization method.
5. Training using the best number of segments for each variable by the Equal frequency (EQF) discretization method.
6. Training using the best number of segments for each variable by the Equal distance (EQD) discretization method.

The main goal in the preceding stages was to evaluate the effectiveness of the proposed combinations in enhancing the predictive capabilities of a binary classifier. To test the proposed combinations, computational experiments were carried out on a system with an 11th Gen Intel Core i7-11800H @ 16x 4.6GHz and 32 GB of RAM, running the Linux operating system with Ubuntu Desktop release 22.04. All computations were performed using Python version 3.8.10. Specific packages and their corresponding versions are detailed in Section 4.

Statistics are computed by taking the validation samples outcome during the 30 cross validation folds for each of the 11 datasets, i.e., the sample size is $11 \times 30 = 330$ for each contrast at the method level and 30 at the individual dataset level.

Next, we analyze the performance for each discretization technique in terms of the average rank achieved for the various datasets. This is done using the Friedman and Holm tests, which have been widely considered in data science for comparing the performance of different methods and assessing statistical significance [60]. The AUC and KS metrics are considered for this analysis.

The Friedman test with Iman-Davenport correction evaluates the null hypothesis of equal ranks for all discretization techniques. The chi squared statistics obtained for the this test were 27.51 and 48.69 for the AUC and KS measures, respectively, rejecting the null hypothesis in both cases with a p value below 0.001. We can conclude that some discretization strategies are able to statistically outperform others.

The Holm post hoc test is used to perform pairwise comparisons between the discretization technique with the highest average rank and those remaining. For a given technique that

Table 2. Holm post hoc test for AUC.

Method	Rank	AUC	*p* value	$\alpha/(j-1)$	Outcome
DCC	1.636	0.769	-	-	-
DEC	1.818	0.769	0.820	0.050	not reject
QUA	3.182	0.755	0.053	0.025	not reject
GAU	3.909	0.635	0.004	0.017	reject
KME	4.864	0.575	0.000	0.013	reject
UNI	5.591	0.530	0.000	0.010	reject

AUC as performance metric for the various discretization strategies.

https://doi.org/10.1371/journal.pone.0289130.t002

Table 3. Holm post hoc test for KS.

Method	Rank	KS	*p* value	$\alpha/(j-1)$	Outcome
DCC	1.545	0.389	-	-	-
DEC	1.727	0.388	0.820	0.050	not reject
QUA	3.091	0.365	0.053	0.025	not reject
GAU	4.000	0.183	0.002	0.017	reject
KME	4.864	0.118	0.000	0.013	reject
UNI	5.773	0.038	0.000	0.010	reject

KS as performance metric for the various discretization strategies.

https://doi.org/10.1371/journal.pone.0289130.t003

achieved the j-th position in the rank, each pairwise test is rejected when the p value is below a significance threshold $\alpha/(j-1)$, with $\alpha = 5\%$ and $j = 2, \ldots, 6$. The results for the Holm post hoc test are provided in Tables 2 and 3 for the AUC and KS measures, respectively. For each discretization technique, these tables also include the average rank and the average value for each metric.

It can be observed in Tables 2 and 3 that the test achieves consistent results for the two metrics. Results show that DCC, DEC, and QUA can be considered statistically equivalent given the null hypothesis rejection of the Holm test and, simultaneously, statistically different from their counterparts GAU, KME, and UNI.

Analyzing in detail, the method with the best average performance is DCC, statistically outperforming GAU, KME, and UNI. On the other hand, the DEC method achieves very similar performance compared to DCC, concluding that these two discretization techniques generally lead to the best results, even if they perform closely to QUA; additional qualitative gains for not restricting the segments to equal frequency can be achieved by using DCC and DEC instead of QUA.

The statistical evidence demonstrates that using a combined discretization strategy based on clustering is a feasible alternative and maintains the model's interpretability and competitiveness for its classification quality.

As for the computational cost (indicated in Table 4), the DEC combination naturally has a higher cost due to the method's comprehensiveness. However, the approximate four-second average has not been deemed a limitation for its implementation. Additionally, the DCC combination does not have significantly different computational times compared with its univariate counterparts.

Table 4. Computational time.

Method	Mean	95% CI
DCC	1.1319	[0.9224, 1.3414]
DEC	4.1383	[3.3257, 4.9508]
EQD	1.1066	[0.8968, 1.3165]
EQF	1.1469	[0.935, 1.3588]
GMM	1.1777	[0.9602, 1.3951]
KME	1.1777	[0.9614, 1.3939]

Seconds per fold comparative chart.

https://doi.org/10.1371/journal.pone.0289130.t004

7 Conclusions and future work

This paper proposed a combination of discretization methods that improve the classification quality of logistic regression as a binary classifier. The combination of discretization methods was conducted using the continuous predictor variables available on each dataset. Subsequently, performance metrics of the proposed combination were obtained in eleven different and diverse datasets through computation experimentation. Experimental results showed that the proposed method has successfully achieved the objectives pursued. Firstly, an alternative discretization through Gaussian mixture models and *k* means proved its competitiveness by maintaining an appropriate classification quality and supplementing widely used conventional discretization methods. Furthermore, combining discretization methods increases classification quality compared to individual methods when applied in various credit risk datasets. On the other hand, since the created discrete segments are univariate, the classifier's explainability is not affected.

The proposed combination method could potentially fill a void in the fields of application where explainability is crucial for a model based on statistical learning simultaneously with permanent efforts to improve classification quality. Similarly, the proposed method provides a balance between modern assembly techniques and the simplicity of linear models, thus benefiting from both approaches' advantages.

Future work concerns assembling discretization methods through genetic algorithms to determine the optimal combination of clusters and methods to be included in the final selection since the search is performed thoroughly and less efficiently in this stage. Alternatively, increasing the number of discretization techniques is plausible, provided that the original explainability objectives are preserved. Findings suggest the proposed method can be successfully incorporated as a discretization alternative in the development of credit scoring models without a significant impact on operation, explainability, and implementation and with a significant contribution to better performance of a logistic regression model used in a classification problem. Finally, the applications of the proposed method scale to every industry where a binary classification problem arises. Examples such as predictive maintenance and predictive quality in manufacturing, customer analytics in retail, or disease detection in healthcare are a sample of compelling use cases worth being explored as an extension of the present work.

Author Contributions

Conceptualization: José G. Fuentes Cabrera, Hugo A. Pérez Vicente, Jonás Velasco.

Data curation: Sebastián Maldonado.

Formal analysis: Sebastián Maldonado, Jonás Velasco.

Funding acquisition: Hugo A. Pérez Vicente.

Investigation: José G. Fuentes Cabrera, Hugo A. Pérez Vicente.

Methodology: José G. Fuentes Cabrera, Sebastián Maldonado, Jonás Velasco.

Project administration: Hugo A. Pérez Vicente.

Resources: José G. Fuentes Cabrera.

Software: José G. Fuentes Cabrera.

Supervision: Jonás Velasco.

Validation: Hugo A. Pérez Vicente, Jonás Velasco.

Visualization: Hugo A. Pérez Vicente, Sebastián Maldonado.

Writing – original draft: José G. Fuentes Cabrera, Hugo A. Pérez Vicente.

Writing – review & editing: Jonás Velasco.

References

1. Lee I. and Shin Y. Machine learning for enterprises: Applications, algorithm selection, and challenges. *Business Horizons*. 63, 157–170 (2020) https://doi.org/10.1016/j.bushor.2019.10.005
2. Hassani H., Huang X., Silva E. and Ghodsi M. Deep learning and implementations in banking. *Annals Of Data Science*. 7, 433–446 (2020) https://doi.org/10.1007/s40745-020-00300-1
3. Ala'raj M., Majdalawieh M. and Abbod M. Improving binary classification using filtering based on *k*-NN proximity graphs. *Journal Of Big Data*. 7, 1–18 (2020)
4. Karaca K. Values and inductive risk in machine learning modelling: the case of binary classification models. *European Journal For Philosophy Of Science*. 11, 1–27 (2021) https://doi.org/10.1007/s13194-021-00405-1
5. Santoso L., Singh B., Rajest S., Regin R. and Kadhim K. A Genetic Programming Approach to Binary Classification Problem. *EAI Endorsed Transactions On Energy Web*. 8, e11 (2020)
6. Wirth, R. and Hipp, J. CRISP-DM: Towards a standard process model for data mining. *Proceedings Of The 4th International Conference On The Practical Applications Of Knowledge Discovery And Data Mining*. 1 (2000)
7. Studer S., Bui T., Drescher C., Hanuschkin A., Winkler L., Peters S. and Müller K. Towards CRISP-ML (Q): a machine learning process model with quality assurance methodology. *Machine Learning And Knowledge Extraction*. 3, 392–413 (2021) https://doi.org/10.3390/make3020020
8. Ashayeri, C., Korjani, M. and Ershaghi, I. Using the Adaptive Variable Structure Regression Approach in Data Selection and Data Preparation for Improving Machine Learning-Based Performance Prediction in Unconventional Plays. *SPE/AAPG/SEG Unconventional Resources Technology Conference*. (2021)
9. Whang S. and Lee J. Data collection and quality challenges for deep learning. *Proceedings Of The VLDB Endowment*. 13, 3429–3432 (2020) https://doi.org/10.14778/3415478.3415562
10. Siddiqi N. Credit risk scorecards: developing and implementing intelligent credit scoring. (John Wiley and Sons,2012)
11. Adeodato, P., Salazar, D., Gallindo, L., Sá, Á. and Souza, S. Continuous variables segmentation and reordering for optimal performance on binary classification tasks. *2014 International Joint Conference On Neural Networks (IJCNN)*. pp. 3720-3725 (2014)
12. Garcia S., Luengo J., Sáez J., Lopez V. and Herrera F. A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning. *IEEE Transactions On Knowledge And Data Engineering*. 25, 734–750 (2012) https://doi.org/10.1109/TKDE.2012.35
13. Mironchyk, P. and Tchistiakov, V. Monotone optimal binning algorithm for credit risk modeling. (Utrecht: Working Paper,2017)
14. Navas-Palencia, G. Optimal binning: mathematical programming formulation. *ArXiv Preprint ArXiv:2001.08025*. (2020)
15. Lund, B., Solutions, M. and Charlotte, N. Binning of Predictors for the Cumulative Logit Model. (2017)
16. Yang, Y. and Webb, G. A comparative study of discretization methods for naive-bayes classifiers. *Proceedings Of PKAW*. 2002 (2002)

17. Silva, W., Fernandes, K. and Cardoso, J. How to produce complementary explanations using an ensemble model. *2019 International Joint Conference On Neural Networks (IJCNN)*. pp. 1-8 (2019)

18. Ogbuabor G. and Ugwoke F. Clustering algorithm for a healthcare dataset using silhouette score value. *International Journal Of Computer Science and Information Technology*. 10, 27–37 (2018) https://doi.org/10.5121/ijcsit.2018.10203

19. Weng S., Reps J., Kai J., Garibaldi J. & Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data?. *PloS One*. 12, e0174944 (2017) https://doi.org/10.1371/journal.pone.0174944 PMID: 28376093

20. Ai Y., He F., Lancaster E. & Lee J. Application of machine learning for multi-community COVID-19 outbreak predictions with wastewater surveillance. *Plos One*. 17, e0277154 (2022) https://doi.org/10.1371/journal.pone.0277154 PMID: 36355921

21. Hartigan J. Clustering algorithms. (Wiley,1975)

22. Moore, A. Clustering with gaussian mixtures. URL Http://www.Autonlab.Org/tutorials/gmm.Html. *Tutorial Slides*. (2004)

23. Nisbet R., Elder J. and Miner G. Handbook of statistical analysis and data mining applications. (Academic Press,2009)

24. Massey F Jr. The Kolmogorov-Smirnov test for goodness of fit. *Journal Of The American Statistical Association*. 46, 68–78 (1951) https://doi.org/10.1080/01621459.1951.10500769

25. Casella G. and Berger R. Statistical inference. (Duxbury Pacific Grove, CA,2002)

26. Kullback, S. Information theory and statistics. (Courier Corporation,1997)

27. Fuentes-Cabrera, J. and Pérez-Vicente, H. Credit scoring model for payroll issuers: a real case. *Mexican International Conference On Artificial Intelligence*. pp. 547-559 (2015)

28. Ai J., Brockett P. and Golden L. Assessing consumer fraud risk in insurance claims: An unsupervised learning technique using discrete and continuous predictor variables. *North American Actuarial Journal*. 13, 438–458 (2009) https://doi.org/10.1080/10920277.2009.10597568

29. Nehrebecka N. Predicting the default risk of companies. Comparison of credit scoring models: LOGIT vs Support Vector Machines. *Econometrics*. 22, 54–73 (2018) https://doi.org/10.15611/eada.2018.2.05

30. Melchiori, M. Credit Scoring Applied to Special Case. *Available At SSRN 3456130*. (2019)

31. Bunker, R., Zhang, W. and Naeem, M. Improving a Credit Scoring Model by Incorporating Bank Statement Derived Features. *ArXiv Preprint ArXiv:1611.00252*. (2016)

32. Nehrebecka N. "COVID-19: stress-testing non-financial companies: a macroprudential perspective. The experience of Poland," *Eurasian Economic Review*, vol. 11, no. 2, pp. 283–319, 2021. https://doi.org/10.1007/s40822-020-00163-0

33. Wang C., Han D., Liu Q. and Luo S. A deep learning approach for credit scoring of peer-to-peer lending using attention mechanism LSTM. *IEEE Access*. 7 pp. 2161–2168 (2018) https://doi.org/10.1109/ACCESS.2018.2887138

34. Zhang Z., Niu K. and Liu Y. A deep learning based online credit scoring model for P2P lending. *IEEE Access*. 8 pp. 177307–177317 (2020) https://doi.org/10.1109/ACCESS.2020.3027337

35. Guo S., He H. and Huang X. A multi-stage self-adaptive classifier ensemble model with application in credit scoring. *IEEE Access*. 7 pp. 78549–78559 (2019) https://doi.org/10.1109/ACCESS.2019.2922676

36. Dua, D. and Graff, C. UCI Machine Learning Repository. (University of California, Irvine, School of Information,2017), http://archive.ics.uci.edu/ml

37. Eggermont, J., Kok, J. and Kosters, W. Genetic programming for data classification: Partitioning the search space. *Proceedings Of The 2004 ACM Symposium On Applied Computing*. pp. 1001-1005 (2004)

38. Drummond, C., Holte, R. and Others C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. *Workshop On Learning From Imbalanced Datasets II*. 11 pp. 1-8 (2003)

39. Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., et al. Scikit-learn: Machine Learning in Python. *Journal Of Machine Learning Research*. 12 pp. 2825–2830 (2011)

40. Szepannek, G. An Overview on the Landscape of R Packages for Credit Scoring. *ArXiv Preprint ArXiv:2006.11835*. (2020)

41. Cano A., Luna J., Gibaja E. and Ventura S. LAIM discretization for multi-label data. *Information Sciences*. 330 pp. 370–384 (2016) https://doi.org/10.1016/j.ins.2015.10.032

42. Saia, R., Carta, S., Recupero, D., Fenu, G. and Saia, M. A Discretized Enriched Technique to Enhance Machine Learning Performance in Credit Scoring. *KDIR*. pp. 202-213 (2019)

43. Wójciak, M. and Dubicka, A. Empirical Comparison of Methods of Data Discretization in Learning Probabilistic Models. *Advances In Computer Science Research*. (2018)

44. Hsieh N. and Hung L. A data driven ensemble classifier for credit scoring analysis. *Expert Systems With Applications*. 37, 534–545 (2010) https://doi.org/10.1016/j.eswa.2009.05.059

45. Vejkanchana, N. and Kucharoen, P. Continuous Variable Binning Algorithm to Maximize Information Value Using Genetic Algorithm. *International Conference On Applied Informatics*. pp. 158-172 (2019)

46. Raymaekers J., Verbeke W. and Verdonck T. Weight-of-evidence through shrinkage and spline binning for interpretable nonlinear classification. *Applied Soft Computing*. 115 pp. 108160 (2022) https://doi.org/10.1016/j.asoc.2021.108160

47. Raymaekers, J., Verbeke, W. and Verdonck, T. Weight-of-evidence 2.0 with shrinkage and spline-binning. *ArXiv Preprint ArXiv:2101.01494*. (2021)

48. A. Iftikhar, M. Alam, R. Ahmed, S. Musa, and M. M. Su'ud, *Risk prediction by using artificial neural network in global software development*, *Computational intelligence and neuroscience*, vol. 2021, 2021, Hindawi.

49. Iftikhar A. Musa S. Alam M. M. Ahmed R. Su'ud M. M. Khan L. M. et al, Risk Classification in Global Software Development Using a Machine Learning Approach: A Result Comparison of Support Vector Machine and K-Nearest Neighbor Algorithms, *Journal of Information Technology Research (JITR)*, vol. 15, no. 1, pp. 1–21, 2022, IGI Global. https://doi.org/10.4018/JITR.299385

50. Khan, S. Microfinance Loan Credit Scoring. (2020), https://www.kaggle.com/shahrukhkhan/microfinance-loan-credit-scoring

51. Adams, T. Give me some credit. *Observer*. pp. 1-2 (2004)

52. Baesens B. Analytics in a big data world: The essential guide to data science and its applications. (John Wiley and Sons,2014)

53. George, N. All Lending Club loan data version 6, february 2018. *See* www.Kaggle.Com/wordsforthewise/lending-club *(accessed 1 October 2018)*. (2018)

54. Zięba M., Tomczak S. and Tomczak J. Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction. *Expert Systems With Applications*. 58 pp. 93–101 (2016) https://doi.org/10.1016/j.eswa.2016.04.001

55. Yeh I. and Lien C. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems With Applications*. 36, 2473–2480 (2009) https://doi.org/10.1016/j.eswa.2007.12.020

56. Krijthe, J. and Maaten, L. Rtsne: T-distributed stochastic neighbor embedding using Barnes-Hut implementation. *R Package Version 0.13*, *URL* https://github.Com/jkrijthe/Rtsne. (2015)

57. Chen, T. and Guestrin, C. XGBoost: reliable large-scale tree boosting system. arXiv. *2016a. ISSN*. pp. 0146-4833 (2016)

58. Joseph V. R. "Optimal ratio for data splitting," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 15, no. 4, pp. 531–538, 2022. https://doi.org/10.1002/sam.11583

59. Marcot B. G. and Hanea A. M. "What is an optimal value of k in k-fold cross-validation in discrete Bayesian network analysis?" *Computational Statistics*, vol. 36, no. 3, pp. 2009–2031, 2021. https://doi.org/10.1007/s00180-020-00999-9

60. Demšar J. Statistical comparisons of classifiers over multiple data set. *Journal Of Machine Learning Research*. 7 pp. 1–30 (2006)