ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2025; 16(1); 30-47

Performance of Classification Techniques on Smaller Group Prediction

Cahit POLAT * Kathy GREEN **

Abstract

Classification techniques allow researchers to analyze data based on groups for the purposes of clustering or making predictions about group membership. Since there are many methods for utilizing classification analyses, such as Linear Discriminant Analysis (LDA), Logistic Regression (LR), and Classification and Regression Trees (CART), it is important to know which techniques perform better under which conditions to affect prediction accuracy. In the context of group prediction, it is crucial to consider the impact of group proportional sizes on prediction accuracy, particularly when comparing smaller groups to larger ones. This study evaluated the small group prediction accuracies of LDA, LR, and CART, controlling for the number of groups, correlation, and the number of predictor variables. Results showed that CART performed best for smaller and overall group prediction in most cases. In addition, a notable difference was observed in overall group prediction accuracy compared to small group prediction accuracy, with the overall group prediction accuracy being greater. Data conditions had a greater impact on LR and LDA than CART, and, in certain instances, LR showed superiority over the other two methods. The number of groups was the most influential factor on small group prediction, while the number of predictor variables, correlation, and method were of decreasing influence. In general, overall group prediction accuracy and small group prediction accuracy were negatively related. However, for the categories with an equal number of groups, the two were positively related.

Keywords: method performance evaluation, group membership, classification accuracy, simulation.

Introduction

Classifying cases into groups is widespread in all fields, and statistical or analytical techniques may perform differently depending on the data conditions. The data structure influences the choice of methods of analysis and sets constraints on the study's scope. Classification serves the purpose of identifying group characteristics and predicting group membership and is a valuable statistical approach in various fields such as social sciences, education, health sciences, and other domains. It is further crucial for researchers to assess the significance of predictors in determining the group or class to which observations belong.

Explanatory models are applied to examine relationships between variables, whereas predictive models are utilized to make predictions about categories using a correlational design. Group discrimination and decisions are assessed using these models (Sainani, 2014). Utilizing predictive models, for instance, one may determine the likelihood of contracting an illness based on the findings of diagnostic tests or the mortality rate of a veteran suffering a stroke within a year at a certain severity level (Bates et al., 2014). By applying such models, it is possible to determine, for example, whether certain predictor variables like the student's positive opinion of their teacher, GPA, whether they lived with their biological parents, and the number of days the student missed from school also predict the dropout status of high school students (Suh et al., 2007).

* Asst. Prof. Dr., Harran University, Faculty of Education, Şanlıurfa-Türkiye, cahitpolat@harran.edu.tr, ORCID ID: 0000-0002-1423-5084
** Professor, University of Denver, Faculty of Education, Colorado, US, Kathy.Green@du.edu, ORCID ID: 0000-0002-1676-3139

To cite this article:
Polat, C. & Green, K. (2025). Performance of Classification Techniques on Smaller Group Prediction, *Journal of Measurement and Evaluation in Education and Psychology, 16*(1), 30-47. https://doi.org/10.21031/epod.1598907
Received: 9.12.2024
Accepted: 19.03.2025

There are various techniques for determining group membership, and logistic regression (LR), linear discriminant analysis (LDA), and classification and regression trees (CART), a more recent technique, are widely used ones (Agresti, 2002; Huberty & Olejnik, 2006; Williams et al., 1999). LDA and LR have historically been used extensively in educational and social science research, but CART is a newer technique (Holden et al., 2011). Additionally, in many recent studies, these techniques are applied simultaneously (Castonguay et al., 2022; Hassan et al., 2024; Hoang et al., 2025; Saboor et al., 2022; Selim et al., 2020; Song et al., 2022; Zampogna et al., 2024).

Though they are widely used, limited information exists regarding the effectiveness of these three techniques in predicting categories of observations, especially for relatively smaller groups, and which perform better in certain data scenarios, such as group size ratios, degree of correlation, number of predictor variables, and number of groups in the outcome variable. Therefore, this study aimed to investigate the performance of LDA, LR, and CART for overall and smaller group prediction in addition to whether prediction accuracies are affected by the correlation between predictor variable strength, number of predictor variables, group size ratios, and number of groups in the dependent variable. Finally, this study explored the relationship between overall group prediction accuracy and small group accuracy. We provide a brief overview of each technique below.

Linear Discriminant Analysis (LDA)

LDA procedure calculates the observation score for jth group $\left(G_j\right)$ as;

$$G_j = c_{j0} + \sum c_{ji}x_i + \ln\left(\frac{n_j}{N}\right) \qquad (1)$$

where c_{j0} represents the constant value for the jth group, c_{ji} denotes the coefficient value of the ith variable within the jth group, x_i is the ith variable, n_j indicates the total number of observations in the jth group, and N represents the total number of all observations.

Moreover, the constant value for the jth group c_{j0} and the coefficient values c_{ji}s are calculated by the formula;

$$c_{j0} = \frac{1}{2}C_j' M_j \qquad (2)$$

where $C_j = W^{-1}M_j$, C_j is the coefficients vector for c_{ji}s, W is the pooled within-group variance-covariance matrix, and M_j is matrix of the means of the variables for group j.

Upon computing the observation scores for each group, the observation is allocated to the group with the highest score. LDA models are exclusively linear functions and assume the absence of multicollinearity and singularity, as well as homogeneity of variance-covariance matrices and multivariate normality (Tabachnick & Fidell, 2013).

Logistic Regression (LR)

LR starts with calculating linear regression model *u* as;

$$u = B_0 + \sum B_j X_{ij} \qquad (3)$$

where B_0 represents the linear regression model's intercept., and B_j indicates the jth variable's coefficient, X_j.

Then $\widehat{Y_i} = \frac{e^u}{1+e^u}$ is calculated as the probability that the ith observation is a member of a group rather than a reference group. It can be seen easily that the natural log of the probability of the odds ratio being in one group versus another reference group is equal to *u* such as;

$$\ln\left(\frac{\hat{Y}}{1-\hat{Y}}\right) = B_0 + \sum B_j X_{ij}. \qquad (4)$$

In many statistical applications, the default threshold for determining observation membership is set at 0.5; hence, if the logit equals or exceeds 0.5, the observation is classified within the group. The cut point may also be established at another value (Soureshani et al., 2013). Logistic Regression (LR) distinguishes itself from many techniques by its flexibility, since it does not relay on certain assumptions such as normality.

Classification and Regression Trees (CART)

CART divides data iteratively to classify objects into more homogeneous groups, which are referred to as nodes. The CART algorithm initiates by locating all subjects in a single node. Subsequently, it assigns them to other nodes by utilizing predictor variables to establish the most homogeneous groups (Breiman et al., 1984). This procedure continues until an ideal group split achieves the desired degree of group membership homogeneity. To mathematically apply this, the node deviances are minimized, and the deviance for i^{th} node (D_i) is computed as;

$$D_i = -2\sum\sum n_{ik} ln(p_{ik}) \tag{5}$$

where n_{ik} denotes the number of subjects from group k in node i, and p_{ik} indicates the proportion of subjects from group k within node i.

The sum $D = \sum D_i$, is used as a measure of homogeneity once the deviances of each group have been calculated; smaller Ds signify higher homogeneity. The procedure continues until either the requirement for stopping iterations is met or the reduction in Ds from one step to the next becomes trivial.

Related Research

This section includes a summary of the literature review of related studies. In a comparison of the overall performance of LDA and LR, one study found that LR had a higher prediction accuracy for group membership (Barön, 1991), while others found little or no difference between the two methods (Dey & Astin, 1993; Hess et al., 2001; Meshbane & Morris, 1996). Further, the statistical methods LDA and LR exhibited comparable performance to CART (Dudoit et al., 2002; Ripley, 1994). However, other studies have demonstrated that LDA and LR outperform CART (Preatoni et al., 2005; Williams, 1999) or that CART outperforms LR and LDA (Holden, 2011; Hao et al., 2022). Lastly, while some results indicated that CART performed better than LDA in terms of group membership prediction accuracy (Grassi et al., 2001), others indicated that LR and CART performed similarly (Schumacher et al., 1996). These conflicting findings may be due to different configurations of the data analyzed. In this regard, the overall performance of any method is uncertain in the absence of an assessment of the data's specific characteristics.

While certain studies compared the accuracies of the methods, the comparison results were not generalizable beyond the scope of the research. Hence, some researchers utilized simulated data to compare the performance of techniques rather than utilizing real data from content areas. A substantial advantage of simulated data is the researcher's capacity to manage the data conditions. As a result, numerous studies have compared the performances of methods under controlled conditions. Numerous data factors may have an impact on how well classification techniques perform. For classification accuracy the following conditions have been shown to have an effect: sample size (Bolin & Finch, 2014), group size ratios (Finch & Schneider, 2006; Lei & Koehly, 2003), effect size (Holden et al., 2011), predictor distributions (Pai et al., 2012; Pohar et al., 2004), and homogeneity of variance-covariance matrices (Fan & Wang, 1999; Lei & Koehly, 2003). On the other hand, less researched but important for comparing the methods are correlations between predictor variables (Kiang, 2003), number of variables (Holden & Kelley, 2010), number of groups in the dependent variable (Zavroka & Perret, 2014), model complexity (Holden et al., 2011), dynamic structure of the data, linearity, presence of outliers (Pai et al., 2012), multimodal structure of the data (Kiang, 2003), percent of initial misclassification (Bolin & Finch, 2014), and group separation (Finch et al., 2014).

CART outperforms LDA and LR in various scenarios involving sample size, homogeneity of variance-covariance matrices and effect size, group size ratio, varying model complexities, percentage of initial misclassification, and group separation level (Bolin & Finch, 2014; Finch et al., 2014; Holden et al., 2011); however, it performs less effectively in scenarios involving normal or skewed data (Finch & Schneider, 2006). When the normality and homogeneity of variance-covariance matrices are violated, LR is predicted to outperform LDA (Dattalo, 1994; Ferrer & Wang, 1999; Huberty, 1999). Meanwhile, despite the broad acceptance of the normality assumption for LDA, it may still be resistant to nonnormality (Graf et al., 2023). Under most circumstances, LR and LDA exhibited generally comparable performance, despite some conflicting results (Dey & Austin, 1993; Hess et al., 2011). Kiang (2003) found that when multimodal data and nonlinearity are present, LR performs better than LDA. The dynamic nature of the data and the presence of outliers impact the classification techniques' success (Pai et al., 2012).

The number of groups in the dependent variable (Pohar et al., 2004) and the number of predictor variables (Huberty, 1994; Rausch & Kelley, 2009) had an impact on classification technique performance. The change in the performance of the techniques LDA, LR, and CART were similar when additional groups were included, and the methods' classification accuracies rose as there were more predictor variables. LDA was shown to perform less well under multicollinearity, whereas LR was unaffected by multicollinearity (Pai et al., 2012). Finally, the group size ratio plays an important role in the performance of methods for small and overall group prediction. When proportions are highly unbalanced, small group prediction accuracy tends to be lower while overall group prediction accuracy tends to be larger (Finch & Schneider, 2006). However, the number of studies testing LDA, LR, and CART simultaneously for the effect of data conditions on small prediction accuracy is limited.

Importance of the Study

Although prior research has provided some insight into the parameters influencing the performance of LDA, LR, and CART, further research is necessary to gain a deeper comprehension of the group classification techniques' respective performances. In particular, the number of predictor variables, the number of groups in the dependent variables, and the correlations between predictor variables have not been fully examined. To get more thorough findings, group size ratio should be taken into consideration while evaluating these circumstances. Additionally, classification accuracies of smaller groups should be considered in addition to overall classification accuracy. In cases where data are unbalanced, the prediction of the smallest group may be important. In consideration of this, this study concentrated on the precision of the small group prediction in situations where the sample sizes of the groups were unbalanced. Besides, this study aimed to investigate which of the three methods performs better in terms of smallest group prediction accuracy given varying degrees of correlation between predictor variables, number of groups in the dependent variables, and number of predictor variables. The purpose was to determine whether the number of groups, the level of correlation between predictor variables, the number of predictor variables, and the group size ratios in the dependent variables interact significantly in relation to the classification accuracy of the overall and the smallest group of the three methods. Finally, this study also aimed to investigate the relationship between the accuracy of prediction for small groups and whole groups. Consequently, the research questions for this study are as follows:

1. How do the number of predictor variables, the number of groups, and the correlation between predictor variables affect prediction accuracy for smaller groups?

2. What is the relationship between overall group prediction accuracy and small group prediction accuracy in different data scenarios?

Method

Research Design

Factors associated with data characteristics were controlled in this study. The variables were group size ratio (2 levels: balanced, imbalanced), number of groups (3 levels: 2, 3, 4), correlation between predictor variables (2 levels:.2,.5), and number of predictor variables (3 levels: 2, 5, 10). While the last two conditions are related to the dependent variable, the first two conditions are related to predictor variables. In addition, three distinct analysis techniques (LDA, LR, and CART) were applied. As a result, using each of the three methods, 2x3x2x3 = 36 distinct data conditions were generated and examined. It was considered that all other variables are uncontrollable and random. A fixed sample size of 200 was used, and 1000 simulations were run for each condition. Consequently, the study contained 36x200 = 7,200 simulated observations, each with 1,000 repetitions for each method. For the smallest group prediction, only unbalanced data in terms of group size ratio was applied, while for the overall group prediction, both balanced and unbalanced data were applied. All the predictor variables were simulated as normally distributed with a mean of 0.0 and a standard deviation of 1.0, i.e., a standard normal distribution.

Steps of Data Generation

A Monte Carlo simulation procedure was utilized to produce a dataset with the specified conditions. Monte Carlo techniques apply random sampling to simulate data as it permits the generation of random variables and the management of controlled variables. These techniques involve generating datasets that meet specific criteria using mathematical approximations and probability computations (Paxton et al., 2001).

The function MVRNORM in R software (R Core Team, 2016) was utilized to create data with specific characteristics, ensuring that the predictor variables followed a multivariate normal distribution. Researchers can use the MVNORM package in R to define the correlations among predictor variables and the number of predictor variables. The sample size was set at 200, which is commonly used in simulation studies and a suitable number of observations in quantitative research in the social and educational sciences. Additionally, for LDA, prior probabilities were determined based on the observed group ratios of the respective sample sizes to the total sample size, following the suggestion of Lei and Koehly (2003).

The MVRNORM function generates multivariate normal distribution variables for each group. For example, generating all five predictor variables by MVRNORM yields multivariate normal distributions for each group, but that does not guarantee normality when combining each group for the dependent variable. This function also lets one define predictor variable means and standard deviations for each dependent variable group. However, multivariate normality is not guaranteed for each iteration when creating predictor variables from a multivariate normal distribution for each group and merging them for total datasets.

The groups were designated as group 1, group 2, group 3, and group 4. Groups with lower numerical labels include fewer observations. In unbalanced scenarios, group 1 consistently has the smallest group size. The simulation of a 1000-iteration dataset under appropriate modified and random settings was completed by following the steps outlined below and using the necessary R tools. If data non-convergence occurred during one replication, an additional replication was performed using the R software to compensate, resulting in the completion of 1000 replications. Following the completion of data training, the data were prepared for analysis.

Controlled Variables and Their Patterns

Two degrees of correlation (CORR) were established: 0.2 (indicating low) and 0.5 (indicating medium). Specific values for low correlation (0.2) and medium correlation (0.5) among all predictor variables were entered using the MVRNORM function in R. Adjustments were made to all five predictor variables to achieve a correlation of 0.2 if the correlation coefficient was 0.2. In the same way, in the case where the correlation coefficient was 0.5, the five variables were adjusted to exhibit a correlation of 0.5. However, when predictor variables were simulated, the average correlation was greater in magnitude when compared to the fixed level. Depending on the data context, when correlations were set to 0.5 in MVRNORM, the simulated data correlations were, for instance, 0.58 or a slightly different value. This was due to arrangements regarding group size ratios and effect sizes. To attain the predetermined correlation conditions, lover-level correlations were introduced to the program and the correlation coefficients were progressively decreased in R throughout the data simulation process until the desired average coefficient values of 0.2 and 0.5 were reached for each of the 36 data scenarios.

The levels of the number of predictor variables (NPV) used in the study were based on generated data with two, five, and ten predictor variables. These levels were set automatically by creating correlation matrices. This study splits the number of groups (GN) in the dependent variable into three levels: two, three, and four, which are the most widely used. To create groups, group size ratios were utilized to count and calculate the number of observations for each group. For example, for three groups with a 10:20:70 group size ratio, 20, 40, and 140 observations were simulated for each group because the total sample size was 200 summed across the group size. Different numbers were assigned to categories. For instance, with three groups in the dependent variable, group 1 had 20 cases, group 2 had 40 cases, and group 3 had 140. After simulating and labeling dependent variable groups (from smaller to larger sizes: group 1, group 2, group 3, and group 4) and predictor variable datasets for each iteration, the outcome variable and predictor variables were randomly matched.

Two different levels of group size ratio (GSR) were controlled in this study: balanced group size ratios and unbalanced group size ratios. A balanced group size ratio exists when the dependent variable's groups have the same number of observations. On the other side, an unbalanced group size ratio exists when the number of instances in the groups is unequal and there is a significant discrepancy in the number of observations between the largest and smallest groups. The group size ratios for balanced groups were set to 50:50, 33:33:33, and 25:25:25:25, respectively, when there were two, three, and four groups. As a result, each group had the same number of instances, with 100 cases per group when there were two groups, 67 cases (1 case omitted from the middle group to set the sample size to 200) when there were three groups, and 50 cases per group when there were four groups. Unbalanced group ratios, on the other hand, were set at 10:90, 10:20:70, and 10:15:20:55 for groups of two, three, and four, respectively. Thus, group sizes were 20 and 180 for the case of two groups, 20, 40, and 140 for the case of three groups, and 20, 30, 40, and 110 for the case of four groups.

Simulating Groups of Dependent Variables

To simulate values for groups for dependent variables, the software was programmed to include the means of predictor variables for each group. The effect size, defined as the standardized difference between consecutive groups, was set at 0.5 using the classification of 0.2, 0.5, and 0.8 as small, medium, and large effect sizes, respectively (Cohen, 1988). The overall group mean was set to zero; to meet this criterion, group means were calculated using their group size ratios. The group means μ_1, μ_2, μ_3 and μ_4 for groups 1, 2, 3, and 4 were calculated using the equations explained below.

Group means were determined based on effect sizes so that consecutive groups' mean difference was 0.5 and the overall mean was 0. Therefore, for balanced two-group case equations $\mu_2 - \mu_1 = 0.5$ and $\mu_1 + \mu_2 = 0$ were solved and, $\mu_1 = -0.25$ and $\mu_2 = 0.25$ were found. For the imbalanced two-group case, equations $\mu_2 - \mu_1 = 0.5$ and, $\mu_1 + 9\mu_2 = 0$ were solved and $\mu_1 = -0.45$ and $\mu_2 = 0.05$ were found. For the three-group balanced case, equations $\mu_2 - \mu_1 = 0.5$, $\mu_3 - \mu_2 = 0.5$, $\mu_1 + \mu_2 + \mu_3 = 0$ were solved and, $\mu_1 = -0.5$, $\mu_2 = 0$ and $\mu_3 = 0.5$ were found. For the three-group imbalanced case, equations $\mu_2 - \mu_1 = 0.5$, $\mu_3 - \mu_2 = 0.5$, $\mu_1 + 2\mu_2 + 7\mu_3 = 0$ were solved and, $\mu_1 = -0.8$, $\mu_2 = -0.3$ and $\mu_3 = 0.20$ were found. For four-group balanced case, equations $\mu_2 - \mu_1 = 0.5$, $\mu_3 - \mu_2 = 0.5$, $\mu_4 - \mu_3 = 0.5$, $\mu_1 + \mu_2 + \mu_3 + \mu_4 = 0$ were solved and, $\mu_1 = -0.75$, $\mu_2 = -0.25$, $\mu_3 = 0.25$ and $\mu_4 = 0.75$ were found. Finally, for four-group imbalanced case, equations $\mu_2 - \mu_1 = 0.5$, $\mu_3 - \mu_2 = 0.5$, $\mu_4 - \mu_3 = 0.5$, $2\mu_1 + 3\mu_2 + 4\mu_3 + 11\mu_4 = 0$ were solved and, $\mu_1 = -1.1$, $\mu_2 = -0.6$, $\mu_3 = -0.1$

and $\mu_4 = 0.4$ were found. Therefore, means of groups were calculated and then introduced to the program.

The observations were generated using R's *c(rep())* function after the predictor variables were assigned their determined values based on correlations between predictor variables, group size ratios, and group sizes. Then all the observations were combined by the *data.frame (,)* function with all predictor variables and the dependent variable.

Analysis of Data

After generating the data with the defined parameters, each analysis method was applied with identical datasets with the same data conditions to predict the outcome variables separately. Therefore, the LDA, LD, and CART analyses were conducted using R's *lda*, *multinom*, and *rpart* functions. Then, an algorithm was created to assess the accuracy of the class predictions obtained from three different methods and to count the number of correct predictions.

To evaluate the performance of the methods, two outcome measures were employed: rate of correct classification for all groups (rccA) and rate of correct classification for the smallest group (rccS) in terms of the group's sample size and number of correct group predictions. The calculation of rccA involved dividing the frequency of all correctly predicted observations by the total number of observations (200). Moreover, rccS was calculated by dividing the frequency of correctly predicted observations for the smallest group by the total number of observations in the smallest group. Hence, this study's analyses were based on proportions, following Edwards' (1985) approach, which used the arcsine transformed value of the proportions as a dependent variable, and the results were the same for the proportions and transformed values.

Calculating the correct prediction rates for all and small groups for each iteration, a second set of data for comparing techniques and data conditions was prepared. A five-way (3x2x3x3x2) factorial analysis of variance (ANOVA) focusing on rccA in connection to Method, Corr, NPV, GN and GSR, and a four-way factorial (3x2x3x3) ANOVA focusing on rccS in connection to Method, Corr, NPV, GN were conducted to evaluate the results of the simulation study. The factorial ANOVA and follow-ups were conducted using SPSS statistical software (IBM Corp., 2025).

Because the statistical significance of interactions and main effects is impacted by sample size, and the sample size of 1000 (number of iterations for each combination of the conditions) is quite large, therefore partial eta squared (η_p^2) was used rather than statistical significance to identify interpretable effects. Partial eta squared is a measure that determines the proportion of total sample variation explained by a specified effect while excluding other main and interaction effects (Pierce et al., 2014; Richardson, 2011). It is calculated using the formula: $\eta_p^2 = \frac{SS_{Effect}}{SS_{Total}+SS_{Error}}$ where SS_{Effect} is sum of squares for the particular effect SS_{total} represents the total sum of squares and, SS_{error} indicates the error sum of squares. Partial eta squares values were utilized to evaluate and compare the importance of main effects and interactions.

The assumptions of factorial ANOVA are independence of data, homogeneity of variance (HOV), and normality of predictor variables. The study's design fulfilled the expectations regarding the independence of observations. On the other hand, according to Levene's test, the assumption of homogeneity of variance was not fulfilled because of large sample sizes (number of iterations), variations in group size ratios, group numbers, and distinctive means. Nearly all the cells met the criteria for normality, except for a few unbalanced situations that included two or five predictor variables and a binary outcome variable (skewnesses were still between -2 and +2). This is based on the general rule that skewness should be between -1 and +1. ANOVA, however, is resistant to HOV and violations of normality, particularly when a sizable dataset with a well-balanced design is present. The consequences of these violations were therefore disregarded.

Following the factorial ANOVA results, further analyses were conducted to explore the main and interaction effects for rccS and rccA. For follow-up analyses in the interactions, the dataset was divided based on one of the factors in the interaction, and the effects of the other conditions were assessed based on rccS and rccA. To evaluate prediction accuracies of specified data conditions, average rccS ($\bar{X}_{rccS}$), and average rccA ($\bar{X}_{rccA}$) were defined. $\bar{X}_{rccS}$ refers to the mean rate of correct classification for the smallest group and $\bar{X}_{rccA}$ refers to the rate of correct classification for all groups in the specified conditions. Finally, the relationship between rccS and rccA for different cases was analyzed with the Pearson correlation coefficient $r_{\text{rccS-A}}$.

Results

In this section, results for rccA, rccS and the relationship between rccA and rccS are presented separately.

Results for rccA

The overall factorial ANOVA model was statistically significant ($p < .001, \eta_p^2 = .969$) for the outcome variable rccA. All main effects and interactions were significant ($p < .001$). Based on partial eta squared (η_p^2) values GSR ($\eta_p^2 = .914$) was the most influential main effect, while GN ($\eta_p^2 = .907$), Method ($\eta_p^2 = .702$), NPV ($\eta_p^2 = .683$) and Corr ($\eta_p^2 = .502$) had smaller effects. Among all the two-way interactions, Method*GSR ($\eta_p^2 = .485$) was the most influential one while GN*GSR ($\eta_p^2 = .44$), NPV*GN ($\eta_p^2 = .393$), Corr*GN ($\eta_p^2 = .317$), Corr*NPV ($\eta_p^2 = .278$), NPV*GSR ($\eta_p^2 = .268$), Method*GN ($\eta_p^2 = .183$), Corr*GSR ($\eta_p^2 = .16$), Method*Corr ($\eta_p^2 = .156$) and Method*NPV ($\eta_p^2 = .059$) were decreasingly influential. Moreover, Corr*NPV*GN ($\eta_p^2 = .177$) and Method*GN*GSR ($\eta_p^2 = .169$) were the most influential three-way effects, while all the other three-way effects had partial eta squared values less than .1. Finally, all the four-way interactions and the single five-way interaction (Method*Corr*NPV*GN*GSR) had a partial eta squared value less than .1.

Only main effects for rccA are reported here since the focus of this study was prediction of the smallest group. Mean rccA for all the cases was .694 and mean rccA values for levels of main effects are presented at Table 1.

Table 1
rccA Values for Levels of Main Effects: Method, Corr, NPV, GN and GSR

Method	$\bar{X}_{rccA}$	Corr	$\bar{X}_{rccA}$	NPV	$\bar{X}_{rccA}$	GN	$\bar{X}_{rccA}$	GSR	$\bar{X}_{rccA}$
LDA	.655	.2	.718	2	.649	2	.800	Balanced	.613
LR	.681	.5	.669	5	.695	3	.658	Unbalanced	.774
CART	.745			10	.737	4	.622		

Notes. $\bar{X}_{rccA}$: Average rccA, Corr: Correlation, NPV: Number of the predictor variables, GN: Number of groups in dependent variable

In rccA, all groups of main effects had prediction accuracy of more than 60%. CART was highest performing method with .745 mean rccA, while LR and LDA had .681 and .655 mean rccA, respectively. As it can be seen from Table 1, higher correlation and higher group numbers resulted in lower mean rccA, while higher NPV resulted in higher rccA. Moreover, unbalanced cases had a greater rccA than balanced cases.

In most cases, CART performed better than LR and LDA. On the other hand, in the case of 10 predictor variables when Corr was .2, GSR was unbalanced, and when Corr was .5, GSR was balanced, LR performed better than CART and LDA. Moreover, when GSR was unbalanced and GN was 4, the cases when Corr was .2 or .5 and NPV was 2, 5 or 10 (6 cases) differences between LR, LDA and CART were trivial.

Results for rccS

Details of the overall factorial ANOVA results for rccS are provided in Table 2.

Table 2

Factorial ANOVA Table for the Effects of Method, Corr, NPV, and GN on rccS

Source	df	F	p	η_p^2
Method	2	2319.797	<.001	.079
Corr	1	6434.903	<.001	.107
NPV	2	9079.160	<.001	.252
GN	2	47265.025	<.001	.637
Method * Corr	2	471.969	<.001	.017
Method * NPV	4	72.635	<.001	.005
Method * GN	4	426.726	<.001	.031
Corr * NPV	2	805.290	<.001	.029
Corr * GN	2	654.058	<.001	.024
NPV * GN	4	359.320	<.001	.026
Method * Corr * NPV	4	71.864	<.001	.005
Method * Corr * GN	4	130.099	<.001	.010
Method * NPV * GN	8	189.364	<.001	.027
Corr * NPV * GN	4	56.785	<.001	.004
Method * Corr * NPV * GN	8	29.854	<.001	.004
Error	53946			
Total	53999			

Notes. Corr: Correlation, NPV: Number of the predictor variables, GN: Number of groups in the dependent variable

The overall factorial ANOVA model for rccS was statistically significant and had a meaningful partial eta squared value ($p < .001, \eta_p^2 = .712$). All the interactions and main effects were statistically significant ($p < .001$). Based on partial eta square values, GN ($\eta_p^2 = .637$) was the most influential effect, and NPV ($\eta_p^2 = .252$), Corr ($\eta_p^2 = .107$) and the method ($\eta_p^2 = .079$) were, in order, smaller. Interaction between Method and GN (Method*GN) ($\eta_p^2 = .031$) was the most effective two-way interaction, while interaction between Corr and NPV ($\eta_p^2 = .029$), NPV and GN ($\eta_p^2 = .026$), Corr and GN ($\eta_p^2 = .024$), Method and Corr ($\eta_p^2 = .017$) and, Method and NPV ($\eta_p^2 = .005$) were the two-way effects, in order. In addition, the interaction between Method, NPV, and GN ($\eta_p^2 = .027$) was the strongest three-way interaction, while Method*Corr*GN ($\eta_p^2 = .010$), Method*Corr*NPV ($\eta_p^2 = .005$) and Corr*NPV*GN ($\eta_p^2 = .004$) were smaller. Finally, the only four-way interaction was the interaction between Method, Corr, NPV, and GN had effect size $\eta_p^2 = .004$.

Main Effects in rccS

As stated above GN had a greater effect than the other variables on rccS and Method had the lowest effect. Among all the unbalanced cases the overall mean rccS was .325 and mean rccS values for the method, levels of correlation, NPV, and GN are shown in Table 3.

Table 3

Overall Mean rccS for Levels of Correlation, NPV, GN and Methods

Method	$\bar{X}_{rccS}$	Corr	$\bar{X}_{rccS}$	NPV	$\bar{X}_{rccS}$	GN	$\bar{X}_{rccS}$
LDA	.291	.2	.371	2	.226	2	.099
LR	.302	.5	.278	5	.329	3	.335
CART	.380			10	.418	4	.539

Notes. $\bar{X}_{rccS}$*:* Average rccS, Corr = Correlation, NPV = Number of the predictor variables, GN = Number of groups in dependent variable

Based on the findings, LDA with .291 average rccS demonstrated the lowest overall performance, followed by LR with .302 and CART with .38 average rccS, which was the highest performing method in rccS overall. In terms of correlation, a lower degree of correlation (.2) resulted in better performance than a higher degree of correlation (.5). Moreover, the cases having a larger number of predictor variables had better performance in terms of rccS such as cases of 2 predictor variables had average rccS of .226, while cases of 5 and 10 predictor variables had .329 ad .418 average rccS, respectively. Finally, having more groups resulted in greater average rccS in this setting. Change in method in terms of average rccS from highest accuracy to lowest was .089, while change in Corr was .093, change in NPV was .192 and change in GN was .44. Therefore, it can be observed that data conditions had greater effects than method in terms of prediction accuracy of small groups.

While evaluating mean rccS values for main effects gives an overall idea about prediction accuracy for the smallest groups, it is important to evaluate interactions so that change in prediction accuracy for a main effect when change in other factors occurs may be investigated. Therefore, for the main effect of Method, data were divided into groups, and prediction accuracies were evaluated based on changes in other variables.

Two-way interactions in rccS

All the two-way interactions for rccS were statistically significant but had smaller effect sizes compared to the effect sizes of the main effects. Comparing Method interactions with other variables based on partial eta squared values, it was observed that the interaction of Method with GN ($\eta_p^2 = .031$) had a greater effect than the interaction of Method with Corr ($\eta_p^2 = .017$), and interaction of Method with NPV ($\eta_p^2 = .005$). Mean rccS scores of the methods at the levels of Corr, NPV, and GN are presented in Table 4.

Table 4

Mean rccS Values for Interactions of Method with Corr, NPV and GN

Method	Corr	$\bar{X}_{rccS}$	NPV	$\bar{X}_{rccS}$	GN	$\bar{X}_{rccS}$
LDA	.2	.348	2	.186	2	.048
	.5	.234	5	.295	3	.305
			10	.391	4	.519
LR	.2	.364	2	.193	2	.050
	.5	.240	5	.307	3	.316
			10	.407	4	.540
CART	.2	.402	2	.298	2	.200
	.5	.359	5	.387	3	.384
			10	.456	4	.557

Notes. $\bar{X}_{rccS}$: Average rccS, Corr: Correlation, NPV: Number of the predictor variables, GN: Number of groups in dependent variable.

Increasing Corr resulted in decreases in the mean rccS for all the methods; increasing Corr from .2 to .5 resulted in .114 decrease in LDA, .124 decrease in LR and .043 decrease in CART for mean rccS. Therefore, it can be inferred that CART is the least affected method by the change in Corr, and LR and LDA had similar changes in mean rccS when changing Corr. On the other hand, increasing NPV resulted in increases in the mean rccS for all the methods. Increasing NPV from 2 to 10 resulted in .205 increase in LDA, .214 increase in LR and .158 increase in CART. Thus, CART was the least affected model in the change of NPV and LR and LDA had similar performances in favor of LR. Finally, increasing GN resulted in increases in the mean rccS for all the methods; increasing GN from 2 to 4 resulted in .471 increase in LDA, .490 increase in LR and .357 increase in CART. Thus, the change in GN had a greater impact on LR and LDA than CART. In conclusion, it was observed that LR was most the sensitive method to data conditions, while LDA was the second and CART was the least affected method by data conditions. GN was the most influential data condition on the method's rccS performances, and NPV and Corr were lesser.

Besides two-way interactions, three-way interactions were also analyzed in detail, as the effect size for Method*NPV*GN ($\eta_p^2 = .027$) was close to the effect sizes of two-way interactions. The four-way interaction was not inspected due to the small effect size ($\eta_p^2 = .004$).

Three-way Interactions in rccS

There were four three-way interactions in the design of this study, and the interactions that included Method were evaluated in detail. Mean rccS values for the interaction between Method, NPV, and GN are presented in Table 5.

Table 5

Mean rccS Values of the Methods for the Different Levels of NPV and NG

NPV	GN	$\bar{X}_{rccS}(LDA)$	$\bar{X}_{rccS}(LR)$	$\bar{X}_{rccS}(CART)$
2	2	.013	.013	.102
	3	.174	.181	.297
	4	.372	.386	.493
5	2	.038	.039	.206
	3	.309	.328	.382
	4	.536	.552	.573
10	2	.094	.099	.291
	3	.431	.441	.472
	4	.648	.683	.606

Notes. $\bar{X}_{rccS}(LDA)$: Average rccS in LDA, $\bar{X}_{rccS}(LR)$: Average rccS in LR, $\bar{X}_{rccS}(CART)$: Average rccS in CART, GN: Number of groups in dependent variable, NPV: Number of the predictor variables.

According to the results presented in Table 5, when controlling for Method and NPV, an increase in GN resulted in an increase in mean rccS in all the levels of NPV in the methods. When there were 2 predictor variables, increasing the number of groups from 2 to 4 LDA increased the mean rccS score from .013 to .372 (difference = .359) while LR increased the mean rccS score from .013 to .386 (difference = .373) and CART from .102 to .493 (difference = .391). Thus, CART was the model that was improved most by the change in GN. Moreover, CART was the best performing model for all NPV cases when GN was 2. Similarly, CART was the best performing model in the case when there were 5 predictor variables, and LR was the most improved model in rccS (from .039 to .552, difference = .513). Similarly, in the case when there were 10 predictor variables LR was the most improved model in rccS and it was the best performing model when the number of groups was 4. On the other hand, when the numbers of the groups were 2 and 3, CART was the best performing method. Thus, increasing NPV and GN produce results in favor of LR and LDA rather than CART.

For three-way interaction Method*Corr*GN, average rccS values for the methods at different levels of Corr and GN are presented in Table 6.

Table 6

Mean rccS Values for the Methods at Different Levels of Corr and GN

Corr	GN	$\bar{X}_{rccS}(LDA)$	$\bar{X}_{rccS}(LR)$	$\bar{X}_{rccS}(CART)$
.2	2	.065	.068	.219
	3	.376	.396	.401
	4	.603	.629	.586
.5	2	.032	.033	.181
	3	.234	.237	.366
	4	.435	.451	.529

Notes. $\bar{X}_{rccS}(LDA)$: Average rccS in LDA, $\bar{X}_{rccS}(LR)$: Average rccS in LR, $\bar{X}_{rccS}(CART)$: Average rccS in CART, Corr: Correlation, GN: Number of groups in dependent variable

According to the results in Table 6, at the fixed levels of Corr, an increase in GN resulted in an increase in average rccS for all the methods. When correlations between variables were .2, increasing the number of groups 2 to 4, LDA improved mean rccS from .065 to .603 (difference = .538) while LR improved mean rccS score from .068 to .629 (difference = .561) and CART from .219 to .586 (difference = .367). Hence, LR was the most affected model by the change in GN. Moreover, while CART was the best performing model in cases when there were 2 or 3 groups, LR was the best performing model for the case when there were 4 groups.

Table 7

Mean rccS Values for the Methods at Different Levels of Corr and NPV

Corr	NPV	$\bar{X}_{rccS}(LDA)$	$\bar{X}_{rccS}(LR)$	$\bar{X}_{rccS}(CART)$
.2	2	.207	.214	.310
	5	.356	.371	.402
	10	.480	.508	.494
.5	2	.165	.172	.286
	5	.233	.242	.372
	10	.303	.307	.419

Notes. $\bar{X}_{rccS}(LDA)$: Average rccS in LDA, $\bar{X}_{rccS}(LR)$: Average rccS in LR, $\bar{X}_{rccS}(CART)$: Average rccS in CART, Corr: Correlation, NPV: Number of the predictor variables.

Fixing Corr at .2, an increase in NPV resulted in an increase in mean rccS for all the methods; increasing NPV from 2 to 10 mean rccS in LDA increased from .207 to .48 (difference = .273), in LR from .214 to .508 (difference = .294), and in CART from .31 to .494 (difference = .184). Hence, in the cases when Corr was .2, CART was the least affected method by the change in NPV and it was notable that LR exceeded the CART in terms of mean rccS at the highest level of NPV. On the other hand, for the cases when Corr was .5 change in NPV from 2 to 10 resulted in similar changes in rccS's of LDA (difference = .138), LR (difference = .135), and CART (difference = .133). Furthermore, CART's performance was superior to the other two methods when Corr was .5 at all the different levels of NPV.

Relationship Between rccS and rccA

To analyze the relationship between the smallest group prediction accuracy and prediction accuracy for all groups, the Pearson correlation coefficient was first employed for all the cases together, then for the different levels of main effects, and finally, for different levels of main effects at different levels of GN.

The whole data for rcsS and rccA were normally distributed based on skewness values between -1 and 1. Besides, the rccS and rccA values demonstrated a normal distribution for the main effects and their respective levels within the GN levels, with skewness ranging from -1 to 1. However, exceptions occurred when GN was 2, where skewness values for rccS and rccA ranged from 1 to 2. Specifically, when GN was 2 and NPV was 2, the skewness for rccS reached 2.686, while for rccA it was 2.282. The outcomes of these cases were carefully analyzed and compared with Spearman correlation coefficients. The Spearman and Pearson correlation coefficients were close to each other, and differences between these values did not change the direction of the analyses, so only Pearson correlations are reported.

There was a notable difference between the overall rccS and rccA values: for all the unbalanced cases the overall mean rccS was .323, while the overall mean rccA was .774. Moreover, the correlation between rccS and rccA for all the cases was -.461, which means there was a negative and medium correlation between rccS and rccA. Besides, correlation values for different levels of main effects are presented in Table 8.

Table 8

Correlation between rccS and rccA at Different Levels of GN, Method, NPV and Corr

GN	$r_{\text{rccS}-\text{A}}$	Method	$r_{\text{rccS}-\text{A}}$	NPV	$r_{\text{rccS}-\text{A}}$	Corr	$r_{\text{rccS}-\text{A}}$
2	.797	LDA	-.515	2	-.620	.2	-.444
3	.637	LR	-.486	5	-.629	.5	-.568
4	.676	CART	-.477	10	-.463		

Notes. $r_{\text{rccS}-\text{A}}$: Pearson Correlation between rccS and rccA, Corr: Correlation, NPV: Number of the predictor variables, GN: Number of groups in dependent variable.

When there were 2 groups, the correlation between rccS and rccA was .797 while it was .637 and .676 for the cases of group number were 3 and 4, respectively. In LDA, $r_{\text{rccS}-\text{A}}$ was -.515 while it was -.486 and -.477 in LR and CART, respectively. Moreover, it was -.620, -.629 and -.463 when the number of predictor variables was 2, 5, and 10, respectively. Finally, in the case when the correlation between variables was .2, the correlation between rccA and rccS was -.444, while it was -.568 when the correlation between predictor variables was .5. Since the correlation between rccS and rccA was negative for the groups of method, NPV and Corr and it was positive for GN, a more detailed analysis was conducted by splitting data into GN for further analysis. Correlation values between rccS and rccA at the levels of the method, NPV, and Corr into levels of GN are presented in Table 9.

Table 9

Correlation between rccS and rccA at the Levels of Method, NPV and Corr for Fixed Levels of GN

GN	Method	$r_{\text{rccS}-\text{A}}$	NPV	$r_{\text{rccS}-\text{A}}$	Corr	$r_{\text{rccS}-\text{A}}$
2	LDA	.762	2	.792	.2	.794
	LR	.791	5	.787	.5	.798
	CART	.818	10	.775		
3	LDA	.766	2	.532	.2	.623
	LR	.775	5	.489	.5	.593
	CART	.457	10	.537		
4	LDA	.767	2	.431	.2	.705
	LR	.784	5	.489	.5	.483
	CART	.352	10	.684		

Notes. $r_{\text{rccS}-\text{A}}$: Pearson Correlation between rccS and rccA, Corr: Correlation, NPV: Number of the predictor variables, GN: Number of groups in dependent variable.

On splitting data by GN, $r_{\mathrm{rccS-A}}$ was positive for all levels of Method and NPV and Corr even though it was negative before splitting. This demonstrates the impact of GN on the relationship between rccS and rccA. In the case when GN was 2, for the methods, the highest correlation between rccS and rccA was for CART ($r_{\mathrm{rccS-A}} = .818$) and the lowest correlation was for LDA ($r_{\mathrm{rccS-A}} = .762$). For different degrees of NPV and Corr correlations between rccS and rccA were high and there were trivial differences in terms of $r_{\mathrm{rccS-A}}$. In the case when GN was 3, there was no notable difference between LR ($r_{\mathrm{rccS-A}} = .766$), LDA ($r_{\mathrm{rccS-A}} = .775$) and CART ($r_{\mathrm{rccS-A}} = .457$) in terms of $r_{\mathrm{rccS-A}}$. Moreover, for different levels of NPV and Corr when GN was 3, there were not important differences in terms of $r_{\mathrm{rccS-A}}$. Finally, when GN was 4, the difference between LR and CART in terms of $r_{\mathrm{rccS-A}}$ became greater since $r_{\mathrm{rccS-A}}$ was .784 for LR and .352 for CART. Moreover, increasing NPV resulted in increase in $r_{\mathrm{rccS-A}}$ while increasing Corr resulted in a decrease in $r_{\mathrm{rccS-A}}$. Finally, differences in $r_{\mathrm{rccS-A}}$ between cases of .2 Corr and .5 Corr when GN was 2, 3, and 4 were .004, .03, and .222, respectively. Thus, when GN was 4 the difference was notably greater than the cases when GN was 2 and 3.

Discussion

This study delved into the comparative effectiveness of three prevalent classification methods CART, LDA, and LR to evaluate their performance in predicting group membership specifically for proportionally small groups across various controlled conditions. Even though there were certain instances in which LR performed better, one of the primary findings that emerged from this research was that CART consistently displays superior performance across most settings. LR tended to outperform LDA and CART in the cases with a high number of predictor variables, low correlation between variables, and an abundance of groups. Consistent with these results, specifically in the simulation studies, the superiority of CART is supported by existing research (Finch et al., 2014; Holden et al., 2011). In addition to that, for most of the cases, LR and LDA had similar performances, though in almost every case LR showed slightly better accuracy. Hence, even though there are conflicting findings indicating that LDA performs better than LR (Williams, 1999), the finding that LR performs better than LDA (Barön, 1991) or CART, particularly when assumptions for LDA are satisfied, and that there are insignificant differences between LR and LDA (Hestie et al., 2009) are supported by the literature.

This research demonstrates that an important component affecting prediction accuracy is the ratio of group sizes, especially when evaluating smaller groups' predictions. This emphasizes the unequal impact that group size can exert on classification accuracy. Moreover, the number of groups is identified as a significant determinant of accuracy. In agreement with previous studies, an increase in the number of groups resulted in a decrease in overall prediction accuracy (Finch & Schneider, 2007; Pohar et al.,2004). On the other hand, this study also demonstrated that the prediction accuracy of small groups was enhanced as the number of groups increased.

By the design of this study, the number of groups is engaged with degrees of group separation. Since groups were separated by a determined mean difference between consecutive groups, cases with a higher number of groups had greater levels of group separation. For example, for the two group cases, the mean difference between large and small groups was .5 while for the four group cases difference between large and small groups was 1.5. Therefore, differences between large and small groups might affect discrimination and prediction of small groups. When group sizes are unbalanced and group separation is large, small groups can be recognized more accurately. Still, this research highlights that smaller group classification accuracy benefits from an augmentation in the number of groups. Consequently, a larger group separation makes it easier to predict smaller groups and smaller groups are more readily discriminated from larger groups. On the other hand, overall group prediction may be decreased due to the members of larger groups predicted as in the smaller groups. Besides, the performance of methods for overall classification diminishes with an increase in the number of groups, signifying that managing multi-group situations continues to be difficult. It was concluded that all the controlled conditions had a greater impact on small group prediction than on overall prediction accuracy in terms of the percentage of correctly predicted observations. Finally, the results showed that all the controlled data conditions had a greater impact on the accuracy of small group prediction than on overall group predictions.

In this study, it was found that an increase in the number of predictor variables improved the classification accuracy across all the methods and data conditions. This finding aligned with Finch and Schneider (2007) who stated that the accuracy of group membership prediction is improved with the addition of new predictor factors. This pattern appears stronger in LDA and LR compared to CART, indicating that these two methods might have a superior ability to utilize complicated, high-dimensional data.

In line with earlier studies, correlation influenced classification accuracy (Kiang, 2003). Furthermore, the results of this study align with the observation made by Pai et al. (2012) concerning the ineffectiveness of multicollinear factors, as increased correlation diminishes the contributions of additional variables. The maximum correlation level for this study was .5; so, at higher values, negligible or little contributions may be anticipated. This study revealed that the impact of correlation was diminished for CART compared to LR and LDA regarding overall and small group prediction accuracy. When predictor variables demonstrate minimal correlations, predictive accuracy often increases, benefiting all three techniques, especially CART and LR. This enhancement is particularly important for smaller groups, where precise classifications are critical. Furthermore, CART demonstrates superior robustness in managing imbalanced datasets compared to LDA and LR, which often encounter difficulties in such scenarios. Nonetheless, LR exhibits optimal performance when the data is balanced and evenly distributed among groups.

This study indicates that overall prediction accuracy is remarkably greater than that of small group prediction accuracy, a conclusion corroborated by Chiang (2021). This study also highlights the correlation between the accuracy of predictions for all groups and the accuracy of predictions for the smallest groups. In all the situations, a moderate negative correlation was found; however, for the same number of groups, a significant positive correlation was found. Therefore, the impact of group size and degree of separation on the relationship between small and overall group prediction accuracy was examined. It was concluded that small group and overall group prediction accuracies have parallel characteristics at the same number of groups, while for mixed numbers of groups they tend to have inverse characteristics.

This study makes useful suggestions for practitioners: Less than 10 predictors and smaller groups are best suited for CART, but larger datasets with more groups and predictor variables are better suited for LR. However, unless certain requirements are satisfied, such as equal covariance and normality, LDA is not advised.

While this study offers a thorough evaluation of the performance of CART, LDA, and LR in terms of small group prediction, in addition to the effect of the data conditions on prediction accuracy, it recognizes a few limitations. Since the study uses simulated data, it might not accurately represent actual circumstances. For instance, the data's group separation was maintained at fixed standardized mean differences, which restricts the study's generalizability to situations with non-normal distributions or variable group separation. Further research is encouraged to investigate the consequences of varied sample sizes, non-normal data distributions, and variable levels of group separation. The complex nature of numerous controlled circumstances necessitated the simulation of data under the assumption of multivariate normality for each category, representing an additional restriction of this work. Additionally, factors such as the presence of categorical predictor variables, multimodality, varying sample sizes between groups, and heterogeneity of variance-covariance matrices were not addressed in this work.

It is advised to look at more recent approaches that may provide better results in specific situations, like support vector machines, random forests, and neural networks, as well as investigating more sophisticated classification methods outside of CART, LDA, and LR. The handling of unbalanced datasets and methods for improving the classification of smaller groups are two areas of special interest for further study. This is particularly important because smaller groups frequently have less prediction accuracy, which can produce biased results in practical applications. Furthermore, particular attention should be paid to how existing techniques might be enhanced to optimize accuracy and judgment in progressively difficult classification tasks, thus promoting the field of predictive modeling.

In summary, the study offers a comprehensive analysis of three widely used classification techniques, highlighting their performance in controlled settings. CART is notable for its adaptability, yet in high-dimensional, multi-group situations, LR proves to be a formidable competitor. For LDA to work effectively, stricter requirements must be met. Practitioners looking to select the best approach for their data classification requirements might benefit from the study's insights. We encourage future developments in classification techniques, especially when handling unbalanced data and smaller groups, indicating the significance of ongoing research and development in the predictive modeling space.

Declarations

This paper was adapted from the first author's doctoral dissertation. A part of this paper was presented at the 2019 annual meeting of the American Educational Research Association (AERA).

Conflict of Interest: The authors declare no conflicts of interest.

Ethical Approval: This study utilized simulated data; therefore, ethical approval was not required. The authors confirm that all ethical guidelines were followed.

References

Agresti, A. (2013). *Categorical data analysis* (3rd ed.). Wiley.

Barön, A. E. (1991). Misclassification among methods used for multiple group discrimination-the effects of distributional properties. *Statistics in Medicine*, *10*(5), 757-766. doi: https://doi.org/10.1002/sim.4780100511

Bates, B. E., Xie, D., Kwong, P. L., Kurichi, J. E., Ripley, D. C., & Stineman, M. G. (2014). One-year all-cause mortality after stroke: A prediction model. *PM&R*, *6*(6), 473-483. doi: https://doi.org/10.1016/j.pmrj.2013.11.006

Bolin, J., & Finch, W. (2014). Supervised classification in the presence of misclassified training data: A Monte Carlo simulation study in the three-group case. *Frontiers in Psychology*, *5* doi:10.3389/fpsyg.2014.00118

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

Castonguay, A. C., Zoghi, Z., Zaidat, O. O., Burgess, R. E., Zaidi, S. F., Mueller-Kronast, N., ... & Jumaa, M. A. (2023). Predicting functional outcome using 24-hour post-treatment characteristics: Application of machine learning algorithms in the STRATIS registry. *Annals of Neurology*, *93*(1), 40-49. https://doi.org/10.1002/ana.26528

Chiang, Y. C. (2021). *Evaluating the performance of classification and regression trees, random forests, and K-means clustering under controlled conditions* (Doctoral dissertation, Indiana University).

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Earlbaum Associates.

Dattalo, P. (1995). A comparison of discriminant analysis and logistic regression. *Journal of Social Service Research*, *19*(3-4), 121-144.

Dey, E. L., & Astin, A. W. (1993). Statistical alternatives for studying college student retention: A comparative analysis of logit, probit, and linear regression. *Research in Higher Education*, *34*, 569-581. doi: https://doi.org/10.1007/BF00991920

Edwards, A. L. (1985). *Experimental design in psychological research* (5th ed.). NY: Harper & Row.

Fan, X., & Wang, L. (1999). Comparing linear discriminant function with logistic regression for the two-group classification problem. *The Journal of Experimental Education*, *67*(3), 265-286. doi:10.1080/00220979909598356

Ferrer, A. J. A., & Wang, L. (1999). *Comparing the Classification Accuracy among Nonparametric, Parametric Discriminant Analysis and Logistic Regression Methods* (pp. 1-24, Rep.). Montreal: Paper presented at the Annual Meeting of the American Educational Research Association. https://eric.ed.gov/?id=ED432591

Finch, H. W., Bolin, J. E., & Kelley, K. (2014). Group membership prediction when known groups consist of unknown subgroups: A Monte Carlo comparison of methods. *Frontiers in Psychology*, *5*. doi:10.3389/fpsyg.2014.00337

Finch, H., & Schneider, M. K. (2007). Classification accuracy of neural networks vs. discriminant analysis, logistic regression, and classification and regression trees: Three- and five-group cases. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *3*(2), 47-57. doi:10.1027/1614-2241.3.2.47

Finch, W. H., & Schneider, M. K. (2006). Misclassification rates for four methods of group classification. *Educational and Psychological Measurement*, *66*(2), 240-257. doi:10.1177/0013164405278579

Graf, R., Zeldovich, M., & Friedrich, S. (2023). Comparing linear discriminant analysis and supervised learning algorithms for binary classification—A method comparison study. *Biometrical Journal*, *66*(1), 2200098. doi: https://doi.org/10.1002/bimj.202200098

Grassi, M., Villani, S., & Marinoni, A. (2001). Classification methods for the identification of "case" in epidemiological diagnosis of asthma. *European Journal of Epidemiology, 17, 19-29.* doi: https://doi.org/10.1023/A:1010987521885

Hassan, H. A., Hemdan, E. E. D., El-Shafai, W., Shokair, M., & Abd El-Samie, F. E. (2024). Detection of attacks on software defined networks using machine learning techniques and imbalanced data handling methods. *Security and Privacy*, *7*(2), e350. doi: https://doi.org/10.1002/spy2.350

Hastie, T., Tibshirani R., Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

Hess, B., Olejnik, S., & Huberty, C. J. (2001). The efficacy of two improvement-over-chance effect sizes for two-group univariate comparisons under variance heterogeneity and nonnormality. *Educational and Psychological Measurement*, *61*(6), 909-936. doi: https://doi.org/10.1177/00131640121971572

Hao, Z., Yang, B., Ruggiano, N., Ma, Y., Guo, Y., & Pan, X. (2022). Depression prediction amongst Chinese older adults with neurodegenerative diseases: A performance comparison between decision tree model and logistic regression analysis. *The British Journal of Social Work*, *52*(1), 274-290. doi: https://doi.org/10.1093/bjsw/bcaa237

Hoang, M. L., Matrella, G., & Ciampolini, P. (2025). Metrological evaluation of contactless sleep position recognition using an accelerometric smart bed and machine learning. *Sensors and Actuators A: Physical*, *385*, 116309. doi: https://doi.org/10.1016/j.sna.2025.116309

Holden, J. E., Finch, W. H., & Kelley, K. (2011). A comparison of two- group classification methods. *Educational and Psychological Measurement, 71*(5), 870-901. doi:10.1177/0013164411398357

Holden, J. E., & Kelley, K. (2010). The effects of initially misclassified data on the effectiveness of discriminant function analysis and finite mixture modeling. *Educational and Psychological Measurement, 70*(1), 36-55. doi:10.1177/0013164409344533

Huberty, C. J. (1994). *Applied discriminant analysis*. John Wiley & Sons.

Huberty, C. J., & Olejnik, S. (2006). *Applied MANOVA and discriminant analysis* (Vol. 498). John Wiley & Sons.

IBM Corp. (2025). IBM SPSS Statistics for Windows, Version 30.0.0. Armonk, NY: IBM Corp. Retrieved from https://www.ibm.com/us-en/marketplace/statistical-analysis-and-reporting

Kiang, M. Y. (2003). A comparative assessment of classification methods. *Decision Support Systems*, *35*(4), 441-454. doi: https://doi.org/10.1016/S0167-9236(02)00110-0

Lei, P., & Koehly, L. (2003). Linear discriminant analysis versus logistic regression: A comparison of classification errors in the two-group case. *Journal of Experimental Education*, *72*(1), 25-49. doi: https://doi.org/10.1080/00220970309600878

Meshbane, A., & Morris, J. D. (1995). A method for selecting between linear and quadratic classification models in discriminant analysis. *Journal of Experimental Education*, *63*(1), 263-273. doi: https://doi.org/10.1080/00220973.1995.9943813

Pai, D. R., Lawrence, K. D., Klimberg, R. K., & Lawrence, S. M. (2012). Analyzing the balancing of error rates for multi-group classification. *Expert Systems with Applications*, *39*(17), 12869-12875. doi: https://doi.org/10.1016/j.eswa.2012.05.006

Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., & Chen, F. (2001). Monte Carlo Experiments: Design and implementation. *Structural Equation Modeling*, 8(2), 287-312. doi: https://doi.org/10.1207/S15328007SEM0802_7

Pierce, C. A., Block, R. A., & Aguinis, H. (2004). Cautionary note on reporting eta-squared values from multifactor ANOVA designs. *Educational and Psychological Measurement*, *64*(6), 916-924. doi: https://doi.org/10.1177/0013164404264848

Pohar, M., Blas, M., & Turk, S. (2004). Comparison of logistic regression and linear discriminant analysis: A simulation study. *Metodoloski Zvezki*, *1*(1), 143-161. http://mrvar.fdv.uni-lj.si/pub/mz/mz1.1/pohar.pdf

Preatoni, D. G., Nodari, M., Chirchella, R., Tosi, G., Wauters, L. A., & Martinoli, A. (2005). Identifying bats from time-expanded recordings of search calls: Comparing classification methods. *Journal of Wildlife Management*, *69*(1), 1601-1614. doi: https://doi.org/10.2193/0022-541X(2005)69[1601:IBFTRO]2.0.CO;2

R Core Team (2016). R: A Language and Environment for Statistical Computing. R foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.r-project.org/

Rausch, J. R., & Kelley, K. (2009). A comparison of linear and mixture models for discriminant analysis under nonnormality. *Behavior Research Methods*, *41*(1), 85-98. doi: https://doi.org/10.3758/BRM.41.1.85

Richardson, J. T. (2011). Eta squared and partial eta squared as measures of effect size in educational research. *Educational Research Review*, *6*(2), 135-147. doi: https://doi.org/10.1016/j.edurev.2010.12.001

Ripley, B. D. (1994). Neural networks and related methods for classification. *Journal of the Royal Statistical Society: Series B (Methodological)*, *3*(1), 409-456. doi: https://doi.org/10.1111/j.2517-6161.1994.tb01990.x

Saboor, A., Usman, M., Ali, S., Samad, A., Abrar, M. F., & Ullah, N. (2022). A method for improving prediction of human heart disease using machine learning algorithms. *Mobile Information Systems*, *2022*(1), 1410169. doi: https://doi.org/10.1155/2022/1410169

Sainani, K. L. (2014). Explanatory versus predictive modeling. *PM&R*, *6*(9), 841-844. doi: https://doi.org/10.1016/j.pmrj.2014.08.941

Schumacher, M., Rossner, R., & Vach, W. (1996). Neural networks and logistic regression: Part I. *Computational Statistics: Data Analysis*, *21*(1), 661-682. doi: https://doi.org/10.1016/0167-9473(95)00032-1

Selim, G. E. I., Hemdan, E. E. D., Shehata, A. M., & El-Fishawy, N. A. (2021). Anomaly events classification and detection system in critical industrial internet of things infrastructure using machine learning algorithms. *Multimedia Tools and Applications*, *80*(8), 12619-12640. doi: https://doi.org/10.1007/s11042-020-10354-1

Soureshjani, M. H., & Kimiagari, A. M. (2013). Calculating the best cut off point using logistic regression and neural network on credit scoring problem-A case study of a commercial bank. *African Journal of Business Management*, *7*(16), 1414. doi: 10.5897/AJBM11.394

Song, G., Ai, Z., Zhang, G., Peng, Y., Wang, W., & Yan, Y. (2022). Using machine learning algorithms to multidimensional analysis of subjective thermal comfort in a library. *Building and Environment*, *212*, 108790. doi: https://doi.org/10.1016/j.buildenv.2022.108790

Suh, S., Suh, J., & Houston, I. (2007). Predictors of categorical at-risk high school dropouts. *Journal of Counseling & Development*, *85*(2), 196-203. doi: https://doi.org/10.1002/j.1556-6678.2007.tb00463.x

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics*. Boston: Pearson Education.

Williams, C. J., Lee, S. S., Fisher, R. A., & Dickerman, L. H. (1999). A comparison of statistical methods for prenatal screening for Down syndrome. *Applied Stochastic Models in Business and Industry*, *15*(2), 89-101. doi: https://doi.org/10.1002/(SICI)1526-4025(199904/06)15:2<89::AID-ASMB366>3.0.CO;2-K

Zampogna, B., Torre, G., Zampoli, A., Parisi, F., Ferrini, A., Shanmugasundaram, S., ... & Papalia, R. (2024). Can machine learning predict the accuracy of preoperative planning for total hip arthroplasty, basing on patient-related factors? An explorative investigation on Supervised machine learning classification models. *Journal of Clinical Orthopaedics and Trauma*, *53*, 102470. doi: https://doi.org/10.1016/j.jcot.2024.102470

Zavorka, S., & Perrett, J. J. (2014). Minimum sample size considerations for two-group linear and quadratic discriminant analysis with rare populations. *Communications in Statistics-Simulation and Computation*, *43*(7), 1726-1739. doi: https://doi.org/10.1080/03610918.2012.744041