The Effect of Skewness and Error with Discretization on the Predictive Accuracy and Correlation of the Exact Tree

Rouvinen, Robert

The Effect of Skewness and Error with Discretization on the Predictive Accuracy and Correlation of the Exact Tree

Robert Rouvinen

Master's Thesis Psychology,

Methodology and Statistics Unit, Institute of Psychology

Faculty of Social and Behavioral Sciences, Leiden University

Date: December 27. 2022

Student number: [redacted]

Supervisor: Dr. Elise Dusseldorp

Table of Contents

Abstract

The purpose of this simulation study was to understand the Exact Tree algorithm and how different methods of discretization and changes in data, such as adding skewness and more systematic error (noise), affect its predictive accuracy. Furthermore, the second aim was to understand the effect of discretization on the measurement agreement (Spearman rho correlation coefficient) between discretized data and the original simulated data with varying levels of skewness and noise. The simulation included 600 data sets with a binary outcome variable and ten continuous predictor variables. The discretization methods compared in this study were Equal Frequency discretization, Optimal Binning and rounding. The predictors were varied at 3 levels of skewness (skewness = 0, -0.2 and -0.9) and the noise that the predictive model used for generating the outcome variable was varied at two levels. The results suggest the discretization methods differ in the Exact Tree predictive accuracy, but only by a small difference (η^2 = 0.008) and do not seem to affect the overall performance of Exact Tree. The effect of the discretization methods on the accuracy depended on the level of skewness, with a significant interaction effect but small effect size (η^2 = 0.003). The best predictive accuracy and least difference between methods was seen in the more severe condition (skewness = -0-9). The Optimal Binning method had the poorest accuracy in the mild and no skewness conditions. Furthermore, the Optimal Binning method resulted in better measurement agreement between the non-discretized data and the discretized data compared to Equal Frequency and rounding. Overall, the small effects sizes suggest that the choice between Equal Frequency, Optimal binning or rounding does not affect the overall performance of the Exact Tree. Moreover, future research directions are suggested, such as comparing more discretization method with the Exact Tree.

Acknowledgement

This master thesis was performed with the supervision of Dr. Elise Dusseldorp.

1. Introduction

In scientific research and other fields data can become very large and complex, this is especially the case with the rise of the "big data" industry. Combined with current technological advances in computation power, gathering large amounts of data and performing complex calculations have become far easier. With technological progress on the hardware part there has also been progress on data analysis methods. Many machine learning tools have been introduced to the public and are used in many areas of life. One of the many machine learning tools that have been popularized are decision trees. These decision trees, in contrast to other machine learning tools are far easier to interpret due to their ability to be visualized into a tree structure. Generally, machine learning tools need to be trained on data, usually the more data the better the algorithm will perform in terms of predictive accuracy. However, as mentioned before the more data the more computation power is needed. Furthermore, as methods have become more available to use for the general public it is beneficial that users can interpret the results well. Therefore, it is relevant to further study decisions tree algorithms due to their usefulness regarding interpretability and wide use. Yet there are ways to take full use of larger amounts of data and transform it into a less computationally intensive state, such as discretization.

Decision trees are a common tool in machine learning and one of the most used are classification and regression trees (CART). CART are divided into regression trees which are for continuous outcome variables and classification trees that are used for categorical outcome variables. The trees themselves use the same recursive partitioning algorithm but the outcome variables they predict determine the type of problem the tree is meant to answer. CART have become widely used especially due to their useful qualities such as allowing for selection of predictors automatically, discovering interaction effects and they can be considered non-parametric (Ma, 2018). Furthermore, they are useful due to the ease at which they can be interpreted, but only when trees are not too large (James, Witten, Hastie & Tibshirani, 2014). Exact Tree is a decision tree algorithm that unlike many other tree algorithms is not a greedy algorithm. Instead, the Exact Tree is characterized by its ability to find the guaranteed optimal tree with the use of dynamic programming (Meulman, Dusseldorp & Van Os, 2011). The Exact Tree algorithm is computationally intensive and requires a lot of time when predictor variables have many unique values.

Discretization methods are used in a large range of fields from finance (Wang, Shang, Huang & Feng, 2013) to medical research (Sahni, Müller, Jansen, Shephard & Taylor, 2006).

Discretization methods have the goal of reducing the search space. This reduction is realized by taking a continuous variable and transforming it into a nominal variable with a set amount of categories, which means it is a data reduction method. Generally, this is beneficial for algorithms since it makes the computation faster and more efficient. Furthermore, the nominal variables that discretization results in can be better suited for interpretation compared to non-discretized continuous variables (Liu, Hussain, Tan & Dash , 2002). With the benefit of faster computation, it is a very useful method but there is a trade-off between the computation and the loss of data that comes with discretization (Kliegr, 2017). On the one hand, using discretization can improve the speed of computation as there are less unique data points, but it also reduces all data points in a prespecified interval to a single value.

Two of the most common ways different discretization methods are grouped by are static to dynamic and supervised to unsupervised methods. Static methods refer to when the discretization is performed before the machine learning method is applied. Dynamic in turn, means the discretization is performed in conjunction with the learning (Ramirez-Gallego et al., 2015). In supervised discretization methods information about the outcome variable, often referred to as the class label, is used in the process of choosing the categories into which the predictor is discretized. Unsupervised methods in turn does not use information about the class label (Ramirez-Gallego et al., 2015). In unsupervised discretization the data is discretized into categories based on a pre-set criteria. Example of such criteria is dividing a variable into categories that have an equal frequency of observations. There are more ways to divide the different discretization methods, but these two are the most relevant for the current research.

The research so far on the Exact Tree method did not examine the issue of how well the algorithm and the chosen method of discretization works on skewed data. Furthermore, after a literature search, research on this area is very limited and therefore it is interesting to focus on skewness and its effects of the algorithm’s accuracy. Moreover, in psychological research it has been shown that distributions often have quite varying levels of skewness (Cain, Zhang & Yuan, 2017). Therefore, it is important to understand what effect skewness has on different statistical tools that researchers have at their disposal. The primary objective of this paper is the investigation of two factors: skewness and systematic error (noise) on the discretization methods used in combination with Exact Tree. Prior research made on the impact of discretization on common data distributions developed a heuristic according to which generally classification error rates increase as the level of kurtosis and skewness grows

(Ismail & Ciesielski, 2003). This collection of common data distributions included various levels of skewness and kurtosis; in our study, we focused on varying the skewness but kurtosis was kept at a constant. In our analysis we also examined the level of agreement between the original non-discretized data and the discretized data. The discretization methods which were examined were the Optimal binning method, developed concurrently with Exact Tree, Equal Frequency discretization and rounding. Rounding is not considered an actual discretization method, but it reduces the data and makes computation faster but to a far lesser extent than discretization. Furthermore, it is unclear how generalizable the findings from previous research are to the Exact Tree, which justifies further research on the topic of the effects of discretization on skewed data.

To evaluate the accuracy of the algorithm the misclassification rate (MCR) was taken for both the train data and test data predictions. The choice was made because the optimization method that is implemented in Exact Tree uses as default the training data However, we do not know the effect this optimization has on the predictive accuracy thus collecting the misclassification rate (MCR) for the training data was a relevant choice. Moreover, Brier scores were calculated as an alternative method of evaluating predictive accuracy.

The first research question was to determine how the combination of different methods of discretization and changes in data, such as adding skewness and more systematic error (noise), affect the predictive accuracy of Exact Tree. The second research question was to determine the effect of discretization on the measurement agreement between discretized data and the original simulated data with varied levels of skewness and noise.

2. Theoretical framework

CART

The CART method, which was developed by Breiman et al. in 1984, is used to build highly interpretable decision trees. In CART objects are split based on how they relate to the outcome variable (regression and classification). The tree starts at the root node as one group containing all objects. Further down, smaller groups are created and these only contain a few objects.

For regression trees CART works by partitioning the data in a way that minimizes the relative sum of squared errors in the nodes, this is the index used to measure the purity of a node in regression trees. For classification trees a purity index that is often used is the Gini index, this splitting criterion states that the cut-point and splitting variable that bring the largest reduction to impurity are chosen for the next split (James, Witten, Hastie & Tibshirani, 2014). However, this method is also vulnerable to selection bias, as the Gini index tends to give more weight for variables with more categories (Strobl, Boulesteix & Augustin, 2007). Classification trees can use other impurity measures, often times a misclassification rate or entropy is used. For example, in the case of the Exact Tree the misclassification rate is used.

Splits occur first at the parent node, in the example this is the highest split with Petal length in Figure 1. Then the splits occur iteratively lower at daughter nodes until finally the terminal node, which separates a group of observations by how they are predicted on the outcome variable. The splitting rules aim at dividing the objects based on increasing the homogeneity of the objects to the outcome variable (Breiman, Friedman, Olshen & Stone, 1984). Stopping rules also determine when and how a split occurs, one rule that is often used is to set a minimum number of objects in any given node.

Furthermore, there are rules on how object or subjects are assigned to the terminal node or leaf, these rules are different for classification and regression trees. For classification trees the rule is called the “majority vote”, this works by assigning each object in a node to the most frequently occurring category of the outcome. For regression trees it is determined by taking the mean value of the outcome variable for all objects within a node (James, Witten, Hastie & Tibshirani, 2014).

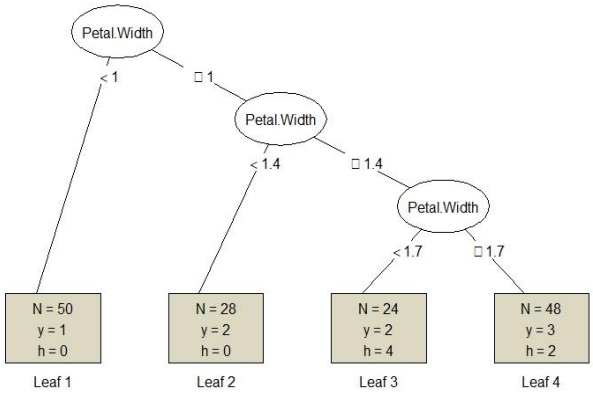
The Exact Tree

The Exact Tree algorithm developed by Van Os uses dynamic programming to build a globally optimal decision tree (Van Os, 2000). The algorithm delivers the optimal tree given the sample size stays under 500, and the number of predictor variables is not too large and the depth of the tree is under or equal to seven. The general aim of the algorithm is to improve the accuracy, stability and interpretability without increasing the complexity of the decision tree (Meulman, Dusseldorp & Van Os, 2011).

The dynamic programming method works by searching every possible split at a single node, and then finding the optimal subtree for both nodes at each split. This search is repeated at every split and nodes of the tree. At each next step when the algorithm moves to the

subsequent node, the best group of subtrees is evaluated and picked. The purity index, which is a criterion that decides when splitting should stop, for the Exact Tree in the case of regression is measured by what split minimizes the sum of squared error. In the case of classification trees purity is measured by what split minimizes the misclassification rate (Meulman, Dusseldorp & Van Os, 2011). To demonstrate the Exact Tree method, it was run on the Iris data, which is commonly used in statistical analyses as an example (Fisher, 1936). The data includes different species of the Iris flower and their plant characteristics (Anderson, 1936). In this example we predicted species of Iris flowers based on four variables, the plants' petal length, petal width, sepal width and sepal length. Figure 1 shows the result of the Exact Tree method for the Iris data. The terminal nodes (gray boxes) show how the objects divide on the outcome variable, the species of Iris by: 1(Setosa), 2 (Versicolor) and 3(Virginica). The circles show which variables determined the splits at each node. Based on this example the most important variable in predicting species was Petal width. In this paper we created 600 Exact Tree models each of them was a unique tree such as the tree diagram shown in the Figure 1.

Figure 1 Results of Exact Tree for the Iris data. Predicting species of flower based on plant variables: petal width, petal length, sepal length and sepal width. Model settings were the same as with the other Exact Tree analyses in this paper, but the Iris data was not discretized.



Discretization methods

In this study, we compared three discretization methods. The first discretization method is called Optimal Binning, which was developed in conjunction with the Exact Tree algorithm (Meulman, Dusseldorp & Van Os, 2011). The method finds a way to distribute the discretized observations in an optimal way for a pre-defined number of categories. The method does this by first ordering the observations on a predictor variable from low to high

and then dividing the observations in a sequence of categories so that the squared error between the observations and their respective category means is minimal. Moreover, the method is implemented using the Exact Tree package developed for R (Meulman, Dusseldorp & Van Os, 2011).

The second method was Equal Frequency discretization where the data is discretized so that all of the desired number of categories have and equal number of observations, forming a uniform distribution. This method is quite arbitrary and does not reveal new information about the data, however its strength is in its simplicity to understand. Equal Frequency discretization is a part of the unsupervised discretization methods, meaning it does not use category information when transforming continuous variables into nominal variables. Generally, this method has been shown to be less sensitive to outliers. However, a weakness of Equal Frequency method is that in some cases observations with same value can be placed in different categories. Moreover, a drawback is that it can be challenging to determine the best amount categories to discretize the data into, resulting in the choice being largely arbitrary often (Hacibeyoglu & Ibrahim, 2018). To perform equal frequency discretization the package funModeling is used in R (Casas, 2020).

The third method used was rounding; here the observations are not divided into categories that are set before the procedure, but the observations are rounded to a certain decimal. This method was chosen as an alternative to no discretization, since it does speed up the process but without high amount of data reduction that the above-mentioned methods have. Rounding off values can often improve the performance of machine learning model because it reduces noise and prevents overfitting. Therefore, rounding can result in the model being more generalizable (Senavirathne & Torra, 2019).

3. Evaluation Criteria

Predictive accuracy

To evaluate the performance of classification methods, the Area under the ROC curve (AUC) and Brier score were calculated. The ROC refers to the receiving operating characteristic, which is a term first used to classify signals in combat. The curve plots a model's rate of false-positives to true-positives. Generally, an area under the ROC curve of 1 is perfect, that it misses all the false positives, but also all of the true positives. Likewise, a rate of 0 is very

weak and it picks all the false positives as true positives. A “guessing model”, that for example randomly predicts either 1 or 2 would have exactly 0.50 as its area on both sides of the curve. Therefore, the evaluation of an AUC value returned suggests that a binary outcome value is sampled randomly with the probability of P(Y= 1) of 0.5 (Goldstein-Greenwood, 2022).

The Brier score is a metric developed by Glen Brier in 1950 to verify weather forecast predictions (Brier, 1950). The Brier score can be used as a metric to evaluate predictive accuracy of a model. A Brier score that is lower indicates better model accuracy, and higher score indicates poorer accuracy; the scores range from 0 to 1, with 0 being perfect accuracy (Rufibach, 2010). In this study of Exact Tree the Brier score is calculated by the mean square error between the outcomes and the predicted class probability of classifying the object as 1 instead of 0 (Harrison, Brady, Parry, Carpenter & Rowan, 2006).

In addition, we used the misclassification rate (MCR) as a measure of predictive accuracy. The misclassification rate is often used to predict model prediction accuracy in classification trees, which is why it is used in this paper.

Measurement Agreement

The measurement agreement between the discretized data and the original data was evaluated using the Spearman rho correlation coefficient. The Spearman rho is known to be better suited when the predictors that are examined are skewed compared to other correlation measures (Mukaka, 2012).

4. Simulation

For this study several scenarios of skewed distributions were simulated that were discretized with the three different discretization methods. The Exact Tree algorithm was then performed on the binned data predicting a binary outcome variable. The methods for this simulation are inspired by Csorba (2020) where the simulated data was based on properties seen in empirical datasets in psychology that are often skewed (Csorba, 2020). To simulate the skewed data the package PearsonDS was used in R (Becker & Klößner, 2022).

For the simulation we examined two design factors, the first one was skewness, varied on three levels and the second factor is systematic error (or noise), varied on two levels. To visualize the design factors, Figure 2 shows how they are combined per cell of the design.

Skewness was varied by, skewness = 0 (no skewness), skewness = -0.2 (mild skewness) and skewness = -0.9 (more severe skewness). The levels of systematic were two-fold, in the first condition there were no predictors that were noise variables and in the high noise condition five of the ten predictors were noise variables (see below for the details).

The dataset had a training sample size of 250 and the test data sample size was 1750. The training dataset should not be too large since it slows down the analysis significantly, as the algorithm is trained based on it. The test dataset can be far larger since it doesn't require the Exact tree algorithm to be used on it. Furthermore, a larger test dataset allows for a better evaluation of how well the algorithm can predict the outcome. However, in this analysis having a larger test dataset still slows down the iterations since the discretization procedures is slower for larger data. The size of the train and test datasets were set following the limitations mentioned by the creators of the Exact Tree method to keep the computation time feasible. The predictor variables were all continuous and were discretized whereas the outcome variable was a binary variable.

The population model for the outcome variable with low noise was a logistic regression model, the model is seen in Equation 1. The model predicted a binary outcome variable, with 10 predictors drawn from three types of distributions, two of which were skewed and one was normally distributed. More detail of these distributions are given below. Furthermore, two interaction terms were added into the model between the variables x_2, x_3 and x_3, x_4. The Beta values or weights were determined so that some have lower and higher weights and the interaction terms were chosen for those with higher weights so that its effect would be more noticeable.

The population model for the high noise condition in the simulation can be seen in Equation 2. In this condition only five of the ten predictors were used to create the binary outcome variable but the same interaction terms were kept as in the low noise condition. Furthermore, to generate the binary outcome variable in the high noise condition error was added to the model. The error was added so that the signal to noise ratio would be 1:2. The error (e) was drawn form a normal distribution with the standard deviation determined by Equation 3, which caused the signal to noise ratio to be 1:2.

Equation 1

$$log\left(\frac{p_i}{1-p_i}\right) = -1 + 0.1x_1 + 3.5x_2 + -0.5x_3 + 3.1x_4 + -0.1x_5 + 0.3x_6 + -0.5x_7 + -0.6x_8 + 0.3x_9 + -0.4x_{10} + 3.5x_2x_3 + 3.1x_3x_4 + e$$

Equation 2

$$log\left(\frac{p_i}{1-p_i}\right) = -1 + 0.1x_1 + 3.5x_2 + -0.5x_3 + 3.1x_4 + -0.1x_5 + 3.5x_2x_3 + 3.1x_3x_4 + e$$

Equation 3

$$sd = \sqrt{2Var\left(log\left(\frac{p_i}{1-p_i}\right)\right)}$$

In each simulation the skewness was the same for all predictor variables in the dataset. To generate a predictor in the more severe skewness condition a random sample was drawn from a Pearson distribution with a mean of 0, variance = 1, skewness = -.9 and kurtosis = 3. For the mild skewness condition the random sample was drawn from a Pearson distribution with a mean = 0, variance = 1, skewness = -.2 and kurtosis = 3. The last condition of no skewness was drawn from a normal distribution with a mean = 0, standard deviation = 1. It is to be noted that kurtosis was not varied and was kept at 3, which is what it is for a normal distribution. This process is repeated until 10 predictors were generated, which means that the intercorrelation between predictors was approximately zero.

Table 1 The table of the 3 x2 simulation design. The table shows which simulation conditions are used per cell of the design.

Low noise conditions	High noise conditions
No skewness + Low noise	No skewness + High noise
Mild skewness + Low noise	Mild skewness + High noise
Severe skewness + Low noise	Severe skewness + High noise

.

Within each cell of the simulation study, which can be seen in Table 1, all three of the discretization methods were applied to the predictors. In each cell of the design, we also evaluated the goodness of fit measures AUC and Brier score, this was done to determine if the data generation process went according to the simulation plan.

The number of categories the simulated datasets are discretized into was determined by previous research on the Exact Tree algorithm and its predictive accuracy. The results suggest that using 25 categories results in the best prediction accuracy (Meulman, Dusseldorp & Van Os, 2011). Therefore, 25 categories were used for the two compared discretization methods. Moreover, the simulated data was rounded to one decimal place in the rounding condition. This choice was taken due to the simulated data having small values where rounding to higher decimal places does not reduce the data by much.

Implementation of Evaluation Criteria

To evaluate if the data generation process went as planned, we checked whether the high noise condition actually results in a model having more noise than in the low noise condition. Therefore, the true AUC per repetition of the simulation was computed. For the estimated models, as mentioned before, we used the Brier score and the MCR to evaluate the accuracy of the Exact Tree algorithm. The MCR and the Brier score was evaluated for both the test and train datasets in each repetition of the simulation. The AUC was calculated using the package pRoc in R (Robin, et al., 2011). To calculate the MCR and the Brier score the predict function of a beta-version of the Exact Tree R package was used.

Moreover, to answer the second research question, the spearman rho was evaluated for each of the discretization methods and the rounding method for every repetition. Since, the correlation coefficient was examined per predictor there would be 10 measures for each repetition. However, for the sake of clarity, the decision was made to only collect the range of the correlation coefficients of the 10 predictors from lowest to highest and the mode. The range was used in the descriptive analysis and the mode will be used for the statistical analysis as well.

Moreover, the first research question was answered with the use of ANOVA (see next section) and the resulting *p*-values and effect sizes were reported, specifically the generalized Eta-squared measure of effect size (ges). For the second research question the Kruskal Wallis test was performed.

Statistical Analysis

To determine how the systematic error and skewness in combination with the discretization methods affect accuracy, an ANOVA was applied. In the ANOVA analysis the MCR of the test sets was the dependent variable. We examined the between subject effects of the systematic noise and the skewness variables, and the within subject effect of the discretization method. Furthermore, the interaction effects between the discretization method and skewness and discretization method and systematic error were examined. The ANOVA was performed with the use of the ez package in R (Lawrence, 2016).

To determine the effect of discretization on the measurement agreement between discretized data and the original simulated data with varied levels of skewness and noise, a Kruskal Wallis Rank Sum test was performed. The test was performed on the mode values of the Spearman rho correlation coefficients. To further examine which groups differed from each other a pairwise Wilcoxon test was performed.

Computational Note

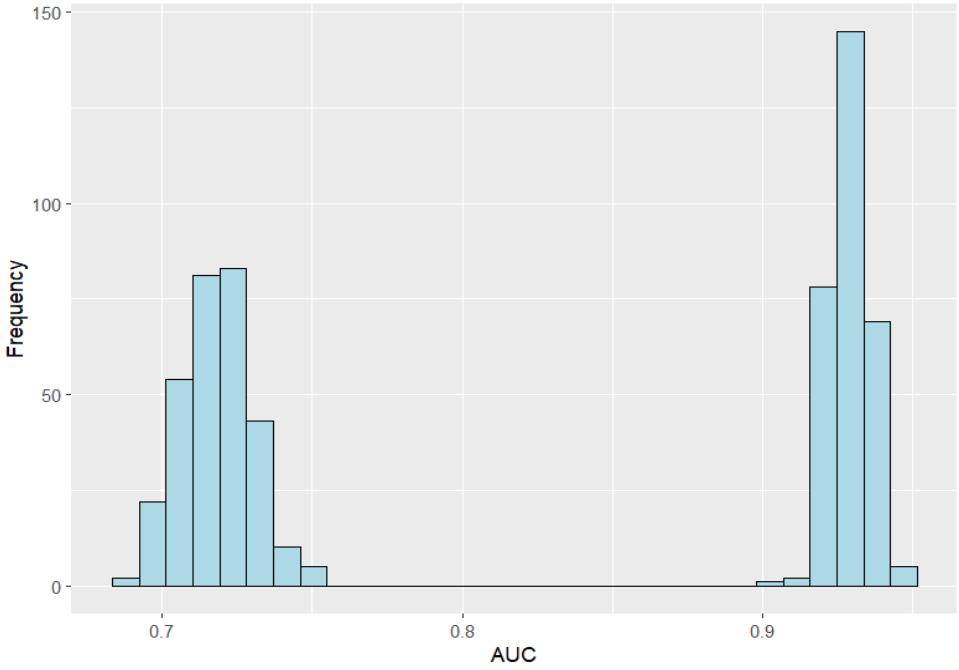
All the necessary programming and simulation was performed in the programming language R and the R-code is provided on the following GitHub page for reproducibility: https://github.com/RobertRou/MSC-Thesis.git . Furthermore, the simulation was performed on the Shark cluster computer provided by Leiden University.

5. Results

Descriptive Results

To check whether the data simulation went according to plan a manipulation check was performed. The simulation study was designed to create two distinct groups of datasets with one having high systematic error (high noise), and the other with low systematic error (low noise). The manipulation check was evaluated using the AUC. The results of the manipulation check can be seen from Figure 2 which shows a histogram of AUC scores that are divided into two groups one with higher predictive accuracies and the other with lower predictive accuracies. The AUC evaluated the predictive accuracy of either the low noise (Equation 1) or high noise (Equation 2) model on the actual simulated outcome variables. The AUC shows whether there was a fall in predictive accuracy in the high noise condition when the outcome variable was generated using the model that had more systematic error and vice versa. In the high noise condition, the model was not as accurate as in the low noise condition, because the generation of the outcome variable involved more error. The AUC calculation was performed once for the entire dataset N = 2000 and this calculation was repeated 600 times in the complete simulation using the true models given by Equation 1 and 2.

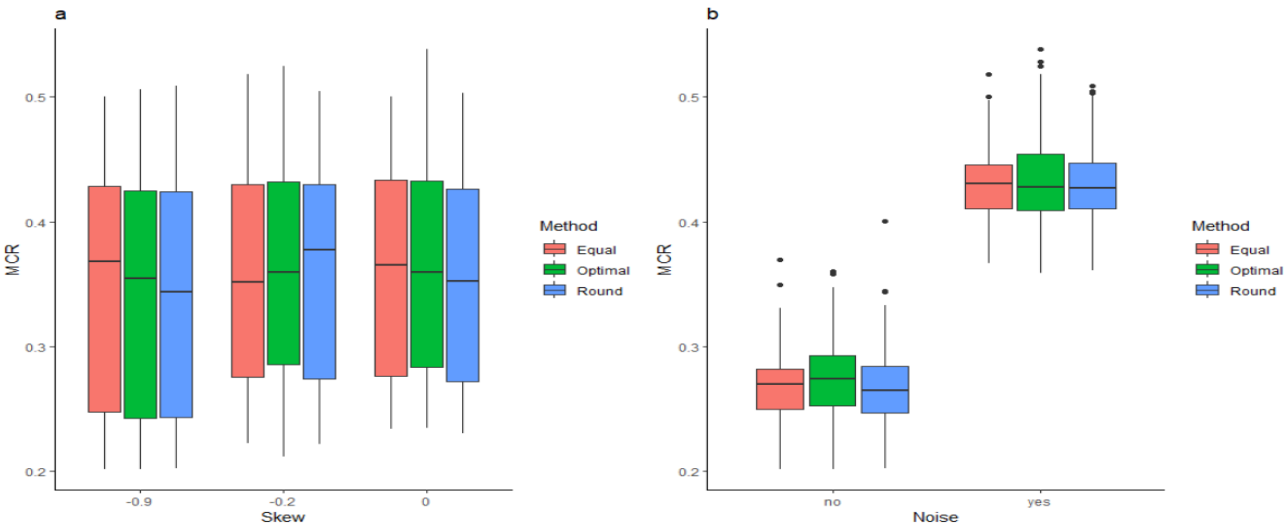
Figure 2 Histogram shows the frequency distribution of the Area Under the Curve values (N = 600).



When using the MCR as the indicator for predictive accuracy, the descriptive analysis suggested that accuracy across all discretization methods differed mostly with changes in noise. The Figure 3b shows a far higher MCR for the Exact trees that were performed in the high noise condition compared to the low noise condition. In contrast, Figure 3a shows that the influence of the changes in skewness on the accuracy of Exact Tree measured with MCR was quite small.

The descriptive analysis did not show that the three discretization methods examined had large differences in terms of the accuracy. The box plots of Figure 3b show that what differentiated the MCR scores the most was whether the prediction was performed on data which was generated using different levels of noise.

Figure 3 Box plots show effect of discretization method, skewness and noise on the misclassification rate.

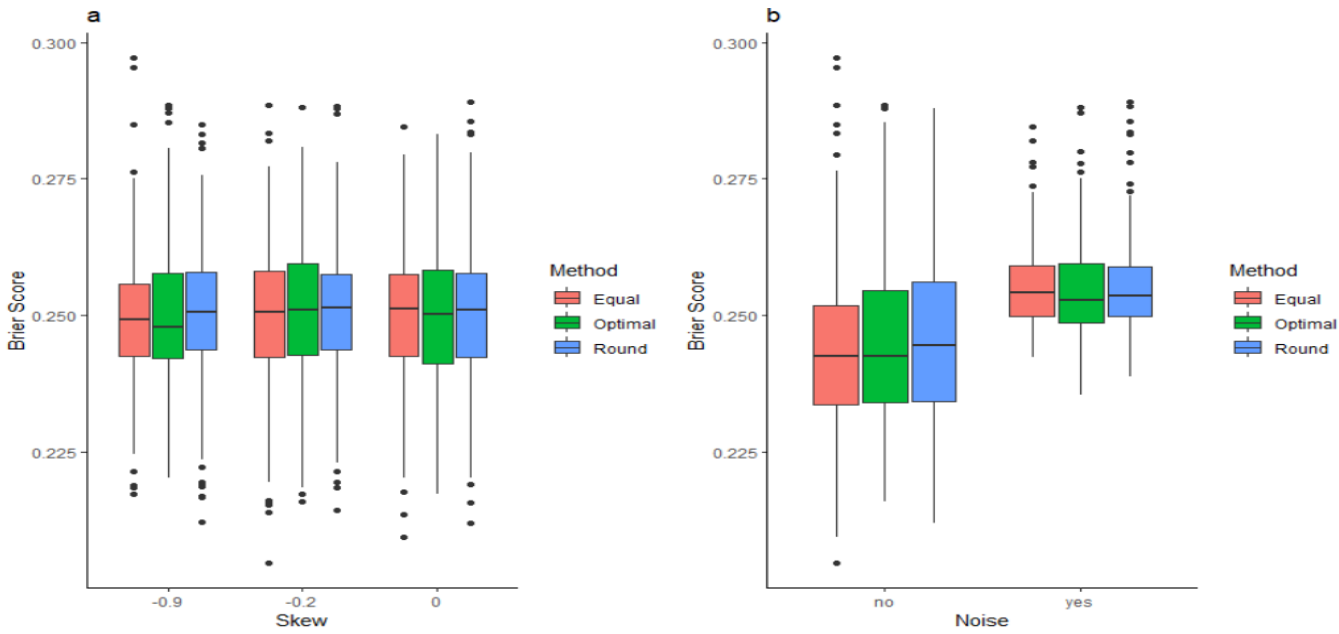


Note. Noise = yes is the high noise condition and Noise = no is the low noise condition

Since the Brier score can also be used to answer the first research question the same descriptive analysis was performed for the Brier scores. The Figure 4 show violin plots that display the relationship between the different discretization methods and simulation conditions – level of skewness and noise. Figure 4a shows that the discretization methods did not differ much in terms of Brier scores and the scores had a similar trend when differentiated by the level of skewness in the data. Figure 4b shows how the Brier scores of each discretization method are differentiated by the level of noise in the data generation; the plot shows that both the high and low noise conditions have quite different Brier scores where the

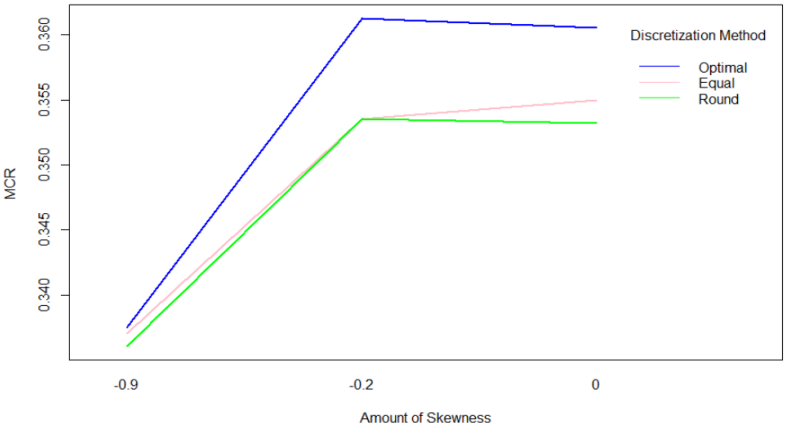
low noise condition has Brier scores that are lower, which indicates better predictive accuracy.

Figure 4 Box plots show effect of discretization method, skewness and noise on the Brier score



Note. Noise = yes is the high noise condition and Noise = no is the low noise condition

Figure 5 Diagram shows interaction effect between the discretization methods and levels of skewness in the data on the misclassification rate of the Exact Tree. It is to be noted that the range of the y-axis is very small.



To understand the effect of discretization on the measurement agreement between discretized data and the original simulated data with varied levels of skewness and noise. The range of the spearman rho correlation coefficients for the predictor variables was calculated

and the mode value was taken. The average of the lower range was 0.9989 and the average of the upper range was 0.9995. Figure 6 shows boxplots for each discretization method and the range (difference between highest and lowest spearman rho for each simulated dataset) of spearman rho coefficients. The boxplot suggests that the Optimal Binning method shows the highest amount of variation in the measurement agreement as the lowest and highest Spearman rho differed the most. Moreover, Figure 7a displays boxplots that show how the discretization methods compared on the mode spearman rho value. Here all three methods showed a good measurement agreement by looking the Y-axis, as it ranges from 0.99 to 1. Of the three methods however, the Optimal Binning method had the best performance when looking at the mode. In Figure 7b the boxplots display the mode Spearman rho Values for each level of skewness, it can be seen that both the mild skewness and no skewness conditions were higher than severe skewness in terms of the median of Spearman rho mode values.

Figure 6 Boxplot showing the computed range of Spearman rho correlation coefficients (difference between the largest and smallest predictor variable per simulated dataset) for each discretization method.

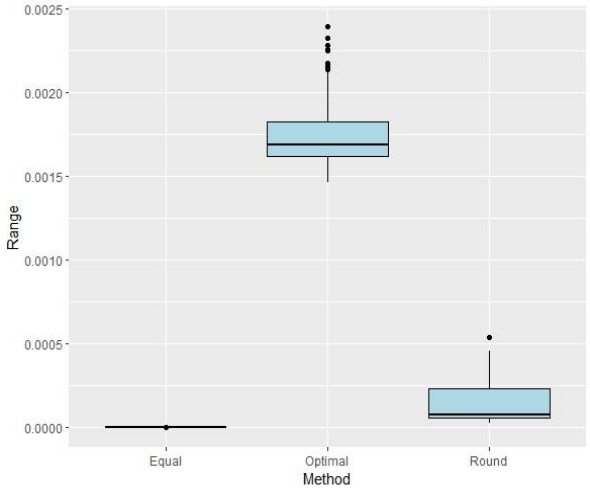
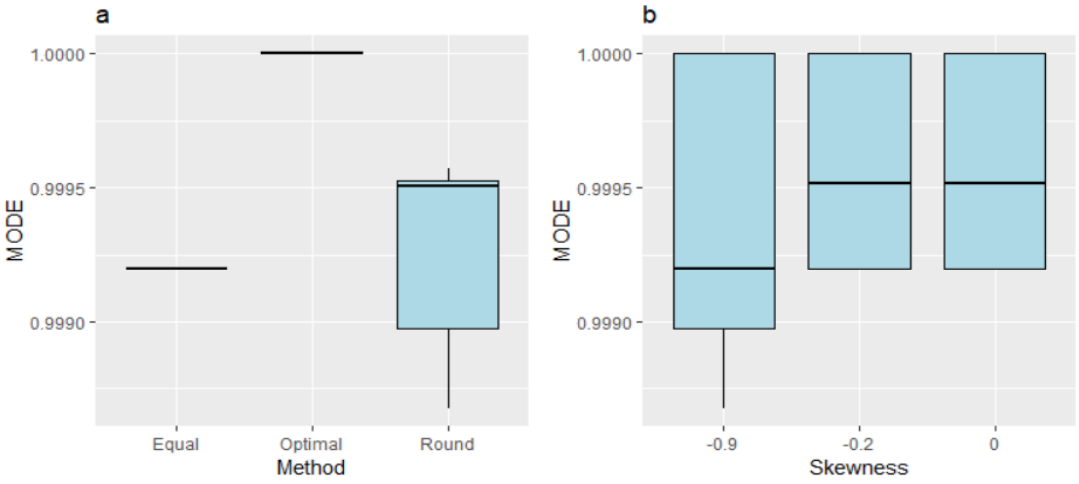


Figure 7 Box plots show the mode Spearman rho values for each method of discretization in (a) and for each level of skewness in (b). It is to be noted that the Y-axis is quite small.



Testing the Differences Between Methods

The ANOVA using the MCR resulted in significant main effect for discretization method with a very small effect size ($F(2,1188) = 14.1$, $p < .05$ and $\eta^2 = 0.008$). The interaction effect between discretization methods and skewness ($F(4, 1188) = 2.4$, $p < .05$ and $\eta^2 = 0.003$) was significant but with a very small effect size. The direction of the effect is further examined in Figure 5 with an interaction diagram. The effect of the discretization methods on the MCR depended on the level of skewness, the diagram shows that with less skewness the discretization methods resulted in higher MCR . Furthermore, the interaction between discretization methods and noise was not significant ($F(2,1188) = 0.67$, $p = .51$ and $\eta^2 = 0.0004$). Moreover, there were significant main effects for noise and with a very large effect size, ($F(1,594) = 8263$, $p < .05$ and $\eta^2 = 0.902$), and skewness with a medium effect size ($F(2,594) = 51.9$, $p < .05$ and $\eta^2 = 0.103$).

Secondly, we performed a similar ANOVA using the Brier scores. The ANOVA resulted in a main effect for discretization method that was not significant with a very small effect size ($F(2,1188) = 0.59$, $p = .55$ and $\eta^2 = 0.0004$). The interaction effects between noise and discretization method was not significant and had a very small effect size ($F(2,1188) = 1.29$, $p = .27$ and $\eta^2 = 0.001$) and between skewness and discretization method was not significant and also had a very small effect size ($F(4,1188) = 0.14$, $p = .96$ and $\eta^2 = 0.0002$). Moreover, the main effect for noise was significant with a large effect size ($F(1,594) = 205.1$, $p < .05$ and $\eta^2 = 0.15$). The main effect for skewness was not significant with a very small effect size ($F(2,594) = 0.13$ $p = .87$ and $\eta^2 = 0.0002$).

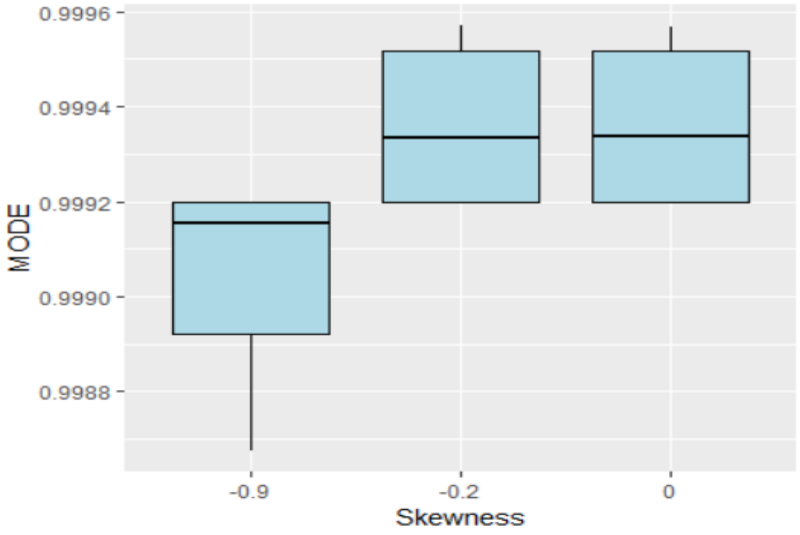
As mentioned before, to understand the effect of discretization on the measurement agreement between discretized data and the original simulated data with varied skewness and noise we computed a spearman rho correlation coefficient. To determine whether there were differences between methods of discretization in measurement agreement a Kruskal Wallis test was performed on the mode Spearman rhos. The results of the Kruskal Wallis test for the different methods showed a significant difference between the methods of discretization on the mode (Chi-squared(2) = 1343.3. $p < .05$). The results of the Kruskal Wallis test examining the differences in Spearman rhos by the three levels of skewness also showed a significant difference (Chi-squared(2) = 143.92 , $p < .05$). The Kruskal Wallis test for examining the difference in spearman rho between the two levels of noise did not show a significant difference between the groups (Chi-squared(1) = 0.019, $p = .88$).

To determine which groups in particular differed from each other we used a pairwise Wilcoxon rank sum test. The results showed the discretization methods all differed from each other significantly on the mode Spearman rho ($p < .05$). The direction of the effect was explored in Figure 7a which shows that Optimal Binning had a higher mode Spearman rho than the other methods. Examining the difference between the three levels of skewness show a significant difference between the non-skewed data and the more severely skewed data ($p < .05$) as well as between the mild skewness and the severe skewness data ($p<.05$). The difference between the non-skewed data and the mild skewness data was not significant ($p = .97$). The direction of the difference can be seen in Figure 7b which shows that in both mild skewness and no skewness the median mode Spearman rho was higher than for the more severely skewed condition.

Furthermore, due to many observations of the Spearman rho correlation coefficient were 1, so perfect correlation, the decision was made to perform the analysis on the non-perfect correlation subset of the mode values. The results were similar to the ones reported above. The Kruskal Wallis resulted in a significant difference between the discretization methods (Chi-squared(1) = 114 , $p < .05$). The Kruskal Wallis test between the three levels of skewness also resulted in a significant difference (Chi-squared (2)= 513, $p < .05$). The difference between the two levels of noise was not significant (Chi-squared(1)= 0.07, $p = 0.78$). To examine which group in particular differed from each other the pairwise Wilcoxon rank sum test was only examined on the groups of skewness. The test shows a significant difference between the non-skewed data and the severely skewed data ($p < .05$) as well as between the mild skewness and the severe skewness data ($p <.05$). The difference between the

non-skewed data and the mild skewness data was not significant (p = .95). The direction of the effect are shown in Figure 8, the direction is mostly the same as without removing the perfect correlation.

Figure 8 Box plots show the mode Spearman rho values for each level of skewness when the perfect correlation has been removed. It is to be noted that the Y-axis is quite small.



6. Discussion

The purpose of this simulation study was to gain a better understanding of the Exact Tree algorithm and how different methods of discretization and changes in data, such as adding skewness and more noise, affect its predictive accuracy. The second aim was to understand the effect of discretization on the measurement agreement between discretized data and the original simulated data with varied skewness and noise. The results suggest that the choice of discretization method does affect the accuracy of the Exact Tree. However, the main effect of type of discretization had a small effect size of Eta-squared = 0.008 (generalized eta-squared) suggesting that the choice of processing the data with Equal Frequency discretization, Optimal Binning or rounding does not have a large influence on the resulting miss classification rates. However, the significant interaction effect between the discretization method and the level of skewness when MCR was used suggests that when making the decision to discretize, skewness should be taken into account (but the effect size was small, Eta-squared= 0.003). In a closer look, the Optimal Binning method had a higher MCR in the mild skewness and non-skewed conditions. This would suggest that when there is less skewness in the data, rounding and Equal Frequency discretization should be preferred to

Optimal Binning. In this study the MCR was lowest at the more severe skewness condition for all three methods, which differed from previous literature where skewness and kurtosis had a positive relationship with classification error (Ismail & Ciesielski, 2003). However, we only manipulated skewness and kept kurtosis at a constant, which could explain the difference in results. Moreover, Equal Frequency discretization was found to be less vulnerable to changes in the distribution of the data, which could also explain why the error rate decreased when it was used in the severe skewness condition (Ismail & Ciesielski, 2003). Moreover, when the same analysis was performed using the Brier scores as a measure of predictive accuracy the choice of discretization method did not have a significant effect on the predictive accuracy.

The second research question was to determine the effect of discretization on the measurement agreement between discretized data and the original simulated data with varied skewness and noise. The results show that all three discretization methods resulted in discretized data which had high measurement agreement with the original non-discretized variables. The measurement agreement was measured using the Spearman rho correlation coefficient between each of the ten discretized predictor variables and original non-discretized variables. The result was summarized as the mode Spearman rho between the ten discretized and the ten non-discretized predictor variables for each of the discretization methods for all 600 simulated datasets. The results suggest that the methods all differed from each other significantly, with the Optimal Binning method having the highest mode of Spearman rho suggesting that on average it resulted in a higher correlation than the other methods. Furthermore, measurement agreement differed significantly between the three levels of skewness, but not between the two levels of systematic error. The results also show that measurement agreement varied a lot more in the rounding method which is not surprising since it is not an actual discretization method and does not reduce the data to the same extent as the two other methods compared. When looking closer at how the three levels of skewness differed from each other, the results showed that the non-skewed and mildly skewed levels did not differ significantly in measurement agreement, but both differed from the severe skewness level.

As a sensitivity analysis, the ANOVA was performed again the mode values which showed perfect correlation because generally having perfect correlation does not add new information. The results without the perfect correlation values also show that there was a significant difference between the discretization methods. Furthermore, the three levels of

skewness differed significantly on the mode values but the two levels of noise did not, which was the same results as in the analysis that included the perfect correlation values.

Certain limitations of this study are that only three methods were compared and a larger amount of discretization methods would provide more information, for example in the research by Ismael and Ciesielski (2003) six discretization methods were compared (Ismail & Ciesielski, 2003). Moreover, the trend seen in the interaction between skewness and discretization method suggested that in the most skewed condition the MCR was lowest for each of the methods, and there was no difference between the methods. In psychological research -0.9 skewness (most skewed condition of this study) is not most severe and higher levels are seen, such as one study found that overall at the 95th percentile of skewness levels in psychological research variables had skewness of 2.77 (Cain, Zhang & Yuan, 2017). Furthermore, in this study the choice was made to discretize the simulated data in to 25 categories, the choice was based on research that used Exact Tree and found using 25 categories to generally have to best accuracy (Meulman, Dusseldorp & Van Os, 2011). As currently the decision to how many categories the data was discretized was quite arbitrary, we suggest future research towards finding a method by which to decide the number of categories data is discretized to. Finally, it should be noted, that we worked with a beta version of the Exact Tree R-package. After the whole simulation study was performed, we found out that the estimated predicted probability (by the predict function of Exact Tree) was only correct for regression problems not for classification problems (such as the ones in this study). Therefore, the results for the Brier score should be interpreted with caution.

In regards to future research we propose incorporating more discretization methods and comparing their predictive accuracy with the Exact Tree algorithm. Furthermore, we suggest that future simulations include more extreme levels of skewness such as 2.77, which are still realistic in psychological research.

In conclusion, with regard to predictive accuracy the Optimal Binning performed worse than the equal frequency and rounding methods, except for situations with severe skewness in which there was no difference in performance between the methods. And with regard to measurement agreement, the Optimal Binning performed a bit better than the other two. The results suggest that different discretization methods produce data that can vary in how well they correlate with the original data. In this case the discretization methods (Equal Frequency and Optimal Binning) produce less variation in the measurement agreement than rounding. However, in this study all three methods still returned high measurement agreement

between the non-discretized and discretized data. Overall, the results of this study suggest that there were only small differences between the methods, therefore the choice of discretization method is not that important to the performance of the Exact Tree.

References

Anderson, E. (1936). The Species Problem in Iris. *Annals of the Missouri Botanical Garden*, *23*(3), 457–509. https://doi.org/10.2307/2394164

Becker, M., Klößner, S. (2022). PearsonDS: Pearson Distribution System. R package version 1.2.2.

Brier, G. W. (1950). Verification of Forecasts Expressed in Terms of Probability, *Monthly Weather Review*, *78*(1), 1-3. https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Wadsworth.

Cain, M. K., Zhang, Z., & Yuan, K. H. (2017). Univariate and Multivariate Skewness and Kurtosis for Measuring Nonnormality: Prevalence, Influence and Estimation. *Behavior Research Methods*, *49*(5), 1716–1735. https://doi.org/10.3758/s13428-016-0814-1

Casas, P. (2020). funModeling: Exploratory Data Analysis and Data Preparation Tool-Box. R package version 1.9.4.

Csorba, I. (2020). *Evaluating methods for the comparison of medians: Which test is the best?*(Masters Thesis). Available from https://hdl.handle.net/1887/3181674

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, *7*(2), 179–188. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x

Liu, H., Hussain, F., Tan, C. L., & Dash, M. (2002). Discretization: An Enabling Technique. *Data Mining and Knowledge Discovery*, *6*(4), 393. https://doi.org/10.1023/A:1016304305535

Goldstein-Greenwood, J. (2022, April 15). *University of Virginia Library Research Data Services + Sciences*. *Research Data Services + Sciences*. Retrieved November 8, 2022, from https://data.library.virginia.edu/roc-curves-and-auc-for-models-used-for-binary-classification/

Hacibeyoglu, M., & Ibrahim, M. H. (2018). EF_Unique: An Improved Version of Unsupervised Equal Frequency Discretization Method. *Arabian Journal for Science and Engineering*, *43*(12), 7695–7704. https://doi.org/10.1007/s13369-018-3144-z

Harrison, D. A., Brady, A. R., Parry, G. J., Carpenter, J. R., & Rowan, K. (2006). Recalibration of risk prediction models in a large multicenter cohort of admissions to adult, general critical care units in the United Kingdom. *Critical Care Medicine*, *34*(5), 1378–1388. https://doi.org/10.1097/01.CCM.0000216702.94014.75

Ismail, K.M., Ciesielski, V. (2003). An empirical investigation of the impact of discretization on common data distributions. *Design and application of hybrid intelligent systems*. IOS Press, NLD, 692–701.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 103). New York: Springer. https://doi.org/10.1007/978-1-4614-7138-7

Kliegr, T. (2017). Quantitative CBA: Small and Comprehensible Association Rule Classification Models. *ArXiv*, https://doi.org/10.48550/arXiv.1711.10166

Lawrence, M. (2016). ez: Easy Analysis and Visualization of Factorial Experiments. R package version 4.4-0.

Ma, X. (2018). *Using classification and regression trees: A practical primer*. Information Age Publishing, Inc.

Meulman, J.J., Dusseldorp, E., Van Os, B.J.: An exact dynamic programming algorithm for regression trees. In: Van der Heijden, M., Koren, B., Van der Mei, R.D., Van Vonderen, J.A.J.(eds.) Jan Karel Lenstra, *the Traveling Science Man*: Liber Amicorum, pp. 198–208. CWI, Amsterdam (2011).

Mukaka, M. M. (2012). Statistics corner: A guide to appropriate use of correlation coefficient in medical research. *Malawi Medical Journal*, *24*(3), 69–71.

Ramírez-Gallego, S., García, S., Mouriño-Talín, H., Martínez-Rego, D., Bolón-Canedo, V., Alonso-Betanzos, A., Benítez, J. M., & Herrera, F. (2015). Data discretization: Taxonomy and big data challenge. *WIREs Data Mining and Knowledge Discovery*, *6*(1), 5–21. https://doi.org/10.1002/widm.1173

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.C., & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*(1), 77–77. https://doi.org/10.1186/1471-2105-12-77

Rufibach, K. (2010). Use of Brier score to assess binary predictions. *Journal of Clinical Epidemiology*, *63*(8), 938–939. https://doi.org/10.1016/j.jclinepi.2009.11.009

Sahni, O., Müller, J., Jansen, K. E., Shephard, M. S., & Taylor, C. A. (2006). Efficient anisotropic adaptive discretization of the cardiovascular system. *Computer Methods in Applied Mechanics and Engineering*, *195*(41), 5634–5655. https://doi.org/10.1016/j.cma.2005.10.018

Senavirathne, N., & Torra, V. (2019). Rounding based continuous data discretization for statistical disclosure control. *Journal of Ambient Intelligence and Humanized Computing*, *1*. https://doi.org/10.1007/s12652-019-01489-7

Strobl, C., Boulesteix, A.-L., & Augustin, T. (2007). Unbiased split selection for classification trees based on the Gini Index. *Computational Statistics & Data Analysis*, *52*(1), 483–501. https://doi.org/10.1016/j.csda.2006.12.030

Van Os, B.J. (2000). *Dynamic Programming in Multivariate Data Analysis*. Leiden University.

Wang, X., Shang, P., Huang, J., & Feng, G. (2013). Data discretization for the transfer entropy in financial market. *Fluctuation and Noise Letters*, *12*(4). https://doi.org/10.1142/S0219477513500193

Appendix

Table 2 Mean misclassification rate of each combination of the simulation factors on test data.

	Low Noise			High Noise		
Skewness	Equal	Optimal	Round	Equal	Optimal	Round
-0.2	0.274	0.286	0.275	0.432	0.436	0.431
-0.9	0.248	0.246	0.244	0.425	0.429	0.428
0	0.277	0.285	0.278	0.433	0.436	0.429