8-1-2018

Performance Evaluation of Logistic Regression, Linear Discriminant Analysis, and Classification and Regression Trees Under Controlled Conditions

Cahit Polat
University of Denver

Performance Evaluation of Logistic Regression, Linear Discriminant Analysis, and Classification and Regression Trees Under Controlled Conditions

Abstract

Logistic Regression (LR), Linear Discriminant Analysis (LDA), and Classification and Regression Trees (CART) are common classification techniques for prediction of group membership. Since these methods are applied for similar purposes with different procedures, it is important to evaluate the performance of these methods under different controlled conditions. With this information in hand, researchers can apply the optimal method for certain conditions. Following previous research which reported the effects of conditions such as sample size, homogeneity of variancecovariance matrices, effect size, and predictor distributions, this research focused on effects of correlation between predictor variables, number of the predictor variables, number of the groups in the outcome variable, and group size ratios for the performance of LDA, LR, and CART. Data were simulated with Monte Carlo procedures in R statistical software and a factorial ANOVA with follow-ups was employed to evaluate the effect of conditions on the performance of each technique as measured by proportions of correctly predicted observations for all groups and for the smallest group.

In most of the conditions for the two outcome measures, higher performances of CART than LDA and LR were observed. But, in some conditions where there were a higher number of predictor variables and number of groups with low predictor variable correlation, superiority of LR to CART was observed. Meaningful effects of methods of correlation, number or predictor variables, group numbers and group size ratio were observed on prediction accuracy of group membership. Effects of correlation, group size ratio, group number, and number of predictor variables on prediction accuracies were higher for LDA and LR than CART. For the three methods, lower correlation and greater number of predictor variables yielded higher prediction accuracies. Having balanced data rather than imbalanced data and greater group numbers led to lower group membership prediction accuracies for all groups, but having more groups led to better predictions for the small group. In general, based on these results, researchers are encouraged to apply CART in most conditions except for the cases when there are many predictor variables (around 10 or more) and non-binary groups with low correlations between predictor variables, when LR might provide more accurate results.

Document Type

Dissertation

Degree Name

Ph.D.

Department

Quantitative Research Methods

First Advisor

Kathy Green, Ph.D.

Second Advisor

Duan Zhang

Third Advisor

Ruth Chao

Keywords

Classification, Classification and regression trees, Group membership, Linear discriminant analysis, Logistic regression, Simulation

Subject Categories

Social and Behavioral Sciences | Social Statistics

Performance Evaluation of Logistic Regression, Linear Discriminant Analysis, and Classification and Regression Trees under Controlled Conditions

_____

A Dissertation

Presented to

The Faculty of the Morgridge College of Education

University of Denver

_____

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

_____

by Cahit Polat

August 2018

Advisor: Kathy Green, Ph.D.

Author: Cahit Polat
Title: Performance Evaluation of Logistic Regression, Linear Discriminant Analysis, and Classification and Regression Trees under Controlled Conditions
Advisor: Kathy Green, Ph.D.
Degree Date: August 2018

Abstract

Logistic Regression (LR), Linear Discriminant Analysis (LDA), and Classification and Regression Trees (CART) are common classification techniques for prediction of group membership. Since these methods are applied for similar purposes with different procedures, it is important to evaluate the performance of these methods under different controlled conditions. With this information in hand, researchers can apply the optimal method for certain conditions. Following previous research which reported the effects of conditions such as sample size, homogeneity of variance-covariance matrices, effect size, and predictor distributions, this research focused on effects of correlation between predictor variables, number of the predictor variables, number of the groups in the outcome variable, and group size ratios for the performance of LDA, LR, and CART. Data were simulated with Monte Carlo procedures in R statistical software and a factorial ANOVA with follow-ups was employed to evaluate the effect of conditions on the performance of each technique as measured by proportions of correctly predicted observations for all groups and for the smallest group.

In most of the conditions for the two outcome measures, higher performances of CART than LDA and LR were observed. But, in some conditions where there were a higher number of predictor variables and number of groups with low predictor variable correlation, superiority of LR to CART was observed. Meaningful effects of methods of correlation, number or predictor variables, group numbers and group size ratio were

observed on prediction accuracy of group membership. Effects of correlation, group size ratio, group number, and number of predictor variables on prediction accuracies were higher for LDA and LR than CART. For the three methods, lower correlation and greater number of predictor variables yielded higher prediction accuracies. Having balanced data rather than imbalanced data and greater group numbers led to lower group membership prediction accuracies for all groups, but having more groups led to better predictions for the small group. In general, based on these results, researchers are encouraged to apply CART in most conditions except for the cases when there are many predictor variables (around 10 or more) and non-binary groups with low correlations between predictor variables, when LR might provide more accurate results.

Acknowledgements

I would like to thank everyone who supported me through all the steps of my education. I am grateful for the sharing, patience, passion, wisdom and uniqueness of everyone I met throughout my Ph.D. studies. I also appreciate the knowledge and talents I gained during this great endeavor. Life fortunately has shown me the results of my efforts.

I would like to recognize the support of RMIS faculty at University of Denver. Drs. Kathy Green, Bruce Uhrmacher, Duan Zhang and Antonio Olmos have had a great influence on my education in research methodologies and applied statistics. I also appreciate the head of RMIS department, Dr. Nick Cutforth, for leading a great community and for his support.

I particularly would like to thank my advisor, Dr. Kathy Green, for her supervision, help, understanding and motivation during my Ph.D. process. This dissertation would not have been possible without her support and wisdom. Dr. Green has become one of the most influential people in my life academically and personally.

I would like to also thank all the kind people of my beautiful country, Turkey. The financial support I had from Turkey's Ministry of National Education enabled me to pursue graduate-level education abroad. I believe that the experiences and education I received in the U.S.A. will allow me to provide a valuable quality of service to my country and the world.

Finally, I would like to thank my family for their support and love despite the difficult conditions we come from. As a son of a mother who has never been in an

educational institution, completing this degree is an important accomplishment for me. I hope that my story of graduate studies will inspire future generations of my family.

This dissertation is dedicated to all people who could not get the education they desire due to the inequalities and hard conditions of their lives as well as to lovely memories of my grandparents and the story of my youth.

Table of Contents

List of Tables

List of Figures

Chapter One

Introduction and Literature Review

Analysis of databases large and small is endemic across disciplines. Under different conditions of the data, statistical/analytical methods may perform differentially. The structure of the data affects the decision about which techniques to apply and hence limitations on directions of the studies. One purpose of data analysis is to determine characteristics of groups (e.g., students who stay in school versus those who drop out; patients who recover quickly from surgery versus those who do not). While many methods exist for identifying group membership of observations, logistic regression (LR) and linear discriminant analysis (LDA) are among the most commonly used (Agresti, 2002; Huberty & Olejnik, 2006) and classification and regression trees (CART) is a more recent method (Breiman et al., 1984; Williams et al.,1999). Despite their extensive use, little is known about how well they classify observations accurately and which perform better under some data scenarios such as number of the groups in the outcome variable, number of the predictor variables, distributions of predictor variables, multicollinearity, and so on.

The purpose of this study is to compare the performance of logistic regression, discriminant analysis, and classification and regression trees under conditions which are common in applied areas and so to address gaps in the literature. Furthermore, this study

aims to give suggestions to applied researchers in terms of which criteria and method to use when dealing with prediction of group membership.

Before introducing the details of LDA, LR, and CART and summarizing the studies which compared these methods, a general overview of statistical learning techniques is presented below.

A General Overview of Statistical Learning Techniques and Classification

Statistical learning techniques can be divided into five main categories based on the research purpose and data properties (Tabachnick & Fidell, 2013). A brief presentation of these categories is presented below; a detailed discussion of classification and group membership techniques is then included as the focus of this research is the comparison of three group membership techniques. The five main categories are:

1) Techniques for investigating degree of relationships between variables: Regression and correlation techniques, multipath frequency analysis, and hierarchical linear models are examples of this category.
2) Techniques for investigation of latent structure: Principal Components Analysis (PCA), Factor Analysis (FA), and Structural Equation Modeling (SEM) are examples.
3) Techniques for investigating the time course of events: Survival analysis and time series analysis are examples.
4) Techniques for investigation of group differences: t-tests, analysis of variance, analysis of covariance, their multivariate versions (MANOVA and MANCOVA and Hotelling T^2) techniques are examples.

5) Group membership techniques: Logistic regression (LR), variations of discriminant function analysis (DFA) such as linear discriminant analysis (LDA), quadratic discriminant analysis (QDA) and mixture discriminant analysis (MDA) techniques, multipath frequency analysis with logits (MFAL), and classification and regression trees (CART) are examples.

Classification is defined as a method of grouping entities by their similarities (Bailey, 1994). The general purpose of classification methods is predicting group membership of cases or grouping variables based on degrees of relationships between predictor and outcome variables. When considering classification techniques, two circumstances for classification should be considered to clarify their differences.

1) Classification techniques for grouping and investigating relationships between variables: When some variables are correlated with each other, it might be useful to reduce the number of the variables by PCA or FA techniques. Moreover, for cases when the variables are not directly observed (latent variable), the relations between latent and other observed variables based on existing literature and researcher assumptions can be explored or confirmed by SEM techniques (Kline, 2016).
2) Classification techniques for clustering observations or predicting group membership of observations: The methods such as latent class analysis (LCA), mixture modeling (MIX), or cluster analysis (CA) aim to identify unobserved group/class membership of observations. These techniques are applied when the predictor variables are observed but not the grouping variable. On the other hand,

as opposed to the cases when there are no observed groups, when the groups are observed and the interest is in predicting group membership based on predictors; i.e., what group the case is likely to belong in, LR, DA, or MFAL, can be applied depending on the choice of the researcher. In addition to these statistical methods, machine learning techniques such as classification trees and regression (CART), or random forests (RF) are applicable the same situation.

Prediction of group membership is a useful statistical learning tool in social, educational, and health sciences, and other applied areas. It plays an important role when the researcher needs to analyze the importance of predictors of the outcome (categorical) variable and more specifically in predicting group/class membership of observations. For example, in health sciences, it is important to predict whether the patient is likely to have cancer or not, based on some health conditions and indicators s/he has since correct diagnosis leads to optimizing the treatment (Valentin et al., 2001). While this is an example of a two-group case (cancer/no cancer), it should be noted that there are situations when more than two groups exist. For instance, a researcher can divide disabilities into groups including speech and language delay, autism, cerebral palsy, down syndrome, nonverbal (visuospatial) learning disability, etc., which creates more than two groups (Lillvist, 2010; Mammeralla et al., 2010).

Similar to the prediction of group membership, investigations of the significance of group differences are also useful statistical learning techniques. Both the evaluation of the group differences and group membership are based on the degree of the relationship between independent and dependent variables which defines the likelihood of

observations belonging to a group of outcome variables or the estimation of variables' importance. On the other hand, statistical group difference and group membership techniques differ as group membership is either an independent variable or the dependent variable: group difference techniques such as MANOVA include group membership variable as an independent variable and group membership techniques, such as LR, use group membership as the dependent variable (Tabachnick & Fidell, 2013). Figure 1 provides a presentation of common classification methods based on the purpose of techniques (predicting group membership of observations, grouping variables and clustering observations) and structure of variable (directly observed, unobserved).

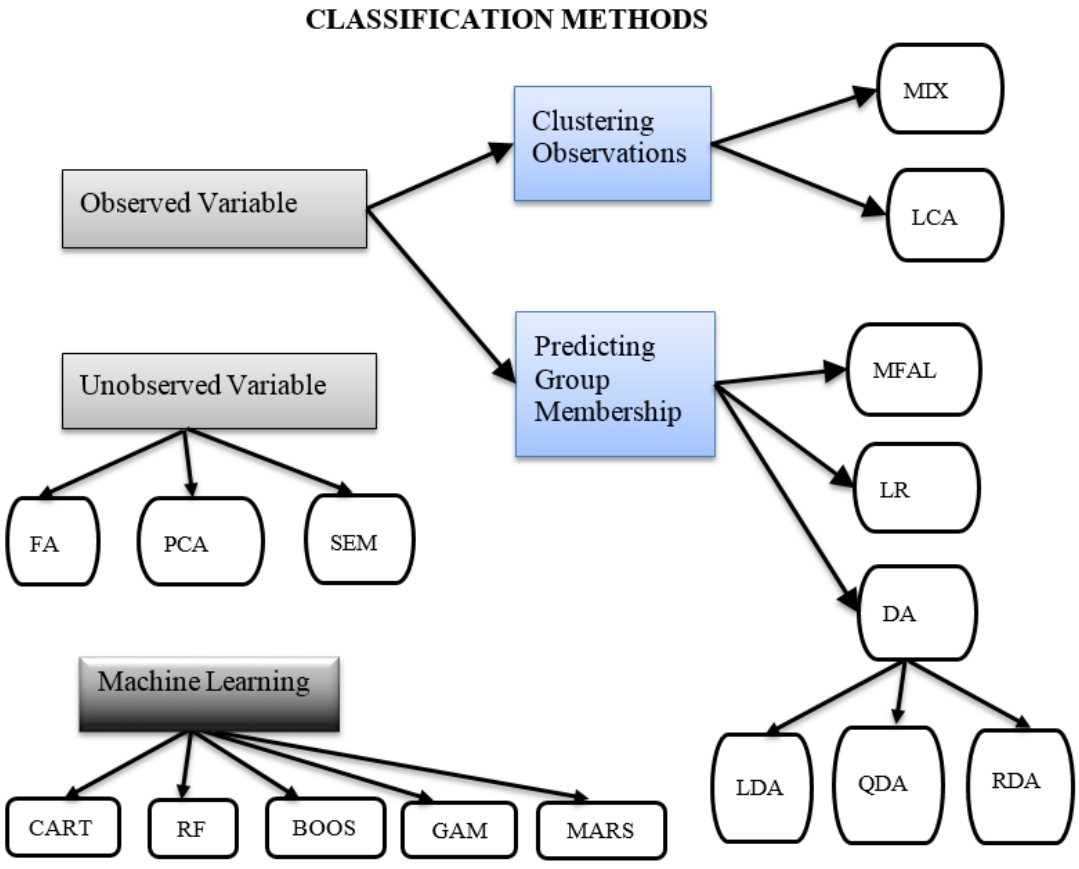

Figure 1

Common Classification Methods

Other Classification Techniques

Before presenting details of LDA, LR, and CART, it should be noted that there are classification methods in addition to those mentioned above. In general, classification techniques can be divided into categories based on the algorithms or formulas they use such as frequentist approaches, linear classifiers, Bayesian procedures, quadratic classifiers, decision trees, neural networks, and feature classification (Swain & Sarangi, 2013). Moreover, classification methods deal with problems such as: group membership (also known as supervised learning in machine learning), clustering (grouping observations to unobserved groups), and dimensionality reduction (reducing the number of variables).

In the categories above, LR and LDA fall into the category of linear classifiers. Additionally, some classification techniques use machine learning algorithms, and while most of the classification techniques are known as statistical techniques or machine learning, there is not a strict division between the ideas of statistics and machine learning in the literature as both techniques are about data analysis (Witten et al., 2017). On the other hand, some institutions such as the National Science Board, Columbia University, and UC Berkeley claim that data science and statistics are different (Ratner, 2017). Machine learning techniques use algorithms to learn from data (such as classification properties of observations) without depending on fixed programing rules and assign observations to groups. Statistical techniques rely on fixed mathematical equations that formulate relationships between variables. Hence, mechanisms of machine learning techniques are different than classical statistical classification techniques.

Generalized additive model (GAM), multivariate adaptive regression splines (MARS), and kth-Nearest neighbor (KNN) are other common statistical classification techniques and neural networks (NNET), classification and regression trees (CART), random forest (RF), and boosting (BOOS) are well known machine learning techniques. In addition, there are other different types of classification techniques such as linear programming (LP) as a mathematical optimization technique and the hybrid method (HM) as a combination of LP and KNN.

While machine-learning techniques for classification are becoming more frequently used techniques among applied researchers, logistic regression (LR) and linear discriminant analysis (LDA) are still the most commonly used techniques in social sciences for observed groups (Holden et al., 2011) while CART is used increasingly. Explanations and details about LD, LDA and CART are presented below.

Linear Discriminant Analysis (LDA)

The purpose of DA is predicting the group membership of cases/observations. It is one of the oldest and most well-known classification techniques, generalized after Fisher (Fisher, 1936; Rauch & Kelly, 2009). Throughout the past century, different discriminant functions were explored but all of them were based on similar logic or purpose. Common types of DA are linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), and regularized discriminant analysis (RDA) (Hastie et al., 2009). LDA specifically requires equality of covariance matrices, multivariate normality, and independence of observations. Additionally, it models only linear functions. On the other hand, QDA is an extended type of LDA which allows for quadratic functions. QDA

also requires multivariate normality and independence of observations but does not have a limitation for homogeneity of covariance matrices (Finch & Schneider, 2007). RDA is a mixture of LDA and QDA where covariance matrices of both methods are combined in a particular way (Friedman, 1989). While RDA and QDA are becoming more widely used, LDA is still the more commonly used classification method among researchers (Holden et al., 2011, Rauch & Kelley, 2009). Therefore, I will focus on LDA rather than the other types of discriminant analysis techniques. More details about LDA are presented below.

The classification mechanism for LDA works by calculating the following formula,

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

where

G_j is the score of the jth group,

c_{j0} is the constant value for the jth group,

c_{ji} is the coefficient value of the ith variable and the jth group,

x_i is the ith variable,

n_j is the number of observations within the jth group, and

N is the total number of observations.

Here, the constant value for the jth group c_{j0} and the coefficient values c_{ji}s are calculated by the formulas,

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

$$C_j = W^{-1} M_j \quad (3)$$

where

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.

After calculating the scores of each case for each group, an observation is assigned to the group for which its score is the highest. For example, suppose the outcome variable has three groups, let us say the group scores were calculated as G_1= 24.65, G_2= 32.09 and $G_3 = 11.40$. In this example, the observation will be assigned to the second group, since it has the highest group score.

Assumptions of LDA.

The LDA technique assumes multivariate normality, homogeneity of variance-covariance matrices (HOCV), linearity, and absence of multicollinearity and singularity. According to Tabachnick and Fidell (2013), discriminant functions are robust against violation of normality when the violation is due to skewness rather than the presence of outliers. They also state that discriminant functions are robust against the violation of HOCV as well, and violation of the assumption of linearity has little effect unless extreme. Multicollinearity and singularity occur when some predictors are redundant with each other, but some computer programs automatically exclude predictors with insufficient tolerance, which prevent analyses from failing due to singularity and multicollinearity. In terms of sample size, Tabachnick and Fidell (2003) stated that each group should have more observations than the number of predictor variables. Finally, they claim that the performance of discriminant function analysis is very sensitive to the presence of outliers.

Logistic Regression (LR)

As another group membership classification method, one purpose of LR is to correctly predict the category of the outcome variable (Agresti, 2002). Therefore, it is related to DFA and MFAL, since they all answer similar types of questions. On the other hand, LR differs from these techniques due to its flexibility, as it does not require satisfaction of some assumptions and it can include both categorical and continuous types of variables as predictor variables. Moreover, the mathematical formulation of LR is different.

To introduce the mathematical background of LR, let *u* be a linear regression model as

$$u = B_0 + B_1X_1 + B_2X_2 + \cdots + B_kX_k = B_0 + \sum B_jX_{ij} \qquad (4)$$

where

B_0 is the intercept of the linear regression model, and

B_j is the coefficient for jth variable, X_j.

Then, the probability of the ith observation to be in a group as opposed to a reference group based on a nonlinear function of the best linear combination of independent variables is

$$\widehat{Y_i} = \frac{e^u}{1+e^u}. \qquad (5)$$

Observe that by some simple mathematical manipulations the regression equation $u = B_0 + \sum B_jX_{ij}$ can be represented by the natural log of the probability of the odds ratio being in one group versus another reference group such as,

$$\ln\left(\frac{\hat{Y}}{1-\hat{Y}}\right) = B_0 + \sum B_jX_{ij}. \qquad (6)$$

The default value of some statistical programs is 0.5 as a cut point to decide membership of an observation, so that if the logit is 0.5 or higher, the observation belongs to the group. The cut point can be set at another value as well (Soureshani et al., 2013).

Classification and Regression Trees (CART)

A more recent method which is an alternative to model-based approaches is classification and regression trees (CART) (Williams, 1999). It is nonparametric since there are no assumptions regarding observations' distributions. While CART produces decision trees for classification, both continuous and categorical variables can be used as dependent variables. However, it should be noted since this research study is focused on group membership and classification, only the case when dependent variable is categorical will be considered.

The mechanism of CART works through iterative division of data which classifies objects into more homogenous groups known as nodes in the CART terminology. The algorithm of CART starts with locating all subjects into one node, then placing them into other nodes based on creating the most homogenous groups by using predictor variables (Breiman et al., 1984). This process continues until an optimal split of the groups reaches a desirable level of homogeneity of groups based on group membership. To evaluate this mathematically, we minimize deviances in the nodes, and each deviance in a node is calculated as

$$D_i = -2 \sum \sum n_{ik} ln(p_{ik}) \qquad (7)$$

where,

D_i is the deviance of the ith node,

n_{ik} is the number of the subjects from group k in node i,

p_{ik} proportion of subjects from group k in node i.

After calculating deviances of each group, their sum, $D = \sum D_i$, is used as the measure of homogeneity where smaller Ds indicate better homogeneity. The process lasts till reduction in Ds from one step to another becomes negligible, or when the criterion for stopping iterations is satisfied.

The process of CART is represented in Figure 2 from Berk (2016). Here, all the observations first go to a root node. Then, the X values are divided into two based on criterion that X values are compared with a value (c_1) where the cases $X > c_1$ go to right and the cases $X \leq c_1$ go to left. The observations on the left are assigned to terminal node 1, and no improvements in fit can be found for them. On the other hand, the observations on the right go to an internal node and they are divided again based on the criterion if $Z > c_2$ and the procedure follows the same pattern. While this is an illustration for two steps with one variable, more complex versions are possible.

While CART has been addressed as an effective classification method with is variations (Holden et al., 2011; Kohavi, 1995; Witten et al., 1999; Quinlan, 1993), it may show a tendency to favor more distinct predictor variables with fewer values or it may create terminal nodes that overfit with observed data (Berk, 2016). Several models such as random forests (RF) (Horthorn et al., 2006) and Bagging (LeBlanck & Tibrishani, 1996) were created to be alternatives to CART to address such problems.

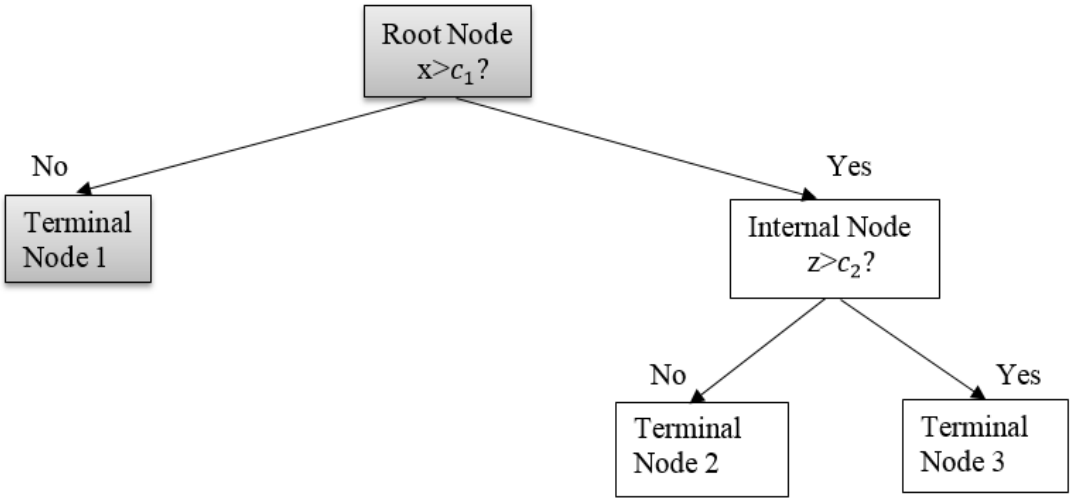

Figure 2

Presentation of a Simple CART Process

Similarities and Differences between LDA, LR, and CART

First, it should be noted that unlike LDA and LR, MFAL cannot be conducted with continuous predictor variables, as it only works with categorical data. Although the LDA and LR methods look the same, as both use the logit ratio of posterior probabilities, a difference arises from the way these techniques estimate coefficients; i.e., the essential difference is how the linear functions fit the data. Moreover, logistic regression is more general and makes almost no assumptions.

When introducing LDA and LR in the book *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Hastie et al. (2013) present them as linear methods for classification. The reason they call these methods linear is because the boundaries of the classes are determined to be linear. Yet, it should be noted that these models can be expanded to their nonlinear versions for classification by adding squares and cross-products of the predictor variables. It is also worth noting that since a predictor

can take values in a discrete set, the input space can be divided into a collection of regions labeled by classification. Thus, class boundaries are the key elements of the classification decisions. Moreover, observations far from decision boundaries play a role in estimating the common covariance matrix, which implies that LDA is not robust to gross outliers (Hastie et al., 2013). On the other side, CART is a nonparametric technique based on decision tree learning logic which provides either classification or regression trees based whether the outcome variable is categorical or continuous. Moreover, it may have stronger resistance to outliers (Timofeev, 2004). Therefore, while all the tree methods differ from each other, CART is by far different from LDA and LR due to the methodology it applies.

Ultimately, Hastie et al. (2013) underscore the difficulty of meeting assumptions in the practice and common use of qualitative variables. They suggest that logistic regression might be a safer choice and more robust than LDA, as well as having fewer assumptions. On the other hand, based on their experience, they mention that both models have very similar performance on classification accuracy in general, even when LDA assumptions are violated. Moreover, CART also requires very few assumptions and performs effectively (Phelps & Merkle, 2008).

Based on the use of variables, LDA and LR can be categorized in three ways. When all the predictors are included in the analyses at the same time, they are defined as direct LR or LDA, but when the order of the variables is specified, they are sequential LDA or LR. Finally, if there is a desire to reduce the number of independent variables but there is no preference on highlighting particular variables, stepwise LR or LDA can be applied for removing some variables by using statistical criteria. On the other hand, the mechanics of

CART does not require selection of variables in advance and it self-determines the important variables and places them into analyses for classification or regression.

Table 1 provides comparisons between LDA, LR and CART in terms of their assumptions and requirements for type of predictor and outcome variables and, decision rules for classification.

Table 1.

Comparison Between LDA, LR and CART

	LDA	**LR**	**CART**
Assumptions	Normality, Absence of outliers, HOCV, Linearity, Absence of Multicollinearity and Singularity, Independence of observations	No Assumptions (minimal sample size requirement)	No Assumptions
Predictor Variables	Continuous	Continuous, Categorical	Continuous, Categorical
Outcome Variable	Categorical	Categorical	Continuous, Categorical
Decision Rule	Highest Group Score	Cut Score (probability, generally 0.5)	Homogeneity through total deviance

Examples of the Application of LDA, LR, and CART

Classification methods have been applied in various areas including social and physical sciences (Arabie & Soete, 1996). Due to the important nature of classification, many applied researchers wish to determine the importance of variables for different groups of observations as outcome variable or they want to be able to predict membership of observed or non-observed cases. For example, in educational research studies risk levels for kindergarten future reading difficulties (Catts et al., 2001), students' learning

disability status (Cook et al., 2015; Dunn, 2007; Keogh, 2005; Lillvist, 2010; Mammarella et al., 2010;), their preferences on instruction types (Clayton et al., 2010) or career choices (Russell, 2008), identification of individuals on the basis of language impairment (Kapantzoglou et al., 2012), or decisions regarding admissions to academic programs (Remus & Wong, 1982) are some of the topic of interest in which LDA, LR, or CART are applied. In psychology and related fields these methods were applied for identification of individuals with psychiatric diagnoses (Zigler & Philips, 1961) or anxiety disorders (Clark et al., 1994); in behavioral sciences, applications have been to study individuals' risks for addiction (Flowers & Robinson, 2002), tobacco consumption (Lei et al., 2015), or prediction of whether male juvenile offenders commit crimes (Glaser et al., 2002) and more.

To give examples from different areas: in health sciences, identification of patients with chronic health failure (Udris et al., 2001) or lung cancer (Phillips et al., 2003) and evaluation of different diagnoses of Alzheimer's patients (Rodriguez et al., 1998); in finance, predicting bankruptcy (Jo et al., 1997), in astronomy, classifying stars (Bidelman, 1957), or in zoology, identification of new species or animals (Britzke et al., 2011) are the topics which have benefited from the use of LDA, LR, or CART. To conclude, it should be noted that the use of LDA, LR, and CART are not limited to the disciplines or topics listed above and the methods can be applied in countless areas.

Comparison Studies of Classification Methods' Performances

Over the decades, there have been studies comparing classification methods, especially LR and LDA. This is because it is important for researchers to be able to choose the optimal method for their studies, especially when the methods' purposes are

the same but with different assumptions (Pohar et al., 2004). However, the discussion of the optimal choice of classification method continues, since there are many specific conditions to explore and some methods perform better than others under certain conditions. To add another dimension, statisticians continue to invent new classification techniques. Hence, over a variety of research areas there can be many different conditions of data such as sample size, group size ratio, or predictor variables' distributions. These different data conditions may impinge on the effectiveness of the classification methods and when the methods' optimal performances occur. Thus, which conditions have significant influence on the classification accuracy of which methods should be explored carefully.

In general, comparison studies between the classification methods has fallen into two different methodologies: comparison studies with "real" data and comparison studies with simulated data. An exception is that by using mathematical techniques and an approximation approach on angles, discrimination boundaries, and key formulas of LR and discriminant function, Efron (1975) states that under multivariate normality and homogenous variances LR performs better than LDA by evaluating expected error rates (Efron, 1975).

Some results of the research from these comparisons along with general definitions are presented below.

Comparison Studies with Real Data.

While most of the studies using real data have a focus on understanding the nature of the classification for the problem in which they are interested, some apply several

classification techniques at the same time and provide a comparison between results of the methods. Comparison with real data is a common way to compare performance of the methods in statistics and related applied areas by using collected empirical data. For example, Dattalo (1995), Meshbane and Morris (1996), and Ferrer and Wang (1999) used real data to compare results of LR and LDA and their focus was on performance of the methods, these studies reported comparable performances of LDA and LR. On the other hand, while focusing on predicting coronary heart disease (Kurt et al., 2008), predicting species distributions (Manel et al., 1999), prediction of dementia (Maroco et al., 2011), or prediction of cardiovascular risk (Colombet, 2000), the researchers used real data and applied at least two of the methods to reach more precise results and recommendations for better methods.

It should be noted that while the studies with real data that focus on the content provide some comparison between methods, and create suggestions for optimal methods for their topics, they are limited to the data they used, and their ability to control data conditions such as sample size, predictor distributions, or effect size is absent. Therefore, the results of studies with a focus on performance evaluation of the methods is reported here since they are not limited to any content area.

Simulation Studies Comparing LDA, LR, and CART.

Simulation studies are also commonly used to compare statistical methods which have the same purpose for analysis, such as the classification methods LDA, LR, and CART. It is important to note that simulation studies have become more common over the past decades for comparison since with them researchers have the ability to

manipulate or control the data, so the evaluation of performance of the methods can be evaluated under different conditions. In simulation studies, data are generated based on specified controlled conditions such as sample size, equality of variance, or number and strength of relationship of predictor variables. Thus, whichever conditions are controlled, researchers can evaluate if the specific data conditions and their interactions affect the performance of the methods, as well as observing which method or methods perform better under certain scenarios. The uncontrolled conditions of the data are assumed to be random.

Many simulation studies use the Monte Carlo technique to simulate data. Particularly, comparison studies between classification techniques including LR, LDA, and CART have applied this technique. The results from simulation studies and real data studies focusing on evaluation of methods' performances based on controlled variables such sample size, group size, and other conditions are presented in the following sections.

Results from Existing Comparison Studies

In this section, results of comparison studies of the performances of LDA, LDA, and CART are reported. The results can be introduced into two main groups: overall performance of the methods and performance of the methods under controlled conditions. While overall performance of the methods are based on simulated data or real data, most of the studies for the performance of the methods were reported from simulated data. However, a few studies using real data also reported performance of the methods for certain conditions.

Comparison studies' results for overall performance of the methods.

Before discussing effect of conditions such sample size or homogeneity of variance on the performance of the classification methods including LDA, LR, and CART, results for the overall performance of the methods from comparison results are summarized.

First, it should be noted that the comparison studies for performance of classification methods include some conflicting results for the overall performance of the methods and their performance under certain data conditions. One reason for this might be due to the fact while some studies use real data, some others used simulated data and the data were from various disciplines. Moreover, while simulation studies have more flexibility and power to manipulate data conditions, it is still difficult to control and report many conditions of data structure at the same time. Finally, the methods and procedures of simulating data and the design of analyses of the studies may differ, so that might also lead to results that are inconsistent.

When comparing the overall performance of LDA with LR, some results showed that LR has higher prediction accuracy for group membership (Baron, 1991), while other results found little or no difference between the two methods (Dey & Astin, 1993; Hess et al., 2001; Meshbane & Morris, 1996). Some studies showed that the statistical methods LDA and LR have performance comparable to CART (Dudoit et al., 2002; Ripley, 1994), but others showed LDA and LR performed better than CART (Williams, 1999) or CART performed better than LR and LDA (Holden, 2011). On the other hand, some results which showed better performance of LDA (Preatoni et al., 2005) or LR (Arminger et al., 1997) than CART are also available in the literature. Finally, some results also showed

better performance of CART than LDA for group membership prediction accuracy (Grassi et al., 2001) while some other results showed similar performance of LR and CART (Schumacher et al., 1996). Thus, the superiority of overall performance of any one method is unclear without consideration of the specific nature of the data.

Comparison studies' results under certain conditions.

In applied areas, data structure might take different conditions such small or large sample size, number of variables, or predictor variables' distributions. While knowing overall performance of the methods helps practitioners decide which methods to apply in their studies, it is critical to evaluate performance of the methods under controlled data conditions. For example, it is possible that a method could perform better than other methods in terms of overall classification accuracy, but it might show poor performance with small sample sizes or in the presence of multicollinearity. Therefore, overall performance of the methods is not enough to make decisions about the optimal method.

Some previous studies evaluated classification methods including LDA, LR, and CART performance under controlled data conditions and the results are reported below.

Sample size.

Sample size is one of the most commonly used conditions in comparison studies for statistical techniques. It refers to the number of observations collected/simulated for a study. In research studies which include quantitative data, there is a common understanding that smaller sample sizes may provide inaccurate results while a very large sample size may not be needed to obtain reliable results (Zavorka & Perrett, 2013). To be able to judge the efficiency of sample size for different statistical analysis techniques, there are sample size calculation methods for finding the desired statistical power.

However, collecting data with the optimum sample size might be challenging due to time, financial, and measurement considerations (Maas & Hox, 2005). Thus, the effects of sample size for the classification methods should be examined carefully, and limitations and consequences regarding it should be explored.

Some studies that have investigated the effects of sample size on the performance of classification techniques agreed that sample size has a significant impact on the accuracy of classification methods (Bolin & Finch, 2014; Holden et al., 2011; Finch et al., 2014; Holden & Kelley, 2010; Pai et al., 2012a, 2012b; Pohar et al., 2004; So, 2003). On the other hand, some studies found that sample size is not a significant factor on the performance of classification methods (Fan & Wang, 1999; Lei & Koehly, 2003). As this creates a conflict in the literature, some researchers whose results did not reach significance claim that this might be due to the limitations of a study, such as not including a small enough sample size (Lei & Koehly, 2003) or not having a varying number of sample size levels (So, 2003). LDA and LR, in general, have lower misclassification rates for larger samples and higher misclassification rates for smaller samples (Finch et al., 2014; Holden et al., 2011; Holden & Kelley, 2010; Lei & Koehly, 2003; Pohar et al., 2004). Yet, in some studies the smallest sample size was not the case of the lowest classification accuracy for LR (Finch et al., 2014; Pai et al., 2012). On the other hand, while Bolin and Finch (2014) reported the reverse, some studies showed that higher sample sizes lead to higher misclassification rates for CART. This might be due to the fact that the studies included different sample sizes.

In spite of the fact that increasing sample size also increases classification accuracy of LDA and LR in general, there are some other factors that significantly

influence sample size effects on the performance of LDA, LR, and CART such as model complexity, group size ratio, and effect size (Holden et al., 2011; Bolin & Finch, 2014). According to Holden et al. (2011), when sample size and effect size increase, the performance of LDA and LR become more similar. On the other hand, they also concluded that the classification accuracy of both classification methods was poor when the group sizes for outcome variables were close to each other under different sample size scenarios. Yet, when discrepancy for group size ratios was greater, LR performed better than LDA in most instances except in some cases when sample size was at the highest level of the study (1000), in which case LDA performed slightly better. While Fan and Wang (1999) state that LDA is more sensitive to sample size than LR, Pohar et al. (2004) made the comment that when the assumptions for LDA are satisfied, it performs better than LR in almost all different possible sample size levels and other conditions. On the other hand, while performance of CART diminished at larger sample size conditions, it was still the best performing method across different sample sizes (Finch et al., 2014; Holden et al., 2011)

Table 2 lists studies which have compared classification techniques including LDA, LR, and CART with a sample size condition.

Table 2.

Comparison Studies for Sample Size

Study	Methods	Sizes
Bolin & Finch (2014)	LR, LDA, CART, QDA, GAM, NNET, RF, MIXDA	150, 1500
Fan & Wang (1999)	LR, PDA	60, 100, 200, 400
Finch et al. (2014)	LR, LDA, CART, GAM, MDA	150, 300, 750; 100, 200, 500
Harrell & Lee (1985)	LR, LDA	50, 130
Holden & Kelley (2010)	LDA, QDA, FFM	100, 1000
Holden et al. (2011)	LR, LDA, CART, QDA, MDA, NNET, GAM, MARS, BOOST	100, 200, 500, 1000
Lei & Koehly (2003)	LR, LDA	100, 400
Pai et al. 2012(a)	LR, MDA, NNET, KNN, LP, HM	100, 200, 400, 500
Pai et al. 2012(b)	LR, DA, MP, HM, NNET, KNN, INT	100, 200, 400, 500
Pohar et al. (2004)	LR, LDA	40, 60, 100, 200, 1000
So (2002)	LR, LDA, LPM, KM	200, 400

Group size ratio, prior probabilities, cut score, and sample representativeness.

Group size ratio.

Group size ratio refers to the proportion of the group's sample size within the outcome group membership variable. In application, group sizes are not generally equal, and cases when groups are somewhat balanced or very imbalanced are more common. In terms of efficiency and accuracy of classification methods, researchers studied the effect of group size ratios, mostly when the ratios were 50:50, 75:25, or 90:10. In this notation,

50:50 refers to both of the groups having 50% of the total sample and 75:25 refers to one group having 75%, the other group having 25% of the total sample.

Group size ratio, that is greater imbalance in group sizes, has a strong impact on the classification accuracy of group membership techniques (Bolin & Finch, 2014, Holden et al., 2011) and it is a significant source of variation in error rates (Finch & Schneider, 2006). Comparing performances of LDA, LR, CART, or other classification techniques in terms of group size ratio, different types of interactions can be observed, so researchers should clarify if their research interest is based on the smaller group, larger group, or total sample classification accuracies (Lei & Koehly, 2003). In general, having greater inequalities between dependent variable group sizes (for example the case of 10:90 versus 25:75) leads to lower overall misclassification rates (Breckenridge, 2000; Craen et al.,2006; Finch & Schneider, 2006; Holden et al., 2011; Holden & Kelley, 2010; Lei & Koehly, 2003). However, greater inequalities in group sizes might have different effects for the different groups. For instance, increasing the group size ratio might increase classification accuracy for the whole sample and for a larger group while it may lead to lower classification accuracy for a smaller group (Bolin & Finch, 2014). In fact, when increasing group size ratio, the classification methods do not misclassify the groups equally and show tendencies to classify in favor of the larger group, although increasing model complexity reduces misclassification rates of both small and large groups (Holden et al., 2011). On the other hand, when the groups are somewhat balanced, misclassification rates for the smaller group are low (Finch & Schneider, 2007).

According to some researchers, group size ratio has a significant interaction with sample size (Bolin & Finch, 2014; Holden et al., 2011), model complexity, effect size

(Holden et al., 2011), and variance ratio (Finch & Schneider, 2007) in terms of classification accuracy for some statistical classification techniques. For example, increasing inequality in a group's variance results in worse performance of LDA and LR and better performance of CART. However, when group variances are highly disproportional and group sizes are balanced, the smaller group classification accuracy is higher for the methods (Finch & Schneider, 2006). Moreover, when group sizes are unbalanced and effect sizes are high, the smaller group classification accuracy does not change very much for LR and LDA (Finch & Schneider, 2007). Finally, when group sizes are disproportional, the effect of the factors sample size, effect size, and covariance matrix ratios are minimal for classification accuracy of the larger group (Finch & Schneider, 2006).

In general, when the group sizes are highly imbalanced, classification accuracy for the smaller group is very low (Holden et al., 2011). On the other hand, when the group sizes are balanced, performances of the classification methods were not highly affected by the variation in sample size. Moreover, LDA showed better performance than LR in balanced situations, while CART was the best performing method regardless of whether group sizes were balanced or not (Bolin & Finch, 2014). Furthermore, according to Holden et al. (2010), LR generally performs better than LDA in different group size ratio and model complexity scenarios. On the other hand, Ferrer and Wang (1999) state that the superiority of logistic regression to discriminant analysis was not impressive in their study (Ferrer & Wang, 1999).

Prior probabilities and cut score.

Prior probability is a concept from Bayesian theory and is related to an event occurring before the collection of data (Nicholson, 2014). In the concept of classification, it should be understood as prior information about the likelihood of a random person to be a member of a specific group; i.e., the proportion of members within the group in the true population (Lei & Koehly, 2003). Therefore, it differs from group size ratio conceptually even though both concepts frequently refer to similar information. Unless specified, some statistical packages such as SAS and SPSS use the default settings for prior probabilities and cut score. As an example, for a two-group case, default prior probabilities are 50:50 and the cut score is 0.5. However, the more general practice for specification of prior probabilities is using the group size ratios gathered from sample size ratios of the groups or the population's group size ratios (Ferrer & Wang, 1999).

Some researchers say that it is important to consider prior probabilities and specification of *a priori* selection of classes when evaluating the performance of LR and parametric classification methods (Fan & Wang, 1998; Huberty, 1994; Press & Wilson, 1978; Wilson & Hargrave, 1995). Ferrer and Wang (1999) showed that prior probabilities explain an important amount of variation for error rates when using group size ratios as the estimate of prior probabilities.

According to Lei and Koehly (2003), there is a significant interaction between cut-score and prior probabilities for accuracy of performance of LDA and LR. Here, cut score refers to a decision rule based on probability; i.e., what probability should be the rule of thumb for group membership in LR? Even though the default is 0.5, one has the option to assign an observation to a group with a different probability. Their results

showed that LDA with priors specified based on sample sizes of the groups from the sample performs better than LR and LDA without specification of priors (i.e., using default settings). Moreover, LR performs slightly better than LDA without the specifications. For optimal performance of classification, assuming that the sample is representative of the population, they suggest using LDA with proportional priors or LR with a cut score 0.5 when the interest is in reducing total misclassification. Similarly, for the case when the concern is reducing large group misclassification accuracy, their suggestion is using LDA with prior probability specification for extreme inequalities of the group sizes such as 10:90 and a cut score of 0.5 for LR regardless of any other conditions. Finally, the suggestion for small group accuracy is using a cut score 0.1 for LR and LDA with equal prior probability specifications regardless of any other conditions such as true prior probabilities and variance ratios.

While many research studies focus only on classification accuracy of groups of the dependent variable, it is also possible that these groups include subgroups. Therefore, subgroup sizes may have effect on classification performances of the methods. For example, in a study regarding reading disability, the groups for dependent variable could be having a reading disability and not having a reading disability. Moreover, these groups could also have unknown subgroups based on degree of disability or other conditions which may not be known. Due to the limitation of not being able to categorize the subgroups exactly, it is also possible that the subgroups overlap. According to Finch et al. (2014), increasing level of overlap between subgroups leads to higher misclassification rates of the methods including LR, LDA, and CART in general. CART showed the best performance under heterogeneous groups and subgroup overlap.

Sample representativeness.

Besides the effects of group size ratio and prior probabilities, So (2003) also studied the effect of sample representativeness which measures how well the sample represents the population in terms of prior probabilities of the population groups. For example, for a population of groups with prior probabilities 10:90, if the sample has the group size ratio 20:80, it means that the sample is over representative (over-sampled) for the smaller group and under representative (under-sampled) for the larger group. The study showed that sample representativeness is a significant factor for the classification accuracies of classification methods including LDA and LR and significantly interact with prior probabilities. On the other hand, the effect of sample representativeness is negligible when the data hold the condition of equal prior probabilities of population.

Table 3 is presented below to summarize some studies which have compared classification techniques including LDA, LR, and CART for group size ratio, prior probabilities, cut score and sample representativeness.

Table 3.

Comparison Studies for Group Size Ratio, Prior Probabilities, Cut Score or Sample Representativeness

Studies	Methods	Conditions	GSR
Bolin & Finch (2014)	LR, LDA, CART, QDA, GAM, NNET, MIXDA, RF	PM, SS, GSR, ES	50:50:50, 25:25:100
Fan & Wang (1998)	LR, PDA	SS, PP, HOCV	50:50, 75:25, 90:10
Ferrer &Wang (1999)	LR, PDA, NPDA	GSR, PD, HOCV	50:50, 75:25, 90:10
Finch et al. (2014)	LR, LDA, CART, GAM, MDA	SS, GSR, SubS, SubR	50:50, 75:25
Finch & Schneider (2007)	LR, LDA, CART, QDA, NNET	GSR, ES, HOCV, NPV, PD	111,211,221 11111,21111, 22221
Finch & Schneider (2006)	LR, LDA, CART, QDA,	SS, GSR, ES, PD, HOCV	50:50, 75:25, 90:10
Holden et al. (2011)	LR, LDA, CART, QDA, MDA, NNET, GAM, MARS, BOOST	SS, ES, GSR, MC	50:50, 75:25, 90:10
Lei & Koehly (2003)	LR, LDA	HOV, GSR, SS, PP	50:50, 75:25, 90:10
So (2003)	LR, LDA, LPM, K-MEAN	SS, PP, HOCV, GS, SRep	50:50, 75:25, 90:10

Note: ES: Effect Size, GSR: Group Size Ratio, HOCV: Homogeneity of Variance-Covariance Matrices, MC: Model Complexity, NPV: Number of Predictor Variables, PD: Predictors' Distributions, PM: Percent Misclassified, PP: Prior Probabilities, SubR: Subgroup Ratio, SRep: Sample Representiveness, SS: Sample Size, SubS: Subgroup Separation

Predictors' distributions: Normality versus non-normality.

In parametric statistical techniques, normality is a required assumption to ensure reliable results (Ghasemi & Zahediasl, 2012). It refers to the distributional property of outcome or predictor variables such as symmetry and inclusion of a proportion of the observations within the determined standard deviations around the mean based on an empirical rule. Even though some studies state that certain statistical techniques are

robust against violation of normality, it is still one of the most common assumptions to check before conducting parametric statistical analyses. Moreover, violation of normality may cause biased classifications which lead to poor performances of classification techniques (Eisenbeis, 1977; Kiang, 2003).

According to some researchers, normality has a significant effect on the performance of LDA (Pohar et al., 2012) and LR (Pai et al., 2012) and LR was superior to LDA when the normality assumption was not satisfied (Kiang, 2003). However, when the assumptions of LDA, normality, and HOCV were satisfied, the two methods showed similar performances. Additionally, when predictor variables were normally distributed, only violation of HOCV slightly affected the accuracies of the methods. Moreover, when data were skewed, violation of HOCV at higher degrees improved performances of LDA, LR, and CART (Finch & Schneider, 2006).

In some comparison studies for analyzing the accuracy of classification techniques, to be able to observe the effect of normality or non-normality, researchers created limited cases of non-normal data, where predictor variables were skewed or lognormally distributed. These non-normal cases were then compared to the cases when data were normal. However, it should be noted that non-normality is a broad situation and there can be many different degrees of non-normality.

According to Pohar et al. (2004), the cases when skewness is somewhat ignorable at values about ±0.2, LDA performs better, but when level of skewness is increased, LR tends to perform better. In general, LDA performed better when all the predictors were normal while LR was better suited to many different types of distributions (Baron, 1991; Cox, 1989). On the other hand, when only one of two variables was normal and the other

was skewed or asymmetric, LDA was robust; i.e., inclusion of even one normal variable might increase the resistance of LDA against sensitivity to non-normality of other variables. Under the condition when data were skewed and kurtosis was small, both LDA and LR performed at an optimal level. Moreover, under non-normality with a large sample size, performance of both LDA and LR became more similar (Rausch & Kelley, 2009). However, increasing kurtosis also increased the performance of classification in favor of LDA, which contradicts other presented studies in which LDA and LR performed differently under conditions of non-normality.

Even though categorical predictor variables are not preferred for LDA, when all predictor variables were categorical, LR, LDA, and CART showed similar performances for overall misclassification rates and smaller group misclassification rates (Finch, Schneider, 2006). Moreover, when prior probabilities are not equal, LR is expected to perform better than LDA under conditions of non-normality and extreme cases (Dattalo, 1994; Ferrer & Wang, 1999; Hosmer 1989; Huberty, 1999). On the other hand, CART performed worse than LDA and LR when the distribution was either normal or skewed except in the cases when covariance ratios were highly disproportional (Finch & Schneider, 2006). Finally, in the discussion for normality, Ashikaga and Chang (1981) argue that *similarity* of population shapes plays a more important role than *normality* of predictors when assessing the performance of classification techniques.

As an extended version of normality, multivariate normality is also an important property of data to be able to make precise parametric statistical estimates. Particularly, in many multivariate statistical techniques, multivariate normality is assumed, but meeting that assumption is even more difficult than meeting the assumption of univariate

normality. For instance, if all the variables' distributions are not normal, then multivariate normality is not possible. Harrel and Lee (1985) reported that when multivariate normality holds, there is little difference between LDA and LR.

Table 4 is presented below to list some studies which have compared classification techniques including LDA, LR, and CART for predictor variables' distributions.

Table 4.

Comparison Studies for Predictor Variables' Distributions

Authors	**Methods**	**Conditions**
Finch, Schneider (2007)	LR, LDA, CART, QDA, NNET	PD, ES, GN, HOCV, GSR, NPV
Finch & Schneider (2006)	LR, LDA, CART, QDA	PD, HOCV, ES, SS, GSR
Ferrer & Wang (1999)	LR, PDA, NPDA	PD, GSR, HOCV
Harrel and Lee (1985)	LR, LDA	PD, PP, GS
Pai et al. (2012a)	LR, MDA, NNET, KNN, LP, HM	PD, SS, CORR, DD, SP, Lin, Out, HS
Pai et al (2012b)	LR, DA, MP, HM, NNET, KNN, INT	PD, DD, Out, HS, GSR, SS, GSR
Pohar et al. (2004)	LR, LDA	PD, SS, CORR, DBGM, MAHD, GN
Rausch & Kelley (2009)	LR, LDA, LDR, MDA	PD, GSSRNP, GS, PD

Note: CORR: Correlation, DBGM: Distance Between Group Means, DD: Dynamic Data, ES: Effect Size, GN: Numbers of the Groups in the Outcome Variable, GS: Group Separation, GSR: Group Size Ratio, GSSRNP: Group Sample Size to the Number of Predictors, HOCV: Homogeneity of Variance, HS: Homoscedasticity, Lin: Linearity, MAHD: Mahalanobis Distance, NPV: Number of Predictor Variables, PD: Predictors' Distributions, PP: Prior Probabilities, Out: Outliers, SS: Sample Size

Effect size.

Effect size is a measure of the quantified difference between groups or degree of relationship between variables. Conceptually, many indices such as standardized mean difference, Cohen's D, Hedge's G, η^2, or a correlation coefficient can be used to report effect size. In the studies which compare classification methods, effect size is also used for evaluating group separation; i.e., degree of group mean differences and standardized mean differences between groups were used as the index of group separation while some researcher evaluated group separation with different formulas.

Some previous research showed that when the group means of the outcome variable were widely separated (large effect size), LDA, LR, and CART showed higher classification accuracies (Bolin &Finch, 2014; Finch & Schneider, 2007; Holden et al., 2011). Holden and Kelly (2010) also reported a similar result only for LDA as they did not include LR and CART in their study. Moreover, at high levels of effect size, increasing the sample size ratio results in lower classification accuracy for LDA and misclassification in LDA occurs in favor of the larger groups (Holden & Kelly, 2010). Furthermore, when normality and HOV assumptions for LDA were satisfied, LDA and LR showed similar results across different effect sizes while CART performed slightly better. On the other hand, when violating HOV, it was observed that the methods' performances of classification accuracy for LDA and LR declined faster at large effect sizes than when effect sizes were smaller, but CART showed improved performance in the same scenario (Finch & Schneider, 2007).

According to some researchers, there are significant interactions of effect size and group size ratio (Finch & Schneider, 2007; Holden et al., 2011), predictors' distributions,

number of the predictor variables, model complexity and sample size (Holden et al., 2011), and variance ratio for the classification accuracy of classification methods (Finch & Schneider, 2007). Moreover, while Finch and Schneider (2007) found that effect size is the highest source of variation for the classification accuracy of the methods they studied among all the other variables in their study, another study (Finch & Schneider, 2006) reports that effect size has a relatively small effect on the classification methods. Additionally, in general, increasing effect size leads to smaller misclassification rates, but increasing effect size decreased the misclassification rate only about 2-3 % and was not affected by the other manipulated variables. On the other hand, effect size has an important impact on smaller group classification accuracy in all group size ratio conditions while misclassification rates of smaller groups for LDA, LR, and CART improved when groups had higher effect sizes. Moreover, CART showed the highest classification accuracy for the larger group and smaller groups at different effect size levels. Furthermore, when effect size was small for unequal group sizes, a very high percentage of observations from the smaller group were misclassified by LDA and LR, but CART showed a better resistance in this case and had higher classification ratios for smaller groups (Holden et al., 2011). On the other hand, large group classification accuracies were high in general regardless of level of effect size (Finch & Schneider, 2006).

When effect size is large the classification methods have a tendency to predict group membership in favor of the larger group in the outcome variable regardless of true group membership (Finch & Schneider, 2006; Holden et al., 2011) and performance difference between LDA and LR become trivial for overall classification when the effect

size is large (Holden et al., 2011). Therefore, if the purpose is prediction of smaller groups, this result should be considered carefully.

Table 5 is presented below to list some studies which have compared classification techniques including LDA, LR, and CART for effect size.

Table 5.

Comparison Studies for Effect Size

Authors	Methods	Conditions	Effect Sizes
Bolin &Finch (2014)	LR, LDA, CART, QDA, GAM, NNET, MIXDA, RF	PM, SS, GSR, ES	.2, .5, .8, 1.6
Finch & Schneider (2007)	LR, LDA, CART, NNET, QDA,	ES, GN, HOCV, GSR, NPV, PD	.2, .8
Finch & Schneider (2006)	LR, LDA, CART, QDA,	PD, CI, ES, SS, GSR, NPV	.2, .5, .8
Holden & Kelley (2010)	LDA, QDA, FFM	HOCV, PM, SS, SSR, ES	.2, .5, .8, 1.6
Holden et al. (2011)	LR, LDA, CART, QDA, MDA, NNET, GAM, MARS, BOOST	SS, ES, GSR, MC	.2, .5, .8, 1.6

Note: ES: Effect Size, GN: Numbers of the Groups in the Outcome Variable, GSR: Group Size Ratio, HOCV: Homogeneity of Variance, NPV: Number of Predictor Variables, PD: Predictors' Distributions, PM: Percent Misclassified PP: Prior Probabilities, SS: Sample Size

Homogeneity of variance-covariance matrices.

Variance is one of the important statistical learning tools to evaluate distributions of variables and homogeneity of variance-covariance matrices (HOCV) indicates that distribution of observations from different groups have similar degrees of distance from their group means. As previously stated, LDA requires HOCV matrices, but some other classification techniques in general also assume HOCV to yield better analyses (Johnson & Wichern, 1988; Tabachnick & Fidell, 2013).

Some procedures such as Levene's homogeneity of variance test, Bartlett's test or its multivariate version, Box's M test, can be applied to evaluate homogeneity of variance-covariance matricies (Tatsuoka, 1988) though Box's M is affected by non-normality (Huberty, 1994; Meshbane & Morris, 1996). Therefore, evaluation of HOCV under non-normal data is difficult (Lei & Koehly, 2003). On the other hand, Mashbane and Morris (1996) did not find that LDA's performance diminishes under violation of HOCV for normal data which implies that LR does not necessarily perform better than LDA under unequal variances. Yet, normality of data is rare in application, so these results should be evaluated with care.

As an alternative to LDA, QDA was suggested to be more robust against violation of HOCV (Anderson, 1984; Huberty, 1994; Huberty, Lowman, 2000; Johnson & Wichern, 1988). While QDA was thought to perform better than LDA, some researchers found that the overall performance of QDA was not better than LDA in any cases of different covariance equality or inequality for total error rates, but QDA was better for individual group error rates (Meshbane & Morris ,1995); however, a preference for LDA over QDA was suggested to practitioners due to theoretical difficulties of QDA (Hess et al., 2001). Particularly, QDA may not be a better technique under covariance inequality when the assumption of normality is not satisfied (Krzanowski, 1977; McClachlan, 1992; Stevens, 1996). Moreover, similar performances of QDA and LDA were reported by Ferrer and Wang (1999) and they claim that might be due to usage of pooled covariance estimations (Ferrer & Wang, 1999). On the other hand, Finch and Schneider (2006) reported that QDA performs better than LDA under violation of HOCV. Finally, such techniques as QDA or FMM (Finite Mixture Model), which assume inequality of

variances, have a disadvantage for class predictions when population variances are indeed homogeneous (Holden & Kelly, 2010).

Heterogeneity or homogeneity of variance-covariance matrices was found to cause a considerable amount of variation in classification error rates of LDA, LR, and CART by some researchers (Fan &Wang, 1998; Finch & Schneider, 2007). Additionally, heterogeneity of the matrices affects the performance of both LDA and LR negatively and the performance difference of LDA and LR becomes greater in favor of LR when variance inequalities between groups are bigger. Yet, that does not necessarily imply that LR performs significantly better than LDA when variance-covariances are unequal. Under violation of HOCV and when groups have highly unbalanced prior probabilities, LDA predicts classes of observations in favor of the larger group, while LR classifies in favor of the smaller group when cut scores for LR and prior probabilities for LDA were not specified, so the default setting were used (Fan &Wang, 1998). However, specification of prior probabilities and cut scores might change the direction of these results. A study by Lei and Koehly (2003) after Fan and Wang (1998) found that LDA and LR both performed better for smaller group classification (when group sizes on the dependent variable were not equal) under unequal variances than in the case when the variances were homogenous. On the other hand, violation of HOCV decreased accuracy of LDA for the larger group as expected. Finally, they found significant interactions of variance-covariance ratios, cut score, and classification methods (LDA, LR) for classification accuracy of larger and smaller groups (when the group sizes were unbalanced). Nonetheless, variance-covariance ratio, cut score, and classification method (LDA, LR) did not yield a significant interaction for total classification accuracy due to

reverse effects of cut score. Therefore, the importance of HOCV may be not necessarily limited to LDA, but also LR (Lei & Koehly, 2003).

While it was expected that LR would perform better than LDA under violation of HOCV, some studies did not find a superiority of LR to LDA for this condition (Fan & Wang, 1999) especially when the data were normally distributed (Finch & Schneider, 2006). On the other hand, according to Kiang (2003), LR is superior to LDA when the assumption of HOCV does not hold. Moreover, Finch and Schneider (2006) found a considerable contribution of variance ratio for the performance of the classification methods. Their results also showed that both LR and LDA show weaker performance under non-equal variances, but CART improved its performance and it was the best performing method under variance inequality. Moreover, when group variances were equal, the larger group had a very small misclassification rate and the smaller group had a very high misclassification rate. On the other hand, similar to their previous study, Finch and Schneider (2007) found that increasing inequality of variances between groups leads to weaker performance of LDA and LR, but variance inequality had a small effect on larger group classification accuracy when group sizes are not balanced (Finch & Schneider, 2006, 2007). Additionally, for the three-group outcome case, both methods had their highest classification accuracies for the group with the highest mean, the group with the lowest mean had the second highest classification accuracy, and the middle group had the lowest classification accuracy for LDA, LR, and CART. They also report that the effect of unequal variances is small on the performance difference of LDA and LR when the data are distributed normally. On the other hand, when the data are not normally distributed, the performance of the methods under variance-covariance matrix

inequalities depends on the type of non-normality. When predictor variables are skewed, an increasing level of variance-covariance inequality leads to better performance of LDA, LR, and CART. Moreover, Hess at al. (2001) found similar results for LDA and LR under non-normal data regardless of whether the variances were homogenous or not (Hess et al., 2001). They also stated that when variances are not homogenous, the performance differences of LDA and LR depends on group size ratios and increasing sample sizes under extreme variance inequality may not improve precision of the methods. Thus, violation of HOCV in general leads to higher misclassification rates of both LDA and LR, and when the effect size was large the performances declined faster than when the effect size was small (Finch & Schneider, 2007). Finally, while some researchers used HOCV as a case of group separation (Fan & Wang 1998), some other researchers evaluated group separation with different concepts such as Mahalanobis distance or effect size.

Table 6 is presented below to list studies which have compared classification techniques including LDA, LR, and CART for HOCV.

Table 6.

Comparison Studies for HOCV

Studies	Methods	Conditions
Fan & Wang (1998)	LR, PDA	SS, PP, HOCV
Ferrer & Wang (1999)	LR, PDA, NDA	PD, GSR, HOCV
Finch & Schneider (2006)	LR, LDA, CART, QDA	PD, HOCV, ES, SS, GSR
Finch & Schneider (2007)	LR, LDA, CART, QDA, NNET	ES, GN, HOCV, GSR, NPV, PD
Hess et al. (2001)	PDA, LR	GS, HOCV, SS, PD
Kiang (2003)	LR, MDA, NNET, DT, KNN	PD, LIN, DYN, CORR, MMOD, HOCV, GSR, SS
Lei & Koehly (2003)	LR, LDA	HOV, GSR, SS, PP

Note: CORR: Correlation, DYN: Dynamic Environment of Data, ES: Effect Size, GN: Numbers of the Groups in the Outcome Variable, GS: Group Separation, GSR: Group Size Ratio, HOCV: Homogeneity of Variance, LIN: Linearity, MMOD: Multimodal Data, NPV: Number of Predictor Variables, PD: Predictors' Distributions, PP: Prior Probabilities, SS: Sample Size

Multicollinearity: Correlation effect.

Correlation defines a degree of linear relationship between variables. Multicollinearity is the case of when some or all predictor variables are highly correlated with each other. Absence of multicollinearity is an assumption for LDA, while LR and CART do not have any specific limitation regarding it. In general, presence of multicollinearity can be assessed by inspection of the correlation matrix of variables or reviewing tolerance or the variance inflation factor (VIF) (Neter et al., 1996). Kiang (2003) reports that low correlation has a moderate effect on the classification performance of LDA and LR. Previous research also showed that presence of multicollinearity significantly increased classification accuracy of LDA, but the performance of LR was not affected by multicollinearity significantly (Pai et al, 2012). This is an interesting result to consider since LDA requires absence of multicollinearity

as an assumption. On the other hand, for parameter estimation, it is known that multicollinearity causes poor performance of statistical techniques (Meyers et al., 2016). Thus, it can be concluded that multicollinearity can have different directions of impact depending on if the purpose is classification or parameter estimation. Even though they did not discuss implications of the results, Pohar et al. (2004) presented a table which implies superiority of LDA under higher correlations. Moreover, Zavroka and Perret (2014) stated that degree of correlations between variables affect the recommended minimum sample size for LDA and QDA. Finally, none of the reviewed studies compared LDA, LR, and CART at the same time for correlation conditions of predictor variables.

Table 7 is presented below to list some studies which have compared classification techniques including LDA and LR for the effect of predictor correlation.

Table 7.

Comparison Studies for Correlation Effect

Studies	**Methods**	**Conditions**
Kiang (2003)	LR, MDA, NNET, DT, KNN	PD, LIN, DYN, CORR, MMOD, HOCV, GSR, SS
Pai et al. (2012a)	LR, MDA, NNET, k-NN, LP, HM	PD, SS, CORR, DYN, GSR, LIN, OUT, HS
Pai et al (2012b)	LR, DA, MP, HM, NNET, KNN, INT	PD, DYN, OUT, HS, GSR, SS
Pohar et al. (2004)	LR, LDA	SS, PD, CORR, DBGM, MAHD, GN
Zavroka & Perret (2014)	LDA, QDA	NPV, CORR, GSR

Note: CORR: Correlation, DBGM: Distance Between Group Means, DYN: Dynamic Environment of Data, GN: Numbers of the Groups in the Outcome Variable, GSR: Group Size Ratio, HOCV: Homogeneity of Variance, HS: Homoscedasticity, LIN: Linearity, MAHD: Mahalanobis Distance, MMOD: Multimodal Data, NPV: Number of Predictor Variables, PD: Predictors' Distributions, SS: Sample Size

Number of predictor variables.

While a larger number of predictor variables has the potential to increase classification accuracy, the cases when variables do not significantly contribute to group differences or when the number of predictor variables is comparable to the number of the subjects in the study might cause decreasing performance of the statistical models (Hosmer & Lemeshow, 1989; Huberty, 1994). Therefore, studies with a greater number of predictor variables typically require larger sample sizes (Zavroka & Perret, 2014). According to McLachlan and Byth (1979), LDA and LDR perform similarly in terms of classification accuracy under the condition when the number of predictor variables is comparable to sample size. However, when the group sample sizes are small relative to the number of predictors, classification methods may tend to provide inaccurate prediction (Rausch & Kelley, 2009).

Most of the comparison studies evaluated performance of LDA, LR, and CART under a fixed number of predictor variables. However, not having varying numbers of predictor variables in a comparison study creates a situation where one cannot analyze the effect of the number of predictor variables on the performance of classification methods, leaving a gap in the literature. On the other hand, it should be noted that the nature of comparison studies becomes complicated when including more controlled conditions, so that it is possible that researchers avoid complexity in their studies by not including number of predictor variables as a condition. Although Finch and Schneider (2007) found that effect size was the main factor in determining correct classification of these methods, they also found a small contribution of the interaction for number of predictor variables, their distributions and variance ratio. Their results also showed that

increasing the number of predictor variables also increased performance of LDA, LR, and CART overall. Moreover, the methods generally showed similar performances for different numbers of predictor variables.

In addition to this finding, a greater impact of the number of predictor variables was found with more rather than fewer groups in the dependent variable. For example, when having three groups in the dependent variable and increasing the number of the predictor variables from three to seven, the methods (LDA, LR, CART) increased their classification accuracies by 2-6%, but when having five groups in the dependent variable and increasing number of the predictor variables from three to seven, LDA and LR increased their classification accuracies about 18% and CART increased its classification accuracy about 10%. It was also notable in that study that LDA and LR showed higher improvement than CART when increasing number of the predictor variables.

Table 8 is presented below to list some studies which have compared classification techniques including LDA, LR, and CART by the number of the predictor variables they included in the study.

Table 8.

Comparison Studies and Number of the Predictor Variables Included in the Study

Number of the Predictor Variables	**Authors and Year**
1	Bolin & Finch (2014); Holden & Kelley (2010); Hess et al. (2001)
2	Pohar et al. (2004); Kiang (2003)
3	Finch &Schneider (2006); Lei & Koehly (2003); Pai et al. (2012a); Pai et al. (2012b)
4	Holden et al. (2011)
5	Finch et al. (2014); Harrell, Lee (1985)
8	Ferrer & Wang (1999); Rausch& Kelley (2009)
2,4	Zavroka & Perret (2014)
3,7	Finch & Schneider (2007)
3,8	Fan & Wang (1998)

Number of groups in the outcome variable.

Most of the studies compared the classification methods evaluated the methods under the case when the outcome variable had just two groups and in almost all the studies which compared performance of classification methods, the effect of number of the groups in the outcome variable was not discussed in detail. However, both LDA and LR can be applied in the case when the outcome variable has more than two groups (Hosmer & Lemeshow, 1989) and number of the groups might have an effect on the

performance of classification methods. In application, the case when the outcome variable has more than two groups is common.

Finch and Schneider (2007) evaluated the cases for three and five groups separately. For almost all the conditions they tested, there was a decrease in classification accuracy of LDA, LR, and CART when increasing the number of groups. On the other hand, there were very small differences between the methods for the different number of groups of the outcome variable under various data conditions. It should be noted that while most of the comparison studies created different levels of conditions such as sample size or group size ratio and included them in the comparison, that was not the case of this study and the effects of number of the groups were evaluated separately. They report an interesting result that the middle groups (in terms of means) had lower classification accuracy. Moreover, they stated that while they studied three and five groups, their results were similar to the two-group case. While the methods showed comparable results for three- and five-group cases, it was noticeable that CART had better classification accuracy than LDA and LDA for the three group (less groups) case and LR and LDA (with similar results) showed better performance than CART for five group case. Finally, based on their results they suggested minimizing the number of groups without disregarding important groups.

In their study with two normally distributed predictor variables, Pohar et al. (2004) also found that the greater the number of groups (categories) in the dependent variable, the lower the prediction accuracy of LDA and LR. They also claim that the effect of categorization might depend on some other data conditions such as correlation

and number of the variables. Finally, their results show that LDA performs better for a larger number of categories while LR is a better option for the binary case.

Table 9 is presented below to list studies which have compared classification techniques including LDA, LR, and CART for the number of groups they included in their study

Table 9.

Number of the Groups in the Comparison Studies

Number of the Groups	**Authors and Year**
2	Fan & Wang (1998); Ferrer & Wang (1999); Finch & Schneider (2006); Harrell & Lee (1985); Hess et al. (2001); Holden et al. (2011); Holden & Kelley (2010); Kiang (2003); Lei & Koehly (2003); Rausch & Kelley (2009)
3	Bolin & Finch (2014)
2 and 3 (separately)	Finch et al. (2014)
3 and 5 (separately)	Finch & Schneider (2007)
4	Pai et al. (2012)-1; Pai et al. (2012)-2; Zavroka & Perret (2014)
2, 3, 4, 5, infinity	Pohar et al. (2004)

Other conditions.

Unlike conditions of sample size, effect size, and predictor variables' distribution which were discussed with some frequency by researchers in the literature, there are some additional conditions which have been tested to compare classification methods including LDA, LR, and CART. Some results regarding these conditions are introduced below.

Linearity.

Linearity is a mathematical property formulated as $f(x) = Ax + B$ in which $f(x)$ is defined as a linear function and x is a variable. The values of $(x, f(x))$ can be represented as a straight line with a random degree of slope in a scatterplot or in the data points somewhat clustering around a line. Therefore, linearity can be understood as level of straightness for the relationship between dependent and independent variables. While it is assumed for LDA, LR and CART do not require linearity as an assumption. Based on limited research, LR was superior to LDA when the linearity assumption was not satisfied (Kiang, 2003). However non-linearity still has a moderate effect on the performance of LR and the performance LDA significantly decreased in the absence of linearity (Kiang, 2003; Pai et al., 2012)

Model complexity.

Model complexity is the condition related to inclusion of the number of variables, their interactions, and nonlinear versions such as quadratic or cubic forms of the variables in the analysis. In general, inclusion of more variables, interactions, and quadratic or cubic forms of the variables increase model complexity. More complex models show greater prediction accuracies, but in the case when some variables do not significantly

contribute to the model's prediction accuracy or the goodness of fit, more parsimonious options should be preferred. Moreover, complex models might lead to some problems such as multicollinearity or singularity. Therefore, model complexity should not be always taken as an advantage.

Holden et al. (2011) stated that, in general, more complex models have lower misclassification rates and the misclassification rates depends on the classification method. They found significant interactions between model complexity, effect size, and group size ratio for the methods' classification accuracies. Moreover, when group sizes were unequal, the large and small group misclassifications were highly dependent on model complexity and group size ratio. Furthermore, when group sizes were unequal and model complexity was increased, both smaller and larger groups had smaller misclassification rates. In their study, which had three levels of model complexity (linear, simple, complex), for linear and simple models and when groups sizes were equal, LR had lower misclassification rates than LDA while for unequal groups, they both showed very similar performance. On the other hand, for the complex model, LR had better performance than LDA regardless of effect size or group size ratio. Finally, in addition to increasing performance when the effect size increased, CART generally showed the highest prediction accuracy among the three methods for different types of model complexities (Holden et al., 2011).

Dynamic environment of data.

Most of the comparison studies made the assumptions of static data, so that there was no change in data values over time. On the other hand, some researchers considered the dynamic nature of real world data should be included in comparison studies for

classification methods' performance. By using some trigonometric functions to generate change in coefficients for variables over time, it was found that a dynamic environment of data decreases performance for some classification methods including LDA and LR (Kiang, 2003; Pai et al., 2012). In the studies which took account of dynamic data, a time series approach was used, but it should be noted that there might different situations with dynamic data rather than just using a sine function which varies from -1 to +1 with a periodic fluctuation.

Outliers.

An outlier can be defined as an observation which is distant from the other observations in the data. In application, the presence of outliers in a dataset is a common situation and it creates some analytical concerns. In statistics, deletion of the outliers or transformation of data with some mathematical formulas are two ways to deal with outliers. Pai et al. (2012) showed that classification accuracies of the techniques are affected by presence of outliers and in line with the LDA assumptions, the performances are decreased.

Multimodal structure of data.

While many statistical models require a unimodal structure of data, multimodal distributions are also possible. Multimodal data is the case when the distribution has more than one peak or modes. Bimodal (2 modes) and trimodal (3 modes) data are types of multimodal data. Kiang (2003) reports that both LDA and LR performed worse under a multimodal data structure and the results indicate LR performed slightly better than LDA (Kiang, 2003). On the other hand, this result should be reviewed with care since both methods had different base error rates.

Percent of initial misclassification.

When some observations are initially misclassified, performance of classification methods might decrease and having misclassified observations is considered a measurement problem (Betebenner et al., 2008; Ozasa, 2008). There are several types of misclassifications such as random, non-random, differential, and non-differential misclassification (Chhikara & McKeon, 1984; Holden & Kelley, 2010; Lachenbruch, 1966; Lachenbruch, 1974; Ozasa, 2008).

LDA might be affected slightly by initial misclassification, as observations which have a mean close to other classes to which they do not belong have a greater chance of misclassification (Holden, 2009; Holden & Kelly, 2010; Lachenbruch, 1966; Lachenbruch 1974; McLachlan, 1972). Bolin and Finch (2014) reported that initial misclassification proportions, group size ratio, and classification methods interacted significantly. Moreover, their results showed that higher initial misclassification rates caused higher misclassification accuracy (prediction) of the methods including LDA, LR, and CART. Finally, in their study CART performed better than LDA and LR while all showed similar patterns in the presence of initially misclassified data by reducing their classification accuracies.

Group separation.

Group separation, like effect size, defines level of separation between two or more groups which is the extent of overlapping levels of populations. While it has mostly been measured by Mahalanobis distance: $D^2 = (\mu_1 - \mu_2)^T \Sigma^{-1}(\mu_1 - \mu_2)$ (Harrell & Lee, 1985, Lei & Koehly, 2003; Rauch & Kelley, 2009; So, 2003), variance ratios (Fan & Wang, 1998) and some other formulas or algorithms also used to quantify separation of

groups. The D^2 values such 6.7 and 2.2 were considered as large separation (Stevens, 1996; Meshbane & Morris, 1996) and 0.7 or below were considered as moderate separation (Huberty, Wisenbaker & Smith 1987).

Some previous research found that increasing Mahalanobis distance (i.e. level of separation between groups) leads to higher classification accuracies for LDA and LR (Fan & Wang, 1999; Harrell & Lee, 1985, Finch et al., 2014; Lei & Koehly, 2003). Moreover, the results also showed that under the conditions when HOCV and normality held, the superiority of LDA to LR disappeared with increasing group separation (Mahalanobis distance). Furthermore, group separation was found to significantly interact with group size ratio, HOCV, prior probabilities, cut score, and smaller and larger group classification accuracies of the methods including LR and LDA (Lei & Koehly, 2003; So, 2003). On the other hand, the performance difference between LDA and LR for smaller and larger groups' classification accuracies was not affected significantly by different degrees of group separation (Harrell & Lee, 1985). Moreover, in their study which included LR, LR, and CART, Finch et al. (2014) found that under the condition when degree of separation for subgroups was high, CART provided more accurate results.

Summary and Research Questions

When applying explanatory models, researchers aim to investigate causal relations between variables, while usage of predictive models such as LDA, LR, and CART are generally targeted to predict categories based on a correlational rather than causal design. Therefore, these models are used to evaluate group discriminations and determinations (Sainani, 2014). For example, by applying predictive models, one can

estimate probability of having a disease based on results of diagnostic tests, or mortality of a veteran with a stroke of a particular level of severity within one year (Bates et al., 2014). Some predictor variables such as positive perception of teacher, GPA, if the student lived with biological parents or not, and number of the days absent from school can be examined to see if they predict high school students' dropout status by applying LR (Suh et al., 2007), as well as the other two. Traditionally, in educational and social science research, LDA and LR are applied widely, and as a newer method, CART is not applied as often as LDA and LR (Holden et al., 2011).

In the reviewed literature, it is clear that many data conditions may affect performance of the classification methods. Sample size, group size ratios, distributions of the predictors, effect size, and homogeneity of variance-covariance matrices are the most studied conditions which are also important factors to consider for classification accuracy for LDA, LR, and CART. On the other hand, correlations between predictor variables, number of the variables, number of the groups in the dependent variable, model complexity, dynamic structure of the data, linearity, presence of outliers, multimodal structure of data, percent of initial misclassification, and group separation are also important and less studied data conditions for comparison of LDA, LR, and CART. Moreover, there is little known about effect of fundamental data conditions for classification: correlation between independent variables, number of the groups in the independent variable and number of predictor variables for the classification accuracy of the methods. Finally, it is noticeable in the literature that there are conflicts between results of some studies for particular conditions.

While CART shows higher performance than LDA and LR in different levels of sample size, homogeneity of variance covariance matrices and effect size, group size ratio, different model complexities, percent of initial misclassification, and level of group separation, it shows lower classification accuracy than LDA and LR under normal or skewed types of data. LR is expected to perform better than LDA under violation of LDA assumptions such as normality and homogeneity of variance covariance matrices. With non-inclusion of CART, some studies reported better performance of LR under nonlinearity and the presence of multimodal data. Finally, without having detailed comparison results, it is known that the dynamic environment of data and the existence of outliers affects performance of the classification methods.

In addition to these results, LDA, LR, and CART were reported to be affected by the number of predictor variables and the number of groups in the dependent variable. A larger number of groups decreased classification accuracies of the methods, while more predictor variables increased the classification accuracies and the methods LDA, LR, and CART showed comparable results under these conditions. Moreover, no study was found to compare LDA, LR, and CART at the same time for the effect of correlation of predictor variables while LDA was found to perform less efficiently and LR not to be affected significantly by multicollinearity. On the other hand, it should be noted that there are a limited number of studies to compare the methods under these conditions and further investigation is needed.

In summary, while previous research accommodates some level of knowledge about the factors which affect performance of LDA, LR and CART, further study is needed to have a better understanding of the performance of the group classification

methods LDA, LR, and CART. Particularly, the following factors have not been thoroughly investigated: correlations between predictor variables, number of the predictor variables, and number of the groups in the dependent variables. Moreover, these conditions should be evaluated not only for overall classification accuracy but also smaller and larger group classification accuracies, so that group size ratio should be included to produce more detailed results. Therefore, the research questions for this study are:

1) Which of the three methods (LDA, LR, and CART) performs better under different levels of correlation between predictor variables?
2) Which of the three methods (LDA, LR and CART) performs better under different numbers of groups in the dependent variables?
3) Which of the three methods (LDA, LR and CART) performs better under different numbers of the predictor variables?
4) Which of the three methods (LDA, LR and CART) performs better under different group size ratios?
5) Is there any significant interaction between level of correlation between predictor variables, number of the predictor variables, number of the groups, and group size ratios in the dependent variables for classification accuracies of the three methods?
6) Which of the three methods (LDA, LR, CART) performs better overall?

Measures of outcome variables

To be able to address the research questions and evaluate performance of the methods LDA, LR, and CART, two measures of outcome will be applied: overall rate of correct classification, rate of correct classification for the smallest group in terms of the group's sample size, and rate of correct classification for the largest group in terms of the group's sample size.

Rate of correct classification for all groups (rccA): rccA will be calculated based on dividing the number of all correctly predicted observations for their classes by total number of all observations. It is presented by the formula:

$$rccA = \frac{Number\ of\ correctly\ \ predicted\ observations\ for\ their\ classes}{Number\ of\ total\ obsevations} \quad (8)$$

Rate of correct classification for the smallest group in terms of groups' sample sizes (rccS): rccS will be calculated by dividing the number of correctly predicted observations for the smallest group by the total number of observations in the smallest group. It can be presented by the formula:

$$rccS = \frac{Number\ of\ correctly\ \ predicted\ observations\ for\ the\ smallest\ group}{Number\ of\ total\ obsevations\ in\ the\ smallest\ group} \quad (9)$$

As can be seen, rccA will be used for measuring overall classification accuracy and rccS will be used for measuring the classification accuracy of the smallest group.

Definitions

This study focuses on the conditions: correlations between predictor variables, group size ratios, number of the groups in the outcome variable, and number of the predictor variables. Correlation defines degree of linear associations between predictor

variables and it was set to .2 and .5. Group size ratio is the percentages of groups sizes within the whole sample was set to the cases when the groups were balanced and imbalanced in terms of groups' sample sizes. Number of the groups in the outcome variable indicates how many groups the outcome variable has and it was set to levels when there were two, three, and four groups. Number of the predictor variables defines number of the variables used to predict the group membership of observations and it was set to the levels when there were two, five and ten predictor variables. The performances of the methods were evaluated by rate of correct classifications which is the rate between number of correct group predictions and total sample sizes. On the other hand, while focusing on the conditions mentioned above, not being able to include other data conditions such effect size, predictor distributions, or sample size due to computational and time concerns was a limitation of this study. Finally, the fact that the focused data conditions were used with a limited number of levels is another main limitation of this study.

Chapter Two

Method

Chapter Two includes a description of the methodology of the study. First, details of the research design and data generation process are presented. Then, tools and procedures of data analysis will be discussed.

Research Design

In this study, factors regarding data properties will be controlled. The factors are number of the predictor variables (3 levels: 2, 5, 10), correlation between predictor variables (2 levels: .2, .5), number of groups (3 levels: 2, 3, 4), and group size ratio (2 levels: balanced, imbalanced). For the specifications of the balanced and imbalanced groups, see Table 10. The first two conditions are related to predictor variables while the latter are related to the outcome variable. Moreover, three different analysis methods (LDA, LR, and CART) were applied to compare their performances. Therefore, 3x2x3x2 = 36 different data conditions were created and analyzed with each of three methods. All other factors were assumed to be random and are uncontrolled. Sample size was fixed to 200 and for each condition 1000 iterations were simulated. Therefore, 3x36x200 = 21,600 simulated observations each having 1,000 iterations were included in this study.

To conduct the comparison study between LR, LDA, and CART and evaluating their performance under certain conditions summarized in Table 10, a Monte Carlo simulation technique was applied. Data generation via Monte Carlo simulation was conducted with R statistical software (R Core Team, 2016) and analysis of the results were conducted with SPSS statistical software (IBM Corp., 2013).

Monte Carlo methods are data simulation techniques relying on random sampling procedures. It is common to use the Monte Carlo approach to test theoretical hypotheses such as mathematical approximations, probability calculations, or probability distributions' parameter estimations by generating datasets that meet specified conditions (Paxton et al., 2001). In summary, it is a commonly used technique to compare statistical techniques and evaluate their performances. As Monte Carlo simulation allows researchers to generate variables randomly and manipulate desired characteristics (controlled variables), it will be used in this study as the data generation process.

For simplicity of data generation and analysis, all the predictor variables were created with a standard normal distribution (normal distribution with mean 0 and standard deviation 1). Table 10 presents the controlled variables and the levels that were used in this study.

Table 10.

Controlled Variables and Levels for the Study

Controlled variables	**Levels of the variable**		
Method	LDA, LR, CART		
Number of Predictor Variables	2, 5, 10		
Number of Groups	2, 3, 4		
Group Size Ratio	Imbalanced	Balanced	
	10:90	50:50	(2 groups)
	10: 20:70	33:33:33	(3 groups)
	10:15:20:55	25:25:25:25	(4 groups)
Correlation Between Variables	.2, .5		

Data Generation

To generate data with desired properties, the function MVRNORM in R software was used and therefore multivariate normality of predictor variables was satisfied. The MVNORM library in R allows researchers to specify the correlations between predictor variables and the number of predictor variables. The sample size was fixed at 200, as it is a common sample size in simulation studies and it is a reasonable number of observations in applied social science quantitative research. Additionally, effect size was fixed to 0.5 following Cohen's (1988) comment that 0.2, 0.5, and 0.8 are small, medium and large effect sizes respectively. Moreover, for LDA, prior probabilities were specified based on their observed group sizes as recommended by Lei and Koehly (2003). Based on their

suggestions, the groups were assigned to assigned to the probabilities which were ratios of their samples sizes in total sample. For example, for a two-group case with group size ratio 10:90, the prior probability for the smaller and larger groups were .1 and .9, respectively. On the other hand, when groups were balanced, the prior probabilities were 0.5 for both groups.

Controlled Variables and Their Patterns

In this study four variables were controlled: the number of the predictor variable, the number of the groups for outcome variable, group size ratio, and variable correlation. The number of the predictor variables and correlation matrices are qualities about predictor variables while the number of the groups and group size ratio are the qualities related to outcome variable. The details about conditions of each controlled variable are presented below.

Correlation (CORR)

The MVRNORM function in R software allows us to determine correlation between predictor variables for simulated data. Even though it is impossible to create all possible levels of correlations between variables, two levels of correlations (low, medium) between predictor variables were created. For low correlation, the coefficient was 0.2, and for medium correlation the coefficient 0.5 was used.

With 2, 5, and 10 predictor variables, the correlation matrices for low and medium correlation are, respectively, as follows in Figure 3.

Two Predictor Variables

$$\begin{bmatrix} - & - \\ .2 & - \end{bmatrix}, \begin{bmatrix} - & - \\ .5 & - \end{bmatrix}$$

Five Predictor Variables

$$\begin{bmatrix} - & - & - & - & - \\ .2 & - & - & - & - \\ .2 & .2 & - & - & - \\ .2 & .2 & .2 & - & - \\ .2 & .2 & .2 & .2 & - \end{bmatrix}, \begin{bmatrix} - & - & - & - & - \\ .5 & - & - & - & - \\ .5 & .5 & - & - & - \\ .5 & .5 & 5 & - & - \\ .5 & .5 & .5 & .5 & - \end{bmatrix}$$

Ten Predictor Variables

$$\begin{bmatrix} - & - & - & \cdots & - \\ .2 & - & - & \cdots & - \\ .2 & .2 & - & \cdots & - \\ \vdots & \vdots & \vdots & \ddots & - \\ .2 & .2 & .2 & .2 & - \end{bmatrix}, \begin{bmatrix} - & - & - & \cdots & - \\ .5 & - & - & \cdots & - \\ .5 & .5 & - & \cdots & - \\ \vdots & \vdots & \vdots & \ddots & - \\ .5 & .5 & .5 & .5 & - \end{bmatrix}$$

Figure 3

Correlation Matrices Between Predictor Variables with Two, Five and Ten Predictor Variables

When simulating predictor variables, it was discovered that the predictor variables were correlated higher than the fixed level, on average. For example, when fixing the correlations at .5 in MVRNORM, the correlations for the simulated data were a little higher than .5; i.e., .58, depending on the data condition. Therefore, smaller values of correlation coefficients were introduced to the R software during the data simulation process, so that the controlled conditions were satisfied. For all 36 data conditions, the new lower correlation values were tested and ensured to be equivalent in value to the fixed correlation coefficient values of .2 and .5.

Number of predictor variables (NPV)

Number of the predictor variables were automatically determined by the creation of correlation matrices, as when deciding correlations between predictor variables one needs to first decide the number of the variables and the function MVRNORM creates the number of the variables based on the determined correlation matrix. The levels of the number of the predictor variables for study were based on simulated data with two, five, and ten predictor variables.

Number of groups for outcome variable (GN)

In this study, the number of groups in the dependent variable has three levels: two, three or four groups in the dependent variables, which are the common in terms of number of the groups in application. To be able to generate groups, first the number of the observations for each group was counted based on the group size ratios, then for each group, the number of observations were generated. For example, for the case with three groups with a group size ratio 10:20:70, 20, 40 and 140 observations for different groups were simulated since total sample size will be 200 and it was distributed to the group according to the group size ratio. Different groups were labeled with different numbers. For example, with three groups in the dependent variable, groups were labeled as group 1, group 2, and group 3 and following the example above group 1 had 20 cases, group 2 had 40 cases, and group 3 had 140 observations. Once simulating and labeling groups for the outcome variable and simulating predictor variable datasets for each iteration, outcome variable and predictor variables were matched to each other randomly based on a standardized mean of 0.5 which is considered as a medium effect size by Cohen (1998).

Table 11 presents sample sizes for the groups based on group size ratios and the number of the groups.

Table 11.

Number of Groups and Groups Sizes for the Simulation

Number of groups	Groups Size Ratio	Sample Sizes for the groups
2	50:50	100:100
	10:90	20:180
3	10:20:70	20:40:140
	33:33:33	67:66:67
4	10:15:20:55	20:30:40:110
	25:25:25:25	50:50:50:50

Group Size Ratio (GSR)

Group size ratio is an important variable to control since it effects other manipulated variables. In this study, two different levels of groups size ratio were controlled: balanced group size ratios and unbalanced group size ratios. A balanced group size ratio exists when each group in the dependent variable has the same number of observations. An unbalanced level of group size ratio exists when the number of cases in the groups are not equal and there is a large difference between the largest and the smallest group in terms of number of observations.

While the balanced case was simulated to compare the methods' performances under balanced cases for overall prediction accuracies, having the imbalanced case accommodated evaluation of the methods' (LDA, LR, and CART) performances for

smaller group prediction accuracy as well as overall class prediction accuracies. Table 12 presents the number of the groups for the dependent variable and the group size ratios for balanced and unbalanced cases. In the notation 10:15:20:55, 10 is the percentage of observations of the smallest group while 55 is the percentage of observations of the largest group.

Table 12.

Number of Groups and GSR for Balanced and Imbalanced Cases

Number of Groups	**GSR for Imbalanced Case**	**GSR for Balanced Case**
2	10:90	50:50
3	10:20:70	33:33:33
4	10:15:20:55	25:25:25:25

Simulating groups of dependent variable

To simulate categorical outcome variables, first means of predictor variables for each group were introduced to the software. To be able to satisfy the fixed standardized group difference (.5 effect size) between each consecutive group in terms of sample size and the overall group mean to be zero, the group means were calculated based on simple mathematical equation systems. For example, for an imbalanced three group case all variables for the smallest, the second and the largest groups were assigned to the values -.8, -.3 and .2, accordingly. For the full list of all predictor variables for different group numbers and levels of group size ratios, see Table 13. Once the predictor variables were assigned to the determined values, based on correlations between predictor variables, group size ratios, and groups' sizes, the observations were created via the *c(rep())*

function in R, then all the observations were combined by the *data.frame (,)* function with all predictor variables and the dependent variable.

Table 13.

Means of Predictor Variables for Levels of GSR and Group Numbers

		Means of Predictor Variables for Groups			
GN	**GSR**	**Group 1**	**Group 2**	**Group 3**	**Group 4**
#2	10:90	-.45	.05	-	-
#2	50:50	-.25	.25	-	-
#3	10:20:70	-.80	-.30	.20	-
#3	33:33:33	-.50	.00	.50	-
#4	10:15:20:55	-1.10	-.60	-.10	.40
#4	25:25:25:25	-.75	-.25	.25	.75

Steps of data generation and manipulation process

Step 1: For the case when there were 2, 5, or 10 predictor variables, the data matrices were created by the function MVRNORM in R software. The MVRNORM function generates variables for each group based on a multivariate normal distribution. For example, for the case with five predictor variables, by generating all five predictor variables by MVRNORM, the predictor variables within each group follow the multivariate normal distribution which means they all follow the normal distribution individually but that does not necessarily mean when combining each group for the dependent variable, the predictor variables satisfied normality. This function also allows

one to specify the means and standard deviations of the predictor variables for each group in the dependent variable. On the other hand, while generating the predictor variables from multivariate normal distribution for each group, after combining them for whole datasets, multivariate normality is not necessarily satisfied for each iteration.

Step 2: For the cases when the outcome variable has 2, 3, or 4 groups, observations are generated based on their group size ratios. Then the groups were labeled as group 1, group 2, group 3, and group 4. The groups labeled with smaller numbers have a smaller number of observations. For example, for unbalanced cases, group 1 is always the smallest group in terms of group size. For three group case, the group size ratios were 67, 66, 67 for group 1, group 2, and group 3. Size of group 2 was 66 for locating group means easily.

Step 3: Assign the observations of outcome variables to created predictor variables randomly based on standardized group mean difference 0.5 and means of predictor variables zero.

By following the steps above and using required R functions, the dataset of 1000 iterations with desired manipulated and random conditions were simulated. When, non-convergence of data for a replication was encountered, another replication was run to replace it by R software, so that 1000 replications were completed. Then, training of the data was completed and the data was ready to analyze.

Analysis of Data

After generating the data with desired conditions, each method was used with similar datasets in terms of data conditions to predict the outcome variables separately.

The functions *lda, multinom* and *rpart* functions in R were used to conduct the analyses for LDA, LD, and CART. After the class predictions from three methods were obtained, an algorithm which controls if the method predicted the class correctly and finds number of correct predictions was created. Then, the rates of correct classification (rccA and rccS) of methods for each iteration were calculated and the second round of data for comparison of the methods was ready. To analyze results of this simulation study, a 3x2x3x3x2 factorial analysis of variance (ANOVA) procedure for Method X Correlation X Number of the Variables X Number of the Groups X Group Size Ratio was conducted based on the rate of correct classifications. A factorial ANOVA assesses main effects and interactions among factors. SPSS statistical software was used in conducting the factorial ANOVA and follow-ups.

Before conducting the factorial ANOVA, for all conditions, rate of correct classification prediction of each method was generated. As each condition had 200 observations, it was possible to count the success rates of class predictions by comparing observed (simulated) groups and predicted groups for LDA, LR and CART. Then for each iteration, the rate of correct classification was determined by dividing the number of correct classifications by the sample size. Thus, by creating an algorithm which calculates the rate of correct classifications for all the groups (rccA), or smaller group (rccS), then the datasets for the factorial ANOVA was ready.

In his research, Edwards (1985; p. 83) carried out all of the analyses using the arcsine transformed value of the proportions as a dependent variable and the results were identical for the proportions and transformed values. Therefore, following this results, the analyses of this study were based on proportions.

As the statistical significance of interactions and main effects are affected by sample size and the sample size of 1000 (number of the iterations for each combination of the conditions) is relatively large, and additional indices will be employed, partial eta squared (η_p^2). Partial eta squared is an index which calculates proportion of total sample variance explained related to group membership by a determined effect partialling out other main and interaction effects (Pierce et al., 2014; Richardson, 2011). It can be presented by the formula

$$\eta_p^2 = \frac{SS_{effect}}{SS_{total} + SS_{error}} \quad (10)$$

where SS_{effect} is sum of squares for the particular effect SS_{total} is the total sum of squares and, SS_{error} is the error sum of squares. Haase (1982) reported partial eta squared (η_p^2) .083 to be a medium effect in size. Therefore, for the factorial ANOVA, main effects and interactions in this study, which had η_p^2 values equivalent or larger than .083 were considered having at least medium effect.

Factorial ANOVA assumes normality of predictor variables, homogeneity of variance (HOV), and independence of observations. By the design of this study the assumptions, for independence of observations was satisfied. On the other hand, due to large sample sizes (number of iterations), differences in group size ratios, group numbers and their respective means, the assumption of homogeneity of variance was not satisfied based on Levene's test. Moreover, based on a rule of thumb for skewness to be between -1 and +1, almost all of the cells satisfied the requirement for normality except some imbalanced cases with a binary outcome variable and two or five predictor variables (skewnesses were still within -2 to +2). However, ANOVA is robust against violation of

normality and HOV, especially when there is a large dataset with a balanced design. Therefore, effects of these violations were ignored.

After factorial ANOVAs were run and the results obtained, the follow-ups were implemented based on meaningful effects of interactions or main effects. Once an interaction was found to be meaningful, the dataset was split by level on one factor present in the interaction and the effects of the other conditions evaluated. Effects of the conditions were evaluated again based on if η_p^2 values were equivalent or larger than .083 and the effects which had η_p^2 less than 0.083 were not interpreted, except for the method effect on rccA. This process was followed until all the cells in the ANOVA design were investigated. Finally, the mean rccA and rccS values were compared for levels of the conditions.

Chapter Three

Results

In this Chapter, results from the data analysis are reported. Since there were two outcome variables, this chapter has two sections: results for rate of correct classification for all groups (rccA) and for correct classification of the smallest group (rccS). At the end of the chapter, results comparing rccA and rccS are summarized. It should be noted that due to use of a decimal outcome variable, results are reported to three decimal places rather than two as suggested by APA.

Results for rccA

Overview

Based on the factorial ANOVA results, all the interactions and main effects were statistically significant ($p < .001$). However, due to the large sample size, instead of statistical significance, partial eta squared values were employed to evaluate the importance of main effects and interactions. Using Haase's (1982) finding of a medium effect size as 0.083, any partial eta squared (η_p^2) value greater than 0.083 of main effects or interactions was considered meaningful. For ease of reporting, the conditions (correlation, number of predictor variables, number of the groups, and group size ratio) are reported with their abbreviations: Corr, NPV, GN, and GSR.

The overall factorial ANOVA model was statistically significant and had a meaningful partial eta squared value ($p < .001, \eta_p^2 = .969$) for the outcome variable rccA. Following the rule of thumb for meaningful effects of $\eta_p^2 > .083$, all main effects were meaningful. Moreover, all two-way interactions except Method x NPV had meaningful effects on rccA. Finally, three-way interactions: Method x NPV x GN ($\eta_p^2 = .087$), Method x GN x GSR ($\eta_p^2 = .169$), and Corr x NPV x GN ($\eta_p^2 = .177$) had meaningful effects on rccA. All remaining interactions had partial eta square values less than 0.083, and are not interpreted here. Details of the overall factorial ANOVA results are provided in Table 14.

Table 14.

ANOVA Summary Table for the Effects of Method, Corr, NPV, GN, and GSR on rccA

Source	df	*F*	*p*	Partial Eta Squared (η_p^2)
Method	2	127047.3	<0.001	0.702
Corr	1	108719.7	<0.001	0.502
NPV	2	115969.6	<0.001	0.683
GN	2	527124.9	<0.001	0.907
GSR	1	1149617	<0.001	0.914
Method * Corr	2	9936.965	<0.001	0.156
Method * NPV	4	1699.443	<0.001	0.059
Method * GN	4	6056.53	<0.001	0.183
Method * GSR	2	50744.27	<0.001	0.485
Corr * NPV	2	20721.98	<0.001	0.278
Corr * GN	2	25021.86	<0.001	0.317
Corr * GSR	1	20509.31	<0.001	0.160
NPV * GN	4	17473.88	<0.001	0.393
NPV * GSR	2	19755.73	<0.001	0.268
GN * GSR	2	42466.41	<0.001	0.440
Method * Corr * NPV	4	2391.374	<0.001	0.081
Method * Corr * GN	4	2281.087	<0.001	0.078
Method * Corr * GSR	2	2668.998	<0.001	0.047
Method * NPV * GN	8	1278.243	<0.001	0.087
Method * NPV * GSR	4	690.048	<0.001	0.025
Method * GN * GSR	4	5501.345	<0.001	0.169
Corr * NPV * GN	4	5786.564	<0.001	0.177
Corr * NPV * GSR	2	3335.765	<0.001	0.058
Corr * GN * GSR	2	1000.944	<0.001	0.018
NPV * GN * GSR	4	335.823	<0.001	0.012
Method * Corr * NPV * GN	8	967.997	<0.001	0.067
Method * Corr * NPV * GSR	4	532.553	<0.001	0.019
Method * Corr * GN * GSR	4	246.794	<0.001	0.009
Method * NPV * GN * GSR	8	200.174	<0.001	0.015
Corr * NPV * GN * GSR	4	397.885	<0.001	0.015
Method * Corr * NPV * GN * GSR	8	93.965	<0.001	0.007
Error	107892			
Total	107999			

Note: Method = Methods (LDA, LR, CART); Corr = correlation levels (.2, .5); NPV = Number of the predictor variables (2,5,10); GN = Number of groups in dependent variable (2,3,4); GSR = Group size ratio (imbalanced, balanced).

To follow up overall results, first the effects of three-way interactions, then effects of two-way interactions, and finally the main effects are interpreted.

Interaction of Method, Group Numbers, and Group Size Ratio

The method x GN x GSR interaction effect on rccA was statistically significant and greater than medium in size ($\eta_p^2 = .169$). To follow up this result, the data were split into levels by group size ratio creating simple interactions and the method x GN η_p^2 was examined. By doing so, effects of the interaction of the factors (method, GN) under single levels of the third factor can be examined (Myers et al., 2013). At both levels of GSR, when the group size ratios were imbalanced ($\eta_p^2 = .20$) and balanced ($\eta_p^2 = .327$) the method x GN had meaningful interaction effects on rccA. Then, the data were split by level of GN in addition to GSR level to investigate effect of method on rccA. The method effect was significant at all levels of GSR for all the levels of GN. The partial eta squared values of method are reported in Table 15.

Table 15.

Partial Eta Squared Values for Method Effect by Level of GSR and GN

GSR	GN	η_p^2
Imbalanced	#2	.130
	#3	.537
	#4	.482
Balanced	#2	.766
	#3	.888
	#4	.722

Table 16 presents mean rccA for LDA, LR, and CART for different levels of GSR and GN. It should be noted that in all cases CART outperformed LR and LDA, though the difference with two imbalanced groups was trivial.

Table 16.

Mean rccA of LDA, LR, and CART by Level of GSR and GN

		Mean rccA		
GSR	**GN**	**LDA**	**LR**	**CART**
Imbalanced	#2	.902	.902	.907
	#3	.731	.731	.772
	#4	.656	.660	.704
Balanced	#2	.659	.660	.773
	#3	.468	.568	.678
	#4	.514	.566	.634

When the groups were imbalanced and number of the groups was two, the performance difference between LDA, LR, and CART was less than 0.5%. When the number of the groups was three or four, the difference between LDA and LR was less than 0.5%, but CART performed better than these two methods by around 4%.

When the groups were balanced and GN was two, there was a trivial difference (.1%) between LDA and LR, but CART performed better than LDA and LR by 12%. When GN was three, LDA showed the lowest performance while LR showed 10% better performance than LDA and CART was the best performing method by 10% better than LR. Finally, when GN was four, LR showed around 5% better performance than LDA, and CART was the best performing method, about 7% better performance than LR.

When the number of groups increased, all methods showed lower rccA. All methods correctly predicted the group classification better when there were fewer groups. Yet, CART showed somewhat better resistance against increasing GN than LDA and LR.

For example, when the groups were imbalanced, increasing GN from two to four resulted in LDA and LR decreases in rccA of 24.6% and 24.2% while CART decreased rccA by 20.3%. When the groups were balanced, the decrease in rccA for all methods was lower. When the groups were balanced, increasing GN from two to four resulted in LDA, LR, and CART decreased rccA by 14.5%, 9.4%, and 13.9%, respectively. In that case, superiority of LR was observed with a more than 4% difference from LDA and CART. Moreover, the difference between LDA and CART was less than 1%. The only exception was for the balanced case when LDA had higher rccA for the four-group case (.514) than for the three-group case (.468).

The methods had a higher rccA when the groups were imbalanced than when the groups were balanced. In general, when switching from the balanced case to the imbalanced case, LDA, LR, and CART performance increased by 21.6%, 16.6%, and 9.93%, respectively. Furthermore, at higher levels of GN, sensitivity of the methods to the GSR of the groups was smaller. It should be noted that when increasing GN, the group differences between the smallest size group and the largest size group also increased by the design of the study.

Figure 4 provides a graphical presentation of rccA for the methods and different levels of GN when GSR was imbalanced.

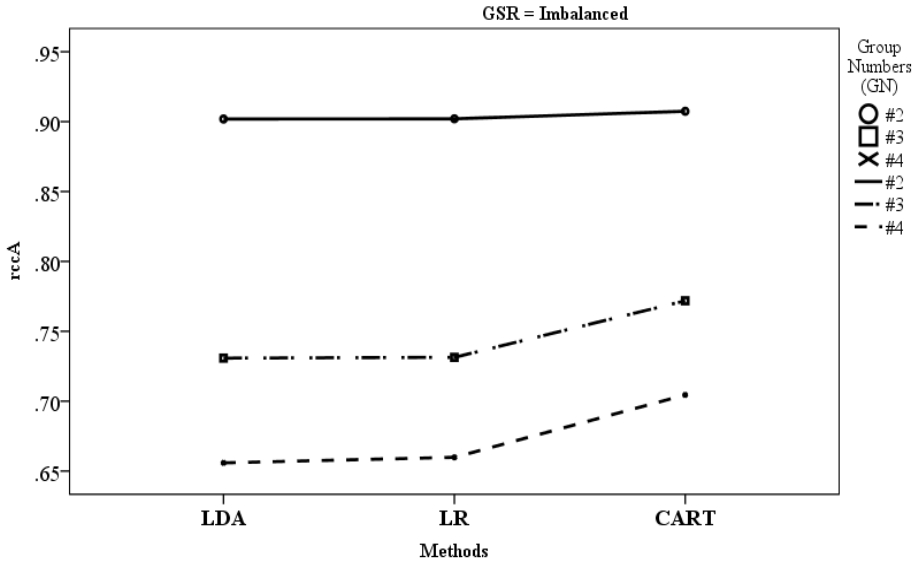

Figure 4

Mean rccA of Method by Level of GN When GSR is Imbalanced

Figure 5 provides a graphical presentation of rccA for the methods at different levels of GN when GSR was balanced.

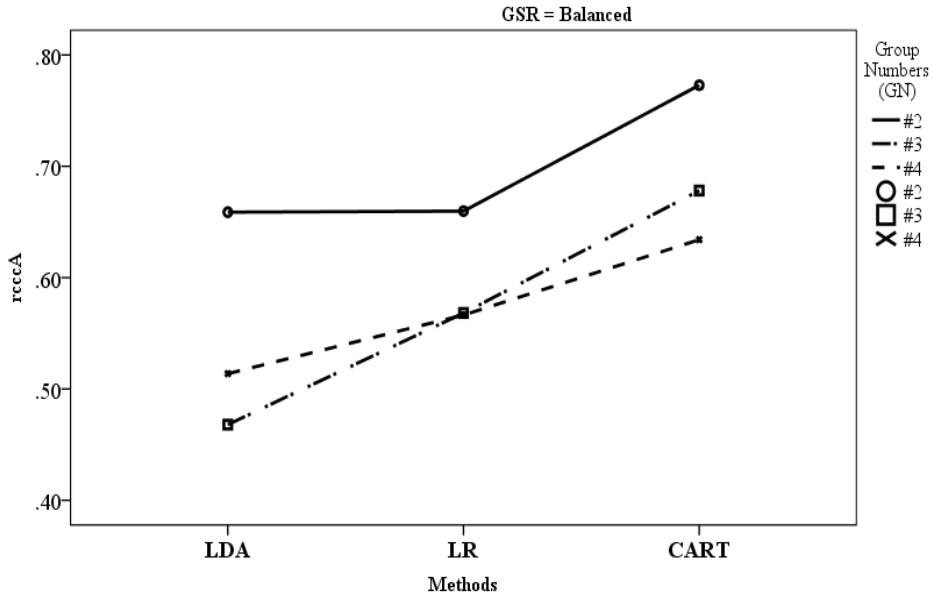

Figure 5

Mean rccA of Method by Level of GN When GSR is Balanced

Interaction of Method, Number of Predictor Variables and Group Number

The Method x NPV x GN interaction was significant and greater than medium in effect size rccA ($\eta_p^2 = .087$). To follow up this result, the data were split by level of GN and method x NPV η_p^2 was examined. Only at the level when GN was four did the method x NPV show a meaningful interaction ($\eta_p^2 = .276$), and at the other two levels of GN, the interaction had an effect size smaller than .083. Since for the cases when there were two groups and three groups the method x NPV interactions were not meaningful, only the case when GN was four is interpreted. The data were split by level of NPV in addition to GN level but only the case of four groups was examined to investigate the effect of method on rccA. The method effect was significant at all the levels of NPV when GN was four. The partial eta squared values of method are reported in Table 17.

Table 17.

Partial Eta Squared Values for Method Effect at Different Levels of NPV when GN was Four

GN	NPV	η_p^2
#4	#2	.789
	#5	.689
	#10	.243

Table 18 presents mean rccA for LDA, LR, and CART for different levels of NPV when GN was four. It should be noted that when NPV was two and five, CART outperformed LR and LDA, but the difference between CART and LR when NPV was ten was trivial.

Table 18.

Mean rccA of LDA, LR, and CART by Level of NPV When GN was four

GN	NPV	Mean rccA		
		LDA	LR	CART
#4	#2	.496	.527	.617
	#5	.581	.607	.678
	#10	.678	.705	.713

According to these results, when NPV was two, LR performed better than LDA by 3.1% and CART performed better than LR by 9%. When NPV was five, LR performed better than LDA by 2.6% and CART performed better than LR 7.1%. Finally, when NPV was 10, LR performed better than LDA by 2.7% and CART performed better than LR by 0.8%. Therefore, CART performed better than LDA and LR at all cases of NPV when GN was four, but when there were ten predictor variables, the difference between CART and LR was trivial.

With increasing NPV, rccA increased for all the methods. However, the contribution of additional predictor variables had different effects on different methods. LR was the best, LDA was second, and CART was the last in terms of increasing rccA by increasing NPV. For instance, by increasing NPV from two to ten, LDA increased its prediction ability by 18.2%, LR by 28%, and CART by 9.6%. It was noticeable that, at the highest level of NPV, the gap between the methods became smaller and CART did not benefit from additional predictor variables as much as the other two methods did.

Figure 6 provides a graphical presentation of mean rccA for the methods at different levels of NPV when GN was four.

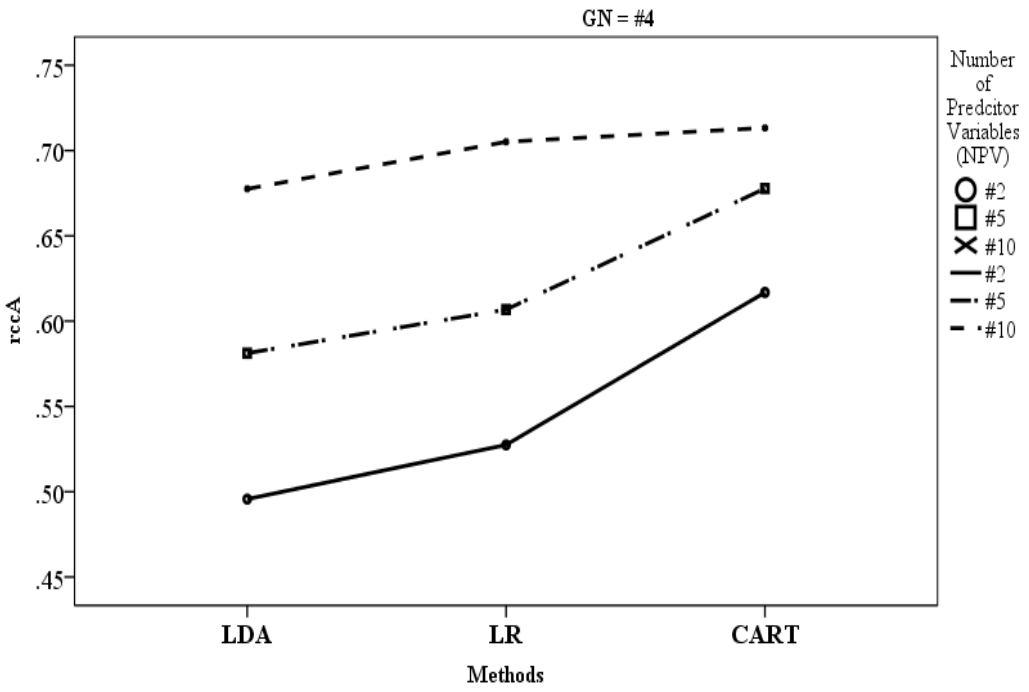

Figure 6

Mean rccAs of Method by Level of NPV with GN Equal to Four

Interaction of Correlation, Number of Predictor Variables, and Group Number

The Corr x NPV x GN interaction effect on rccA was significant and greater than medium in effect size ($\eta_p^2 = .177$). To follow up this result, the data were split by level of Corr and the NPV x GN η_p^2 was examined. At both levels of Corr, when the correlations between predictor variables were .2 ($\eta_p^2 = .612$) and .5 ($\eta_p^2 = .11$), NPV and GN had meaningful interaction effect on rccA. Then, the data were split by level of GN in addition to Corr level to investigate effect of NPV on rccA. The NPV effect was

significant at all levels of Corr for all the levels of GN. The partial eta squared values of NPV are reported in Table 19.

Table 19.

Partial Eta Squared Values for NPV Effect at Different Levels of Corr and GN

Corr	GN	η_p^2
.2	#2	.401
	#3	.747
	#4	.921
.5	#2	.201
	#3	.406
	#4	.581

Table 20 presents mean rccA for levels of NPV at the levels of GSR and GN. It was observed that at the lowest level of GN (#2), the contributions of additional predictor variables for average rccA were trivial while at case when GN was high (#4) and Corr was low, the increase in rccA by the contribution of additional variables was noticeable.

Table 20.

Mean rccAs with NPV of 2, 5, and 10 by Level of GN and Corr

		rccA		
Corr	GN	NPV #2	NPV #5	NPV #10
.2	#2	.785	.810	.828
	#3	.623	.680	.730
	#4	.558	.666	.785
.5	#2	.779	.794	.806
	#3	.613	.641	.662
	#4	.535	.578	.612

At both levels of correlation when GN was two, greater levels of NPV had greater rccAs and the difference was 2.5% or less. At the .2 level of Corr when GN was three, the

difference between the greater level of NPV with the one level smaller NPV in rccA was around 5% and when GN was four the difference was between 10% and 13%. The same differences for the .5 level of Corr when GN was three or four was about 3% or 4%. Even though it was clear that additional predictor variables increased prediction power of group membership methods, the case when Corr was .2 and GN was four had the most noticeable improvement with the addition of more predictor variables.

At both low and moderate levels of correlation and with a different number of predictor variables, rccA was higher for two groups than for three or four. As NPV increased, rccA increased for all conditions of Corr and GN.

Figure 7 presents mean rccAs for different levels of NPV at levels of GN when the correlation between predictor variables was .2.

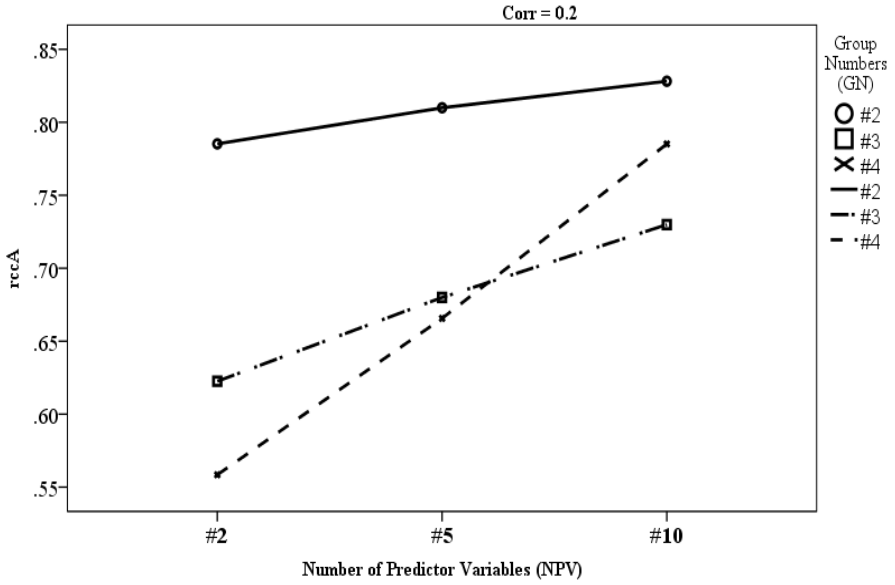

Figure 7

Mean rccAs of Level of NPV by Level of GN When Corr is .2

Figure 8 presents mean rccAs for different levels of NPV at levels of GN when the correlation between predictor variables was .5.

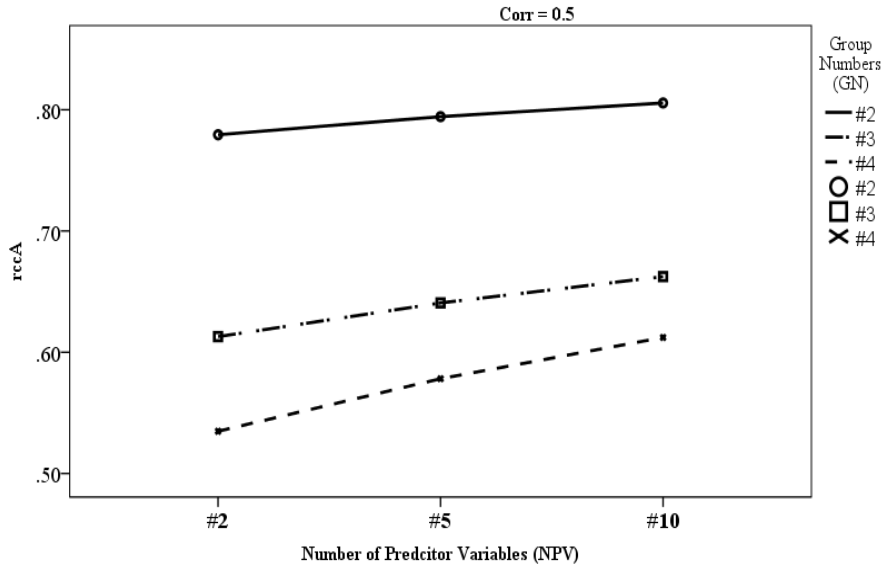

Figure 8

Mean rccAs of Level of NPV by Level of GN When GN and Corr is .5

For a general sense of the effect of NPV, GN and Corr on accuracy of group prediction, Figure 9 presents effects on rccA when increasing NPV, GN and Corr.

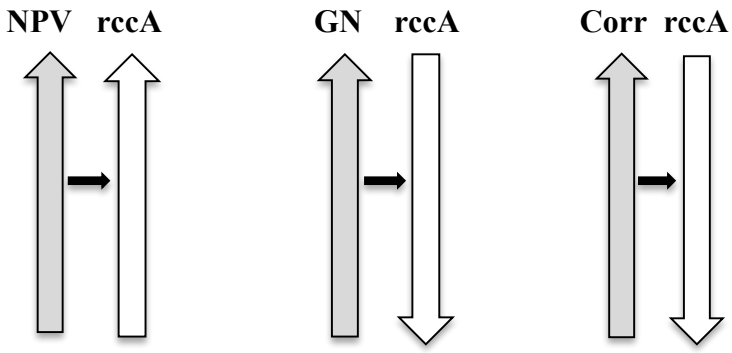

Figure 9

Reactions of rccA to Increase in NPV, GN, and Corr

As presented in Figure 9, increasing the number of predictor variables led to an increase in rccA while increasing the correlation between predictor variables or number of the groups in the outcome variable led to a decrease in rccA except for one case (Corr = .2 and GN = 4).

Interaction of Method and Correlation

The Corr x Method interaction effect on rccA was significant and greater than medium in effect size ($\eta_p^2 = .156$). To follow up this result, the data were split by level of correlation and at both levels when the correlation between predictor variables was .2 ($\eta_p^2 = .55$) and .5 ($\eta_p^2 = .796$), the method condition had a meaningful effect on rccA.

Table 21 presents mean rccA by method at each level of correlation. It was observed that at the low level of correlation, the methods performed better for prediction of group membership. Moreover, the influence of correlation level on LDA and LR for rccA was higher than its influence on CART.

Table 21.

Mean rccAs of Method by Level of Correlation

Corr	Method	Mean rccA
.2	LDA	.688
	LR	.712
	CART	.755
.5	LDA	.621
	LR	.650
	CART	.735

When the correlation was low (.2), the performance difference between LR and LDA in terms of rccA was 2.4% in favor of of LR and CART had a 4.3% better performance than LR. When the correlation was medium (.5), the performance difference

between LR and LDA was 2.9% in favor of LR and CART had a 8.5% better performance than LR.

The methods LDA and LR had 6.7% and 6.2% better performances at the low level of correlation than at the medium level while CART had just a 2% better performance at the low level of correlation. Figure 10 depicts mean rccA of the methods by level of correlation.

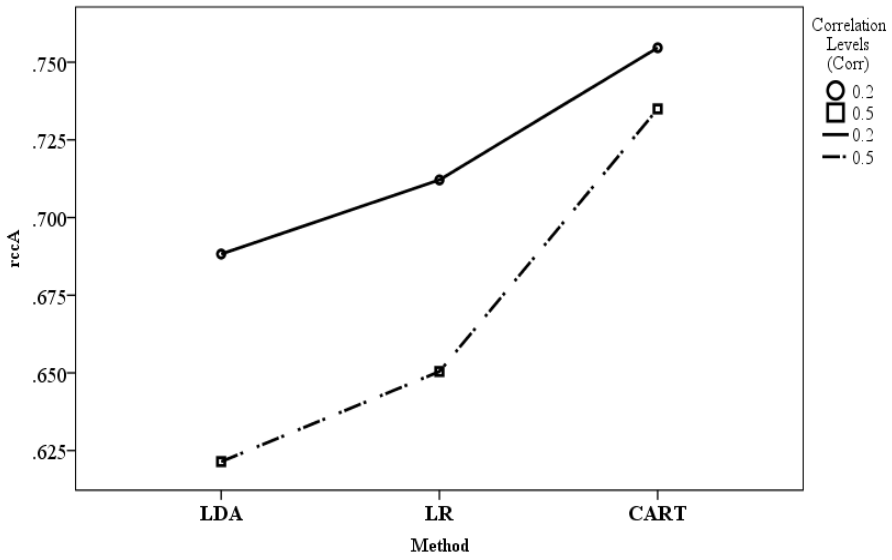

Figure 10

Mean rccAs for Method by Level of Correlation

Interaction of Method and Group Numbers

The Method x GN interaction effect on rccA was significant and greater than medium in effect size ($\eta_p^2 = .183$). To follow up this result, the data were split by level of GN. The method condition had a meaningful effect on rccA at all levels of GN: when there were two groups ($\eta_p^2 = .631$), three groups ($\eta_p^2 = .81$) and four groups ($\eta_p^2 = .628$).

Table 22 presents mean rccA by method by GN. It was observed that at lower levels of GN, the methods performed better for prediction of group membership. Moreover, the influence of group number on performance of the methods was not large in terms of percentage.

Table 22.

Mean rccAs of the Methods at Different Levels of GN

GN	Method	Mean rccA
#2	LDA	.780
	LR	.781
	CART	.840
#3	LDA	.599
	LR	.650
	CART	.725
#4	LDA	.585
	LR	.613
	CART	.669

When there were two groups, the performance difference between LDA and LR was trivial, but CART had a 6% better performance in mean rccA. When there were three groups, LR had a 5.1% better performance than LDA and CART performed 7.5% better than LR. With four groups, LR had a 2.8% better performance than LDA and CART performed 5.6% better than LR.

All the methods decreased prediction accuracies with increasing GN. The differences between the cases of four groups and three groups in rccA for LDA, LR, and CART were 19.5%, 16.8% and 17.1%, respectively. Figure 11 depicts mean rccA by method at levels of GN.

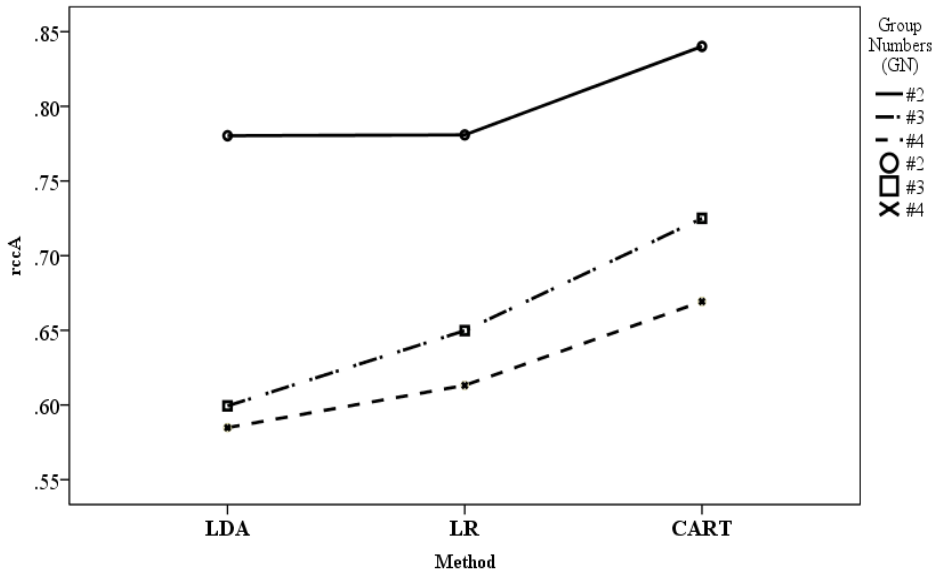

Figure 11

Mean rccAs for Method by Number of Groups

Interaction of Method and Group Size Ratio

The Method by GSR interaction effect on rccA was significant and greater than medium in effect size ($\eta_p^2 = .485$). To follow up this result, the data were split by level of GSR. The method condition had a meaningful effect on rccA at the both levels of GSR when group size ratio was imbalanced ($\eta_p^2 = .42$) and balanced ($\eta_p^2 = .805$).

Table 23 presents mean rccA by method by GSR. At the imbalanced level of GSR, the methods performed better than in the balanced case. Moreover, the influence of GSR on performance of the methods was not different by large percentages for imbalanced GSR.

Table 23.

Mean rccAs of Method by Level of GSR

GSR	Method	Mean rccA
Imbalanced	LDA	.763
	LR	.764
	CART	.795
Balanced	LDA	.547
	LR	.598
	CART	.695

In the imbalanced case, the performance between LDA and LR was trivial, but CART had around 3.1% higher rccA. When the groups in the dependent variable were balanced in terms of their sample sizes LR had a 5.1% higher performance than LDA, and CART had a 9.7% higher rccA than LR.

The methods LDA, LR, and CART showed 21.6%, 16.6%, 10% better performance, respectively, in the imbalanced case than the balanced case. CART has less sensitivity to variation in group size ratio. While reporting these results, it should be noted that in the imbalanced case, the majority of the sample belonged to the largest group, with an advantage in rccA. Figure 12 depicts mean rccA by method by level of GSR.

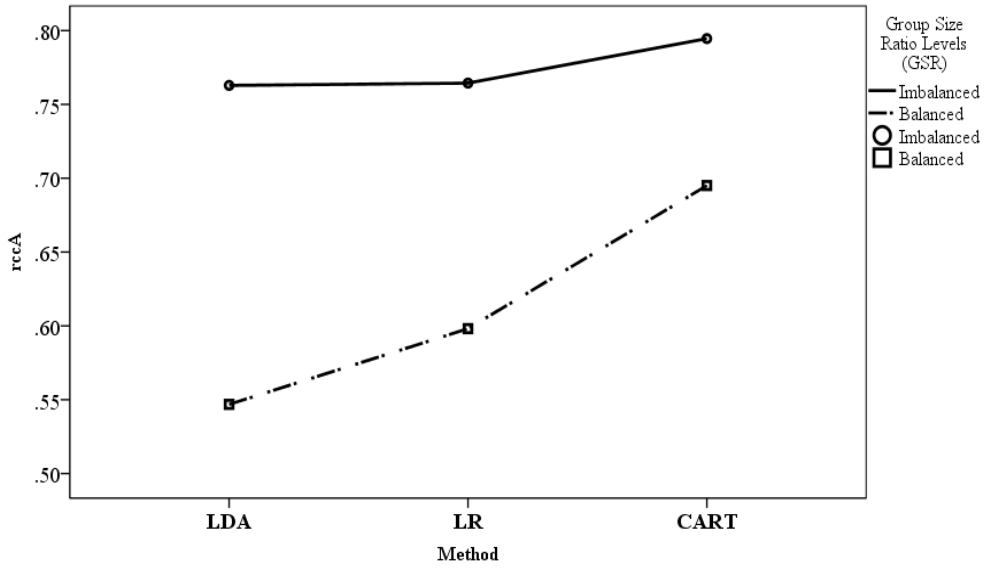

Figure 12

Mean rccAs for Method by Level of GSR

Interaction of Correlation and Number of Predictor Variables

The Corr x NPV interaction effect was greater than medium in effect size ($\eta_p^2 = .278$). To follow up this result, the data were split by level of correlation. When the correlation between the predictor variables were .2 ($\eta_p^2 = .81$) and .5 ($\eta_p^2 = .423$), NPV had a meaningful effect on rccA.

Table 24 presents mean rccA by NPV and level of Corr. It was observed that influence of correlation in rccA was higher at the cases with higher number of predictor variables than lower number of predictor variables. Moreover, rccA values were higher at the cases with lower level of correlation than the cases with medium level of correlation.

Table 24.

Mean rccAs of Level of NPV by Level of Corr

Corr	NPV	Mean rccA
.2	#2	.655
	#5	.719
	#10	.781
.5	#2	.642
	#5	.671
	#10	.693

When Corr was .2, increasing NPV from two to five resulted in a 6.4% increase in mean rccA and increasing NPV from five to ten resulted in a 6.2% increase in mean rccA. On the other hand, Corr was .5, increasing NPV from two to five resulted in 2.9% increase in mean rccA and increasing number of the predictor variables from five to ten resulted in 2.1% increase in mean rccA. Figure 13 depicts mean rccA by level of NPV for levels of correlation between predictor variables.

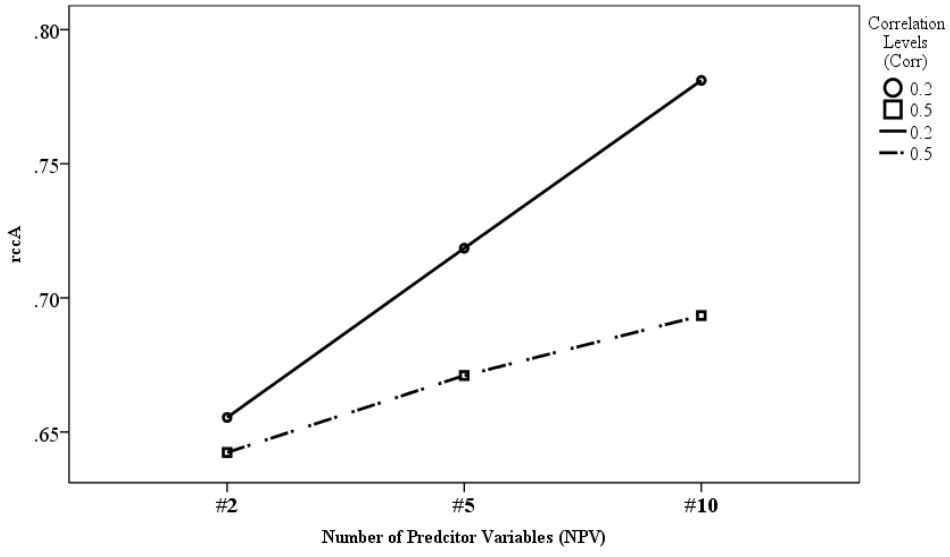

Figure 13

Mean rccAs for Number of Predictor Variables by Level of Correlation

Interaction of Correlation and Group Number

The Corr x GN interaction effect on rccA was significant and greater than medium in effect size ($\eta_p^2 = .317$). To follow up this result, the data were split by level of correlation. When the correlations between the predictor variables were .2 ($\eta_p^2 = .867$) and .5 ($\eta_p^2 = .934$), GN had a meaningful effect on rccA.

Table 25 presents mean rccA by GN and Corr. Increasing GN or Corr resulted in decreases in mean rccA, with the largest decrease for four groups.

Table 25.

Mean rccAs of the Levels of GN at the Different Levels of Corr

Corr	GN	Mean rccA
.2	#2	.808
	#3	.667
	#4	.670
.5	#2	.793
	#3	.639
	#4	.575

At the low level of correlation (.2), increasing GN from two to three resulted to 14.1% decrease in mean rccA while there was a trivial difference between the cases of three groups and four groups. On the other hand, at a medium level of correlation, increasing GN from two to three resulted in a 15.4% decrease in rccA and from three to four resulted in a 6.4% decrease in rccA.

When GN was two, three, and four, the mean rccA differences at the low level of correlation and medium level of correlation were respectively 1.5%, 2.8%, and 9.5%. Figure 14 depicts mean rccA by level of GN by level of correlation between predictor variables.

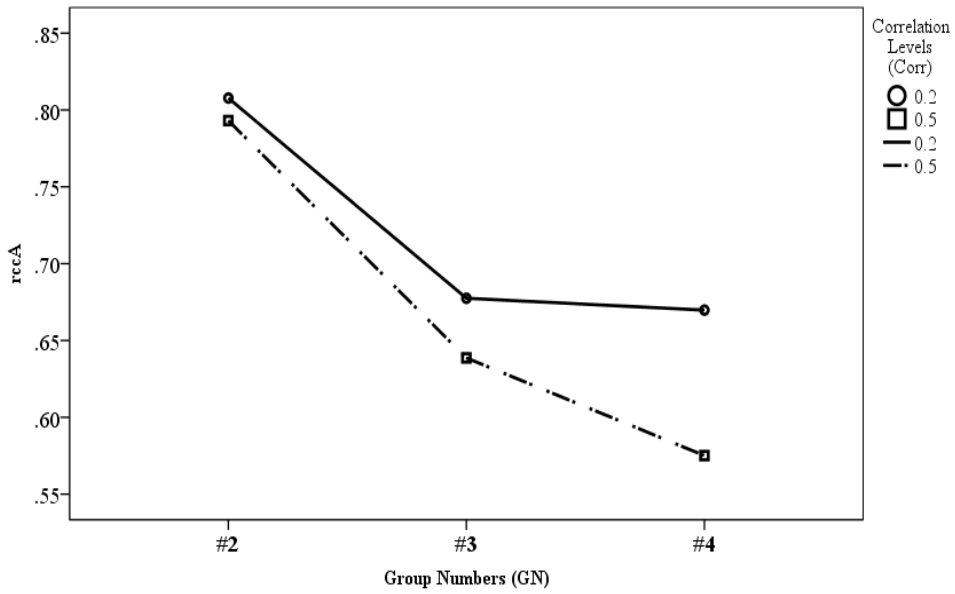

Figure 14

Mean rccAs for Number of Groups by Level of Correlation

Interaction of Correlation and Group Size Ratio

The Corr x GSR interaction effect on rccA was significant and greater than medium in effect size ($\eta_p^2 = .16$). To follow up this result, the data were split by level of correlation. When the correlation between the predictor variables were .2 ($\eta_p^2 = .887$) and .5 ($\eta_p^2 = .933$), GSR had a meaningful effect on rccA.

Table 26 presents mean rccA by level of GSR by level of correlation. When group size ratios were balanced, rccA was affected more by the increase in correlation than when group size ratios were imbalanced.

Table 26.

Mean rccAs of Level of GSR by Level of Corr

Corr	GSR	Mean rccA
.2	Imbalanced	.788
	Balanced	.649
.5	Imbalanced	.760
	Balanced	.578

Figure 15 illustrates mean rccA for GSR by Corr.

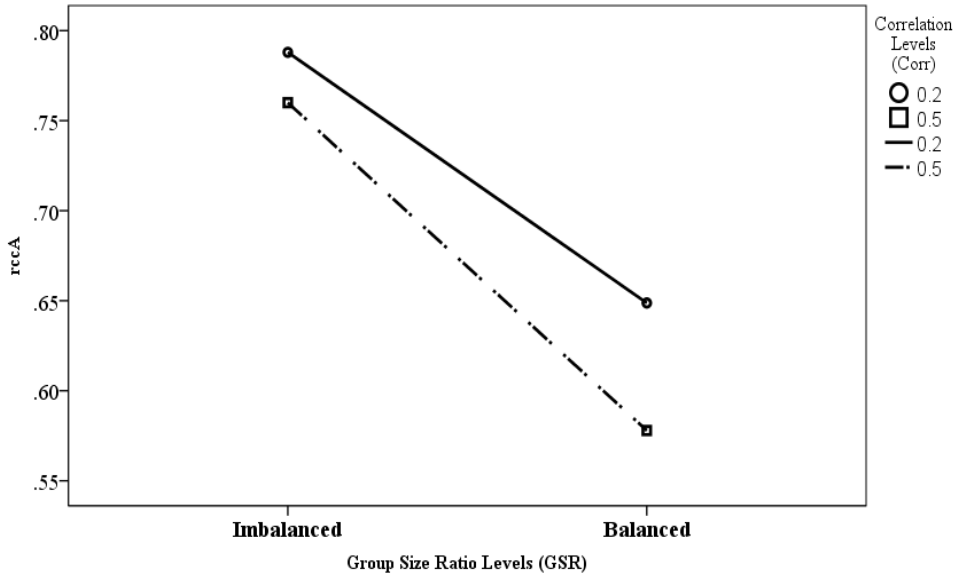

Figure 15

Mean rccAs for Level of GSR by Level of Correlation

Interaction of Number of Predictor Variables and Group Numbers

The NPV x GN interaction effect on rccA was significant and greater than medium in effect size ($\eta_p^2 = .393$). To follow up this result, the data were split by NPV level. At all NPV levels when there were two predictor variables ($\eta_p^2 = .944$), five

variables ($\eta_p^2 = .907$), and ten variables ($\eta_p^2 = .836$), GN had a meaningful effect on rccA.

Table 27 presents mean rccA by level of GN by NPV. Cases with a higher number of predictor variables were less affected by the increase in group numbers.

Table 27.

Mean rccAs of Level of GN by Level of NPV

NPV	GN	Mean rccA
#2	#2	.782
	#3	.618
	#4	.547
#5	#2	.802
	#3	.660
	#4	.622
#10	#2	.817
	#3	.696
	#4	.699

Increasing GN resulted in lower mean rccA except for a trivial difference with 10 predictor variables and 3-4 groups. Increasing group numbers from two to four for the cases when there were two, five or ten predictor variables resulted 23.5%, 18%, and 11.8% decrease in rccA respectively.

On the other hand, increasing the number of predictor variables resulted in higher rccA when controlling for GN. When there were two, three, and four groups in the dependent variable, increasing the number of predictor variables from two to ten resulted in a 3.5%, 4.2%, and 15.2% increase in rccA, respectively. Figure 16 depicts mean rccA by level of group number by number of predictor variables.

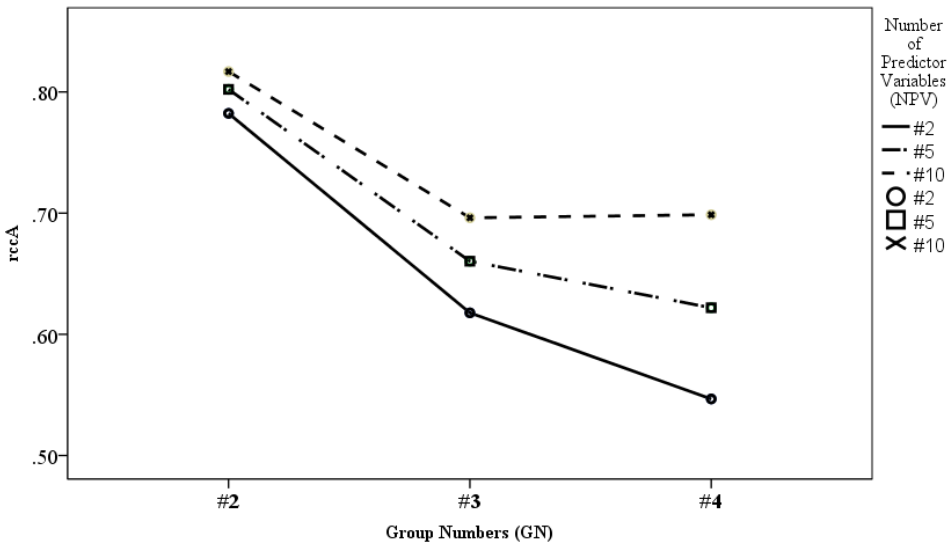

Figure 16

Mean rccAs of Number of Group by Level of NPV

Interaction of Number of Predictor Variables and Group Size Ratios

The NPV x GSR interaction effect on rccA was significant and greater than medium in effect size ($\eta_p^2 = .268$). To follow up this result, the data were split by GSR level. When the groups were imbalanced ($\eta_p^2 = .604$) and balanced ($\eta_p^2 = .739$), the NPV condition had a meaningful effect on rccA.

Table 28 presents mean rccA by level of NPV by level of GSR. The influence of increasing NPV was stronger in the balanced case than in the imbalanced case. Furthermore, the difference in rccA between the levels of GSR was lower when there were more predictor variables.

Table 28.

Mean rccAs of Level of NPV by Level of GSR

GSR	NPV	Mean rccA
Imbalanced	#2	.748
	#5	.773
	#10	.800
Balanced	#2	.550
	#5	.616
	#10	.674

Figure 17 depicts mean rccA by level of GSR by NPV.

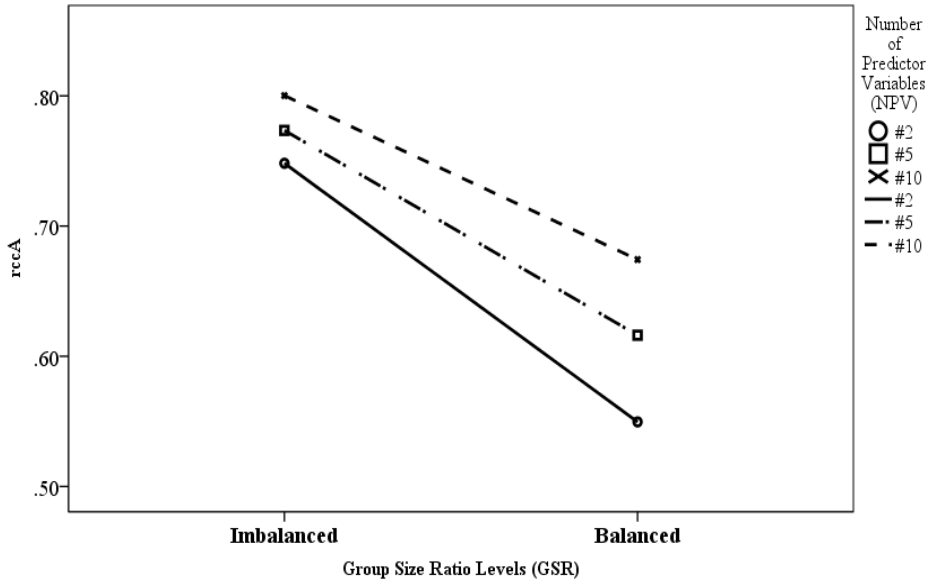

Figure 17

Mean rccAs for Level of GSR by Level of NPV

Interaction of Group Numbers and Group Size Ratio

The GN x GSR interaction effect on rccA was significant and greater than medium in effect size ($\eta_p^2 = .44$). To follow up this result, the data were split by GSR

level. When the groups into the dataset were imbalanced ($\eta_p^2 = .969$) and balanced ($\eta_p^2 = .793$), the GN condition had a meaningful effect on rccA.

Table 29 presents mean rccA by level of GN by level of GSR. Cases with less group numbers were influenced more by change in group size ratio levels than then the cases with more group numbers. Moreover, variation in mean rccA for the cases with different number of group numbers in balanced case was less than imbalanced case.

Table 29.

Mean rccAs of Level of GN by Level of GSR

GSR	GN	Mean rccA
Imbalanced	#2	.904
	#3	.745
	#4	.673
Balanced	#2	.697
	#3	.571
	#4	.571

When groups were imbalanced with two groups, mean rccA was 5.9% higher than when there were three groups, and three groups had a 7.2% higher mean rccA than when there were four groups.

When there were two, three, and four groups, the imbalanced case had 20.7%, 17.4%, and 10.2% higher mean rccAs than the balanced case, respectively.

Figure 18 depicts mean rccA by level of GSR by GN.

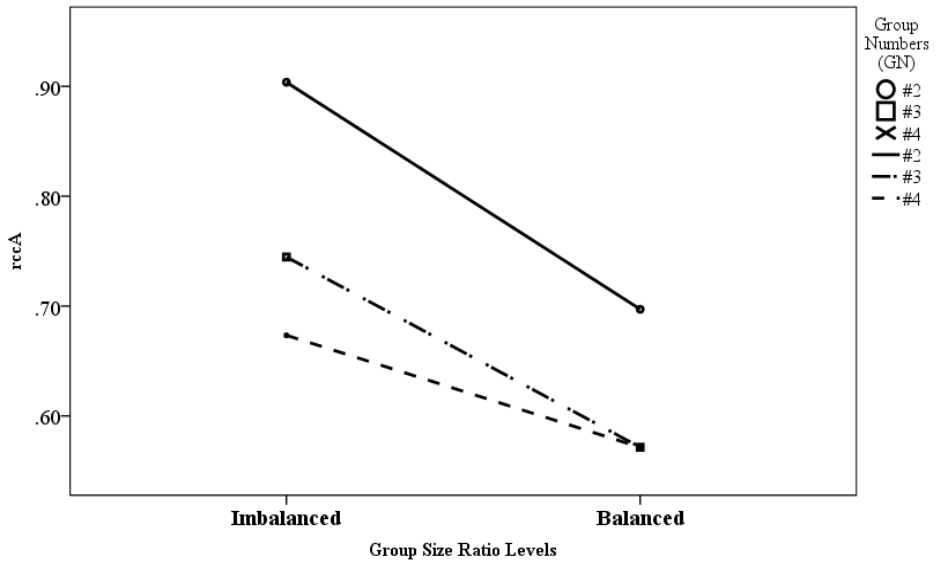

Figure 18

Mean rccAs for Group Number by Level of GSR

Effect of Method in rccA

The method effect on rccA was significant and large in size ($\eta_p^2 = .702$). Table 30 presents the overall mean rccA for the methods, LDA, LR, and CART.

Table 30.

Overall Mean rccA of LDA, LR, and CART

Method	Mean rccA
LDA	.655
LR	.681
CART	.745

LDA with .655 rccA showed the overall lowest performance, LR with .681 rccA was the second, and CART was the best overall performing method with .744 rccA.

While the overall results showed that CART was the best performing method of group membership prediction of the three methods evaluated in this research project, it should be noted that the superiority of CART may not be the case for all the conditions. Of 36 controlled conditions, in two conditions LR showed higher performance than CART and LDA and, these conditions are presented in Table 31. In all other conditions, CART showed better performance than LDA and LR in terms of mean rccA. In one of the conditions, LR showed a 1.8% better performance than LDA and 4.6% better performance than CART in terms of mean rccA. In the other condition, LR showed a 3.4% better performance than LDA and a 7.1% better performance than CART in terms of mean rccA. And, LDA outperformed CART in these two conditions as well, by 2.8% and 3.7%.

Table 31.

Conditions in Which LR Performed Better Than Other Methods

Corr	NPV	GN	GSR	LDA rccA	LR rccA	CART rccA
.2	#10	#4	Imbalanced	.790	.808	.762
.5	#10	#4	Balanced	.799	.833	.762

In addition to the cases where LR performed better, there were conditions in which the prediction accuracy between the methods were trivial (difference being less than 1%) and they are presented in Table 32. These conditions were for the cases when group size ratios were imbalanced and the outcome variable was binary (GN = 2).

Table 32.

Conditions in Which Performance Differences between Methods were Trivial

Corr	NPV	GN	GSR	LDA rccA	LR rccA	CART rccA
.2	#2	#2	Imbalanced	.900	.900	.903
.2	#5	#2	Imbalanced	.902	.902	.908
.2	#10	#2	Imbalanced	.906	.906	.913
.5	#2	#2	Imbalanced	.900	.900	.903
.5	#5	#2	Imbalanced	.901	.901	.907
.5	#10	#2	Imbalanced	.902	.902	.910

A box plot of overall mean rccA of the methods is presented as Figure 19. As can be seen, while overall comparable performance between LDA and LR was observed, CART had a superior overall performance.

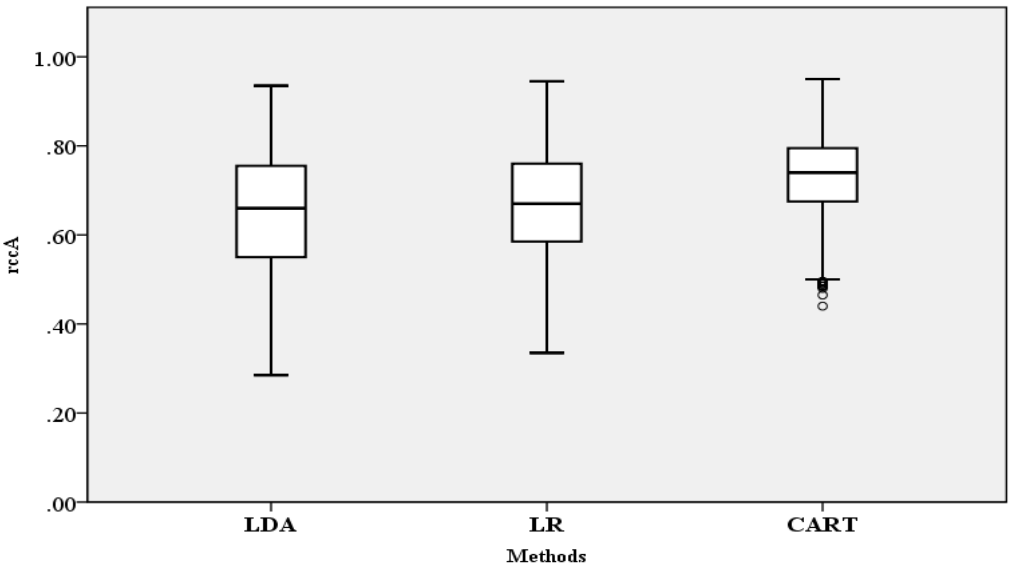

Figure 19

Box-plot for Performance of the Methods on Mean rccA

Effect of Correlation in rccA

The correlation effect on rccA was significant and large in size ($\eta_p^2 = .502$). Table 33 presents the overall mean rccA by level of correlation, .2 and .5.

Table 33.

Overall Mean rccA by Level of Correlation

Corr	Mean rccA
.2	.718
.5	.669

There was a 4.8% difference between the levels of correlation in rccA. The overall prediction power of the methods was higher for the cases when correlations between predictor variables were low than the cases when correlations were medium. However, in some controlled conditions, the difference was trivial (less than 1%) and these are presented in Table 34.

Table 34.

Conditions in which the Difference between Correlation Levels in Mean rccA was Trivial

Methods	NPV	GN	GSR	Corr .2 rccA	Corr .5 rccA
LDA	#2	#2	Imbalanced	.900	.900
LDA	#2	#3	Imbalanced	.714	.710
LDA	#5	#2	Imbalanced	.902	.901
LDA	#10	#2	Imbalanced	.906	.902
LR	#2	#2	Imbalanced	.900	.900
LR	#2	#3	Imbalanced	.715	.709
LR	#5	#2	Imbalanced	.902	.901
LR	#10	#2	Imbalanced	.906	.902
CART	#2	#2	Imbalanced	.903	.903
CART	#2	#2	Balanced	.740	.735
CART	#2	#3	Imbalanced	.747	.747
CART	#2	#3	Balanced	.626	.626
CART	#5	#2	Imbalanced	.907	.908
CART	#5	#3	Imbalanced	.772	.772
CART	#10	#2	Imbalanced	.913	.910

A box plot for overall mean rccA by level of correlation is provided as Figure 20. Overall mean rccA was higher at the low level of correlation.

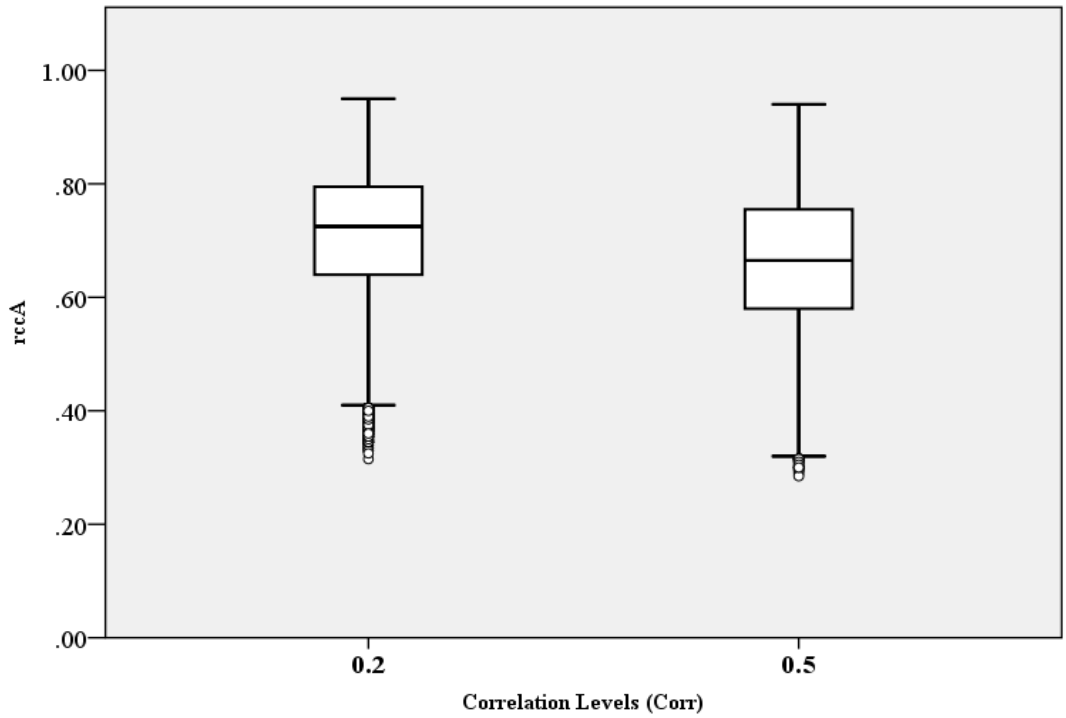

Figure 20

Box-plot for rccA by Level of Correlation

Effect of Number of Predictor Variables in rccA

The effect of NPV on rccA was significant and large in size ($\eta_p^2 = .683$). Table 35 presents the overall mean rccA by NPV.

Table 35.

Overall Mean rccA by Number of Predictor Variables

NPV	Mean rccA
#2	.649
#5	.695
#10	.737

With a higher number of predictor variables, the overall mean rccA was higher. However, in some conditions the effect of number of predictor variables was ignorable; they are presented in Table 36. In every binary imbalanced case, the effect of number of predictor variables was ignorable.

Table 36.

Conditions in Which the Effect of NPV on rccA was Trivial

Method	**Corr**	**GN**	**GSR**	**NPV #2 rccA**	**NPV #5 rccA**	**NPV #10 rccA**
LDA	.2	#2	Imbalanced	.900	.902	.906
LDA	.5	#2	Imbalanced	.900	.901	.902
LR	.2	#2	Imbalanced	.900	.902	.906
LR	.5	#2	Imbalanced	.900	.901	.902
CART	.2	#2	Imbalanced	.913	.908	.903
CART	.5	#2	Imbalanced	.910	.907	.903

While the rccA values in Table 36 showed that the prediction performances of the methods were very high for the binary imbalanced cases, it should be noted that the prediction performance of the methods were actually poor in these conditions. In most cases, the observations from the largest group were predicted correctly and the observations from the smallest were predicted to be in the largest group. One can predict with 90% accuracy without employing any classification technique just by assuming all the observations belong to the larger group. Therefore, the methods did not improve the prediction accuracy for the binary imbalanced cases.

A box plot for rccA by number of predictor variables is presented as Figure 21. A higher rccA at a higher number of predictor variables was observed.

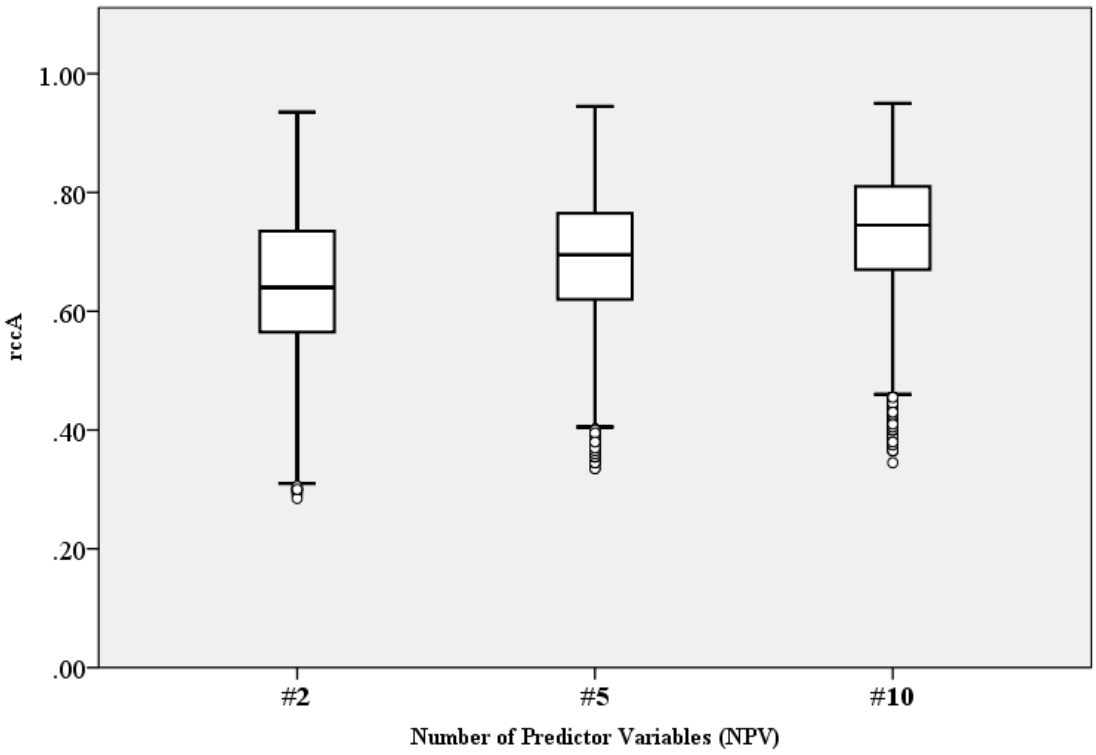

Figure 21

Box-plot for rccA by Number of Predictor Variables

Effect of Number of Groups in rccA

The effect of group numbers on rccA was significant and large in size ($\eta_p^2 =$ .907). Table 37 presents the overall mean rccA by GN.

Table 37.

Overall Mean rccA by Group Number

GN	Mean rccA
#2	.800
#3	.658
#4	.622

The mean rccA when there were two groups was 14.2% higher than the case with three groups and 17.8% higher than the case with four groups. While in general, mean rccA for the cases with two groups was higher than with three or four groups, some results differed (Table 38). Cases where mean rccA for the four-group case was higher than for three-group case or there were trivial differences are presented in Table 39.

Table 38.

Conditions in which Four Groups had the Highest Mean rccA

Method	Corr	NPV	GSR	GN #2 mean rccA	GN #3 mean rccA	GN #4 mean rccA
LDA	.2	#10	Balanced	.715	.609	.799
LR	.2	#10	Balanced	.718	.681	.833

Table 39.

Conditions with Trivial Differences by Level of Group Number and Cases with Four Groups rccA Higher than Three Group rccA

Method	Corr	NPV	GSR	GN #2 mean rccA	GN #3 mean rccA	GN #4 mean rccA
LDA	.2	#2	Balanced	.633	.412	.414
LDA	.2	#5	Balanced	.682	.518	.587
LDA	.2	#10	Imbalanced	.906	.770	.790
LDA	.5	#5	Balanced	.642	.426	.431
LDA	.5	#10	Balanced	.661	.456	.478
LR	.2	#5	Balanced	.682	.606	.625
LR	.2	#10	Imbalanced	.906	.773	.808

Figure 21 presents means rccA when there were two, three, or four groups. It is observed that in general, at lower levels of GN, rccA was higher.

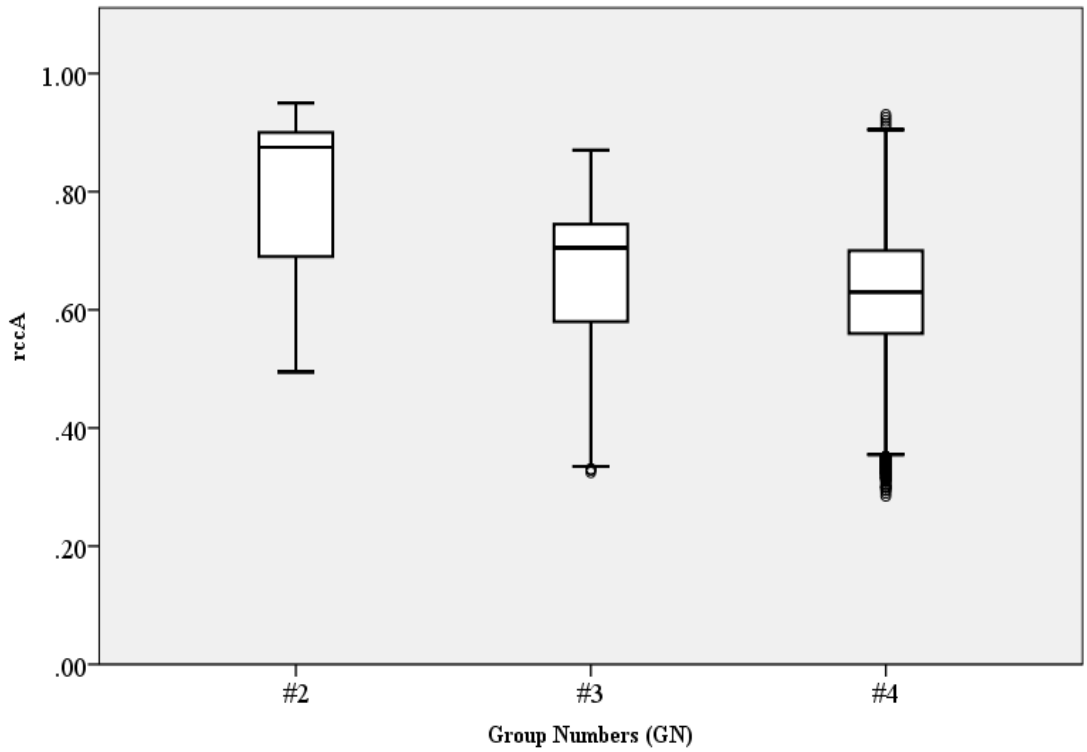

Figure 22

Box-plot for rccA by Level of Group Number

Effect of Group Size Ratio in rccA

The group size ratio effect on rccA was significant and large ($\eta_p^2 = .914$). Table 40 presents the overall mean rccA by level of GSR.

Table 40.

Mean rccA by Level of GSR

GSR	Mean rccA
Imbalanced	.774
Balanced	.613

The imbalanced cases were predicted with 16.1% better accuracy. While in almost every case the imbalanced cases were predicted better than the balanced ones, in two conditions the balanced case was predicted better and they are presented in Table 41.

Table 41.

Conditions in Which Balanced Data was Predicted Better than Imbalanced Data

Method	Corr	NPV	GN	GSR Imbalanced mean rccA	GSR Balanced mean rccA
LDA	.2	#10	#4	.790	.799
LR	.2	#10	#4	.808	.833

Figure 23 presents means rccA for the cases when the groups in dependent variable were imbalanced and balanced. In general, the cases with imbalanced group sizes were predicted with higher accuracy.

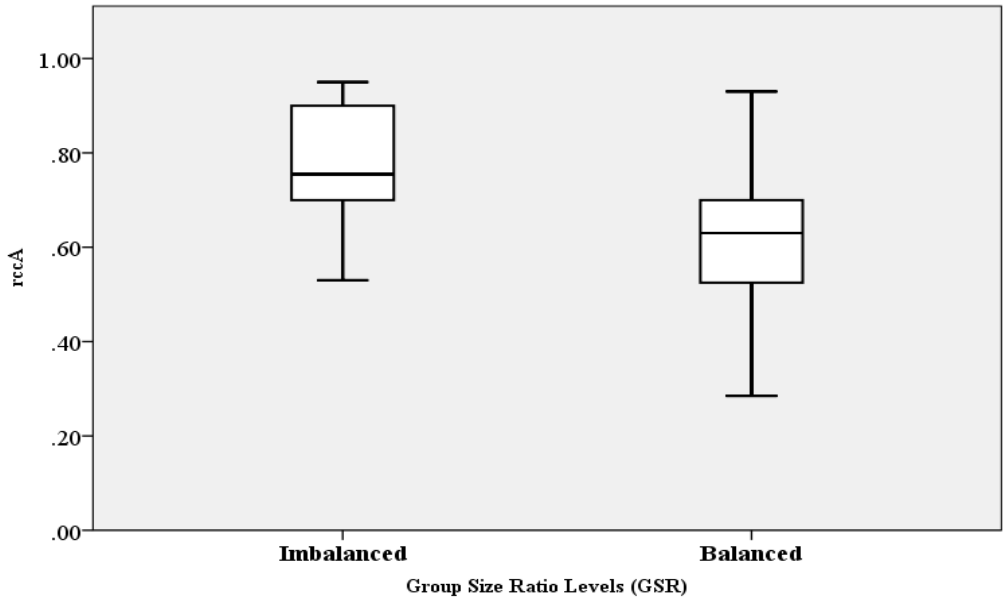

Figure 23

Box-plot for rccA by Level of Group Size Ratio

Results for rccS

As prediction of all groups is important, the prediction of the smallest group in terms of sample size may be equally important when data are imbalanced. For example, if we wish to predict school completion for students who come from the most underrepresented minority group in a school district, it is important to accurately identify the ratio of the underrepresented student group in terms of whole student population in the district and know that prediction of completion for the small group might be difficult. Therefore, in this section accuracy of the smallest group prediction (rccS) is reported just for the cases when data were imbalanced in terms of groups' sample sizes. Therefore, the condition GSR was dropped for rccS, resulting in analysis of the effects of method, Corr, NPV, and GN.

Overview

Based on the factorial ANOVA results, all the interaction and main effects were statistically significant ($p <$. 001). Moreover, the main effects Corr ($\eta_p^2 = .107$), NPV($\eta_p^2 = .25$) and GN ($\eta_p^2 = .637$) were greater than medium in size so they had meaningful effects on rccS. The method effect ($\eta_p^2 = .079$) and all the interactions were statistically significant but smaller than medium in effect size. The overall factorial ANOVA model was statistically significant and had a meaningful partial eta squared value ($p < .001, \eta_p^2 = .712$) for the outcome variable rccS.

While the method effect was the main focus of this study and the partial eta square for it was close to medium in size, in addition to the other main effects, results for

the method effect were also reported. Results for the interaction effects are not reported. Details of the overall factorial ANOVA results for rccS are provided in Table 42.

Table 42.

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

Source	df	*F*	*p*	Partial Eta Squared (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			

Note: Method = Methods (LDA, LR, CART); Corr = correlation levels (.2, .5); NPV = Number of the predictor variables (2,5,10); GN = Number of groups in dependent variable (2,3,4).

Effect of Method in rccS

The method effect on rccS was significant and but the effect was not higher than medium in size ($\eta_p^2 = .079$). According to results, LDA with the .291 of rccS showed the overall lowest performance, LR with .302 rccS was the second, and CART was the best overall performing method with .38 rccS.

Table 43 presents the overall mean rccS for the methods, LDA, LR, and CART.

Table 43.

Overall Mean rccS of LDA, LR, and CART

Method	Mean rccS
LDA	.291
LR	.302
CART	.380

While the overall results showed that CART was the best performing method for smallest group membership prediction of the three methods evaluated in this research project, it should be noted that the superiority of CART was not be the case for all the conditions. Of 18 controlled conditions, in some conditions LR showed higher performance than CART and LDA and these conditions are presented in Table 44. In all other conditions, CART showed at least more than 1% better performance than LDA and LR in terms of mean rccS. It was noticeable that the conditions in which LR performed better than CART were the cases when there were a high number of predictor variables, group numbers, and low correlations between predictor variables.

Table 44.

Conditions in Which LR Performs Better Than Other Conditions in rccS

Corr	NPV	GN	LDA rccS	LR rccS	CART rccS
.2	#5	#3	.392	.417	.382
.2	#5	#4	.623	.642	.601
.2	#10	#3	.533	.561	.525
.2	#10	#4	.636	.828	.781

A box plot for overall mean rccS by method is presented in Figure 24. Similar to rccA results, while overall comparable performance between LDA and LR was observed, CART had a superior overall performance.

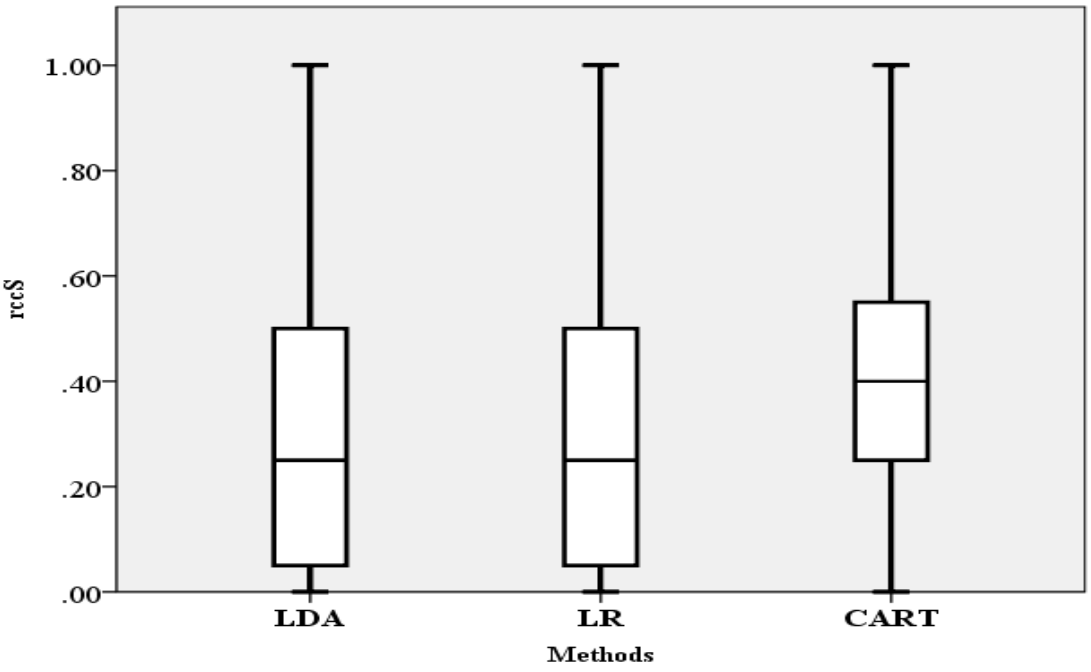

Figure 24

Box-plot for rccS by Method

Effect of Correlation in rccS

The correlation effect on rccS was significant and greater than medium in size ($\eta_p^2 = .107$). Table 45 presents the overall mean rccS by level of correlation, .2 and .5.

Table 45.

Overall Mean rccS Values for Levels of Correlation

Corr	Mean rccS
.2	.371
.5	.278

There was a 9.3% difference between the levels of correlation in rccS. In general, rccS values were higher at the low level of correlation. However, in some controlled conditions, the difference was trivial (less than 1%) and these are presented in Table 46.

Table 46.

Conditions in which the Difference between Correlation Levels in Mean rccS was Trivial

Methods	NPV	GN	Corr .2 rccS	Corr .5 rccS
LDA	#2	#2	.015	.010
LR	#2	#2	.016	.010
CART	#2	#3	.297	.297
CART	#5	#3	.382	.382

A box plot for mean rccS value by level of correlation is provided as Figure 25. Overall mean rccS was higher at the low level of correlation.

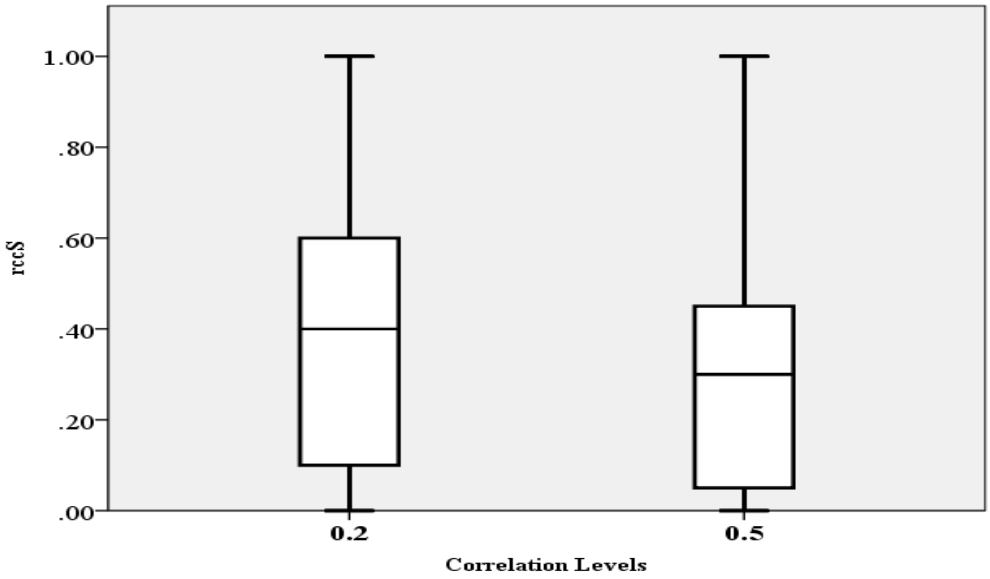

Figure 25

Box-plot for rccS by Levels of Correlation

Effect of Number of Predictor Variables in rccS

The effect of NPV on rccA was significant and large in size ($\eta_p^2 = .252$). Table 47 presents the overall mean rccS by NPV.

Table 47.

Overall Mean rccS by Number of Predictor Variables

NPV	Mean rccS
#2	.226
#5	.329
#10	.418

With a higher number of predictor variables, the overall mean rccS was higher. Between every consecutive level of NPV, the difference in rccS was more than 1%. A

box plot for rccS at the different numbers of predictor variables is presented as Figure 26. Higher rccS at a higher number of predictor variables was observed.

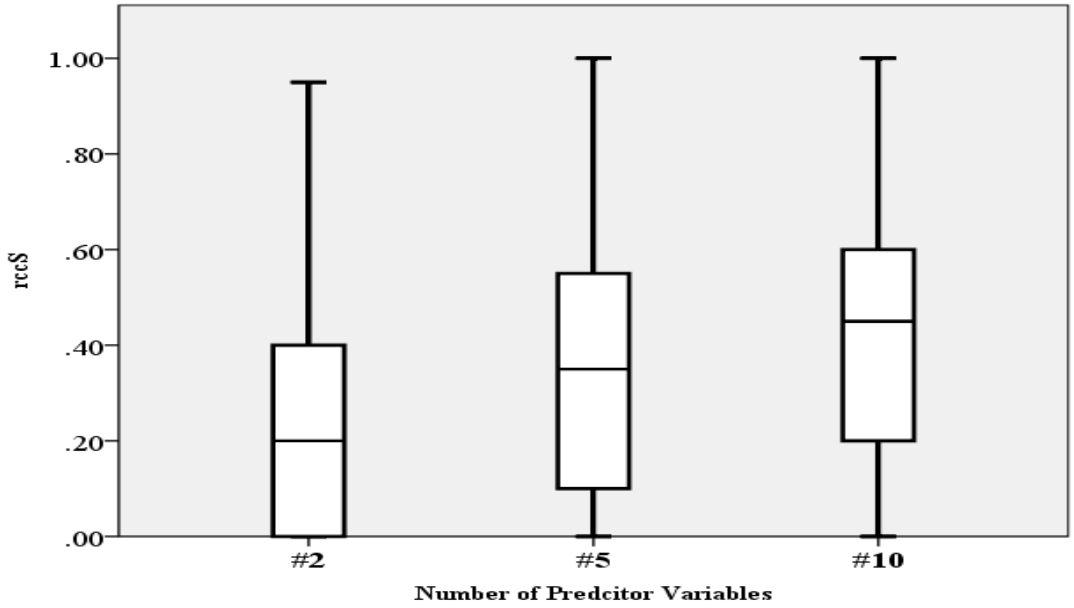

Figure 26

Box-plot for rccS by Number of Predictor Variables

Effect of Number of Groups in rccS

The effect of group number on rccS was significant and large in size ($\eta_p^2 =$ .637). Table 48 presents the overall mean rccS by GN.

Table 48.

Overall Mean rccA by Group Number

GN	Mean rccS
#2	.099
#3	.335
#4	.539

The mean rccS when there were two groups was 44% higher than the case with three groups and 20.4% higher than the case with four groups. Increasing number of groups resulted in an increase in rccS and differences between every consecutive levels of GN were more than 1%. Figure 27 presents means rccS values for the cases when there were two, three, or four groups.

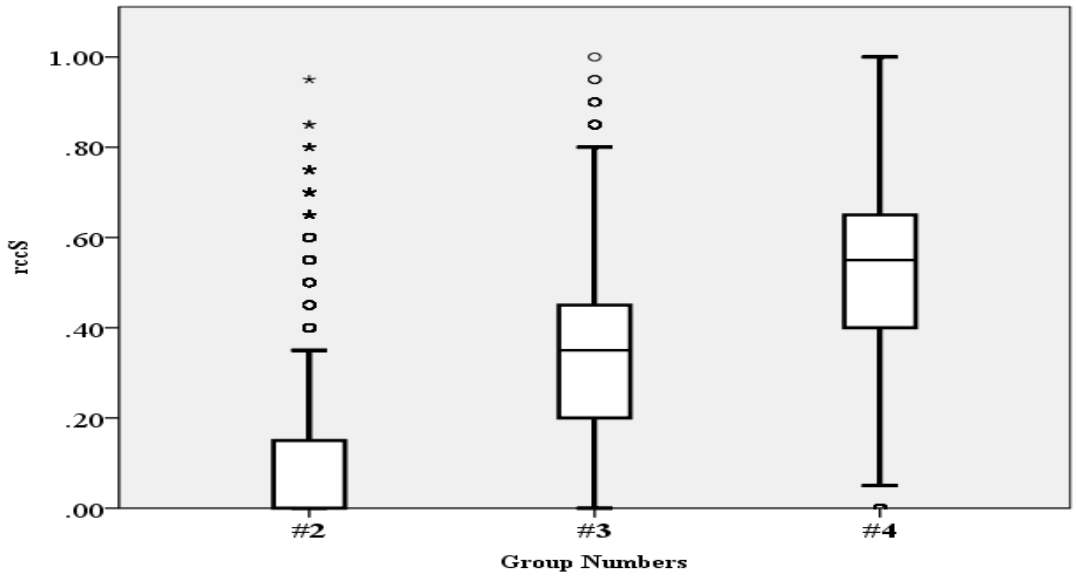

Figure 27

Box-plot for rccS by Group Numbers

Comparison between Results of rccA and rccS

One hundred and eight conditions were evaluated for mean rccA and fifty-four conditions were evaluated for mean rccS (as the balanced cases were dropped for rccS). The mean rccA ranged from .374 to .913 which means that in the condition in which mean rccA was highest, 91.3% of the observations were predicted correctly and in the condition in which mean rccA was the lowest, just 37.4% of the observations were

predicted correctly. On the other hand, the mean rccS values ranged from .010 to .828 and higher variation in rccS than rccA was observed.

One of the important differences between rccA and rccS was the overall mean of rccA and rccS values. While the overall mean for rccA was .694, the overall mean rccS was .324. Thus, there was a 47% difference between prediction accuracy for all groups and for the smallest group. That implies that the methods have weaker abilities to predict smaller groups.

While CART was the overall best performing method in rccA and rccS, in some cases LR performed better than CART or there were trivial differences between these methods for both outcome measures. Especially in cases where correlations between predictor variables were low, and there were a higher number of predictor variables and group numbers, LR had comparable or higher performances than CART for both outcome measures. Overall performance differences in rccA between CART and LR was 6.4% and in rccS it was 7.8%. Moreover, overall performance differences in rccA between CART and LDA was 9% and in rccS it was 8.9%.

The effect size for method was high for rccA, but smaller than medium in size for rccS in terms of partial eta squared. But the methods had greater performance differences on rccS than rccA (the percentages for prediction accuracy), in general. This may be because GSR was included as a factor for rccA but not for rccS. Moreover, having different baselines for rccA and rccS might have affected the main effect of method in terms of partial eta squared.

Both rccA and rccS had higher values at the low compared to the medium level of correlation. For both outcome measures, the effect of correlation was lower for cases

where there were fewer predictor variables. The overall difference in rccA between the cases when the correlation was .2 and .5 was 4.9% and the difference was 9.3% in rccS. Therefore, a greater impact of NPV was observed for rccS than rccA.

Increasing the number of predictor variables yielded an increase in rccA and rccS. However, the contribution of additional predictor variables was more effective for rccS. For example, the difference in mean rccS between the case where there were ten predictor variables and two predictor variables was 19.2% while the difference in rccA was 8.8%. Therefore, the impact of number of predictor variables was greater in rccS.

One of the most noticeable differences between effects on rccA and rccS was in terms of group number. While at higher group numbers mean rccA were lower, mean rccS values were higher. Therefore, the effect of GN had a different direction of impact on rccA and rccS.

Figure 28 is a graphical representation of reactions of rccA and rccS for an increase in number of predictor variables, group number, and correlation between predictor variables.

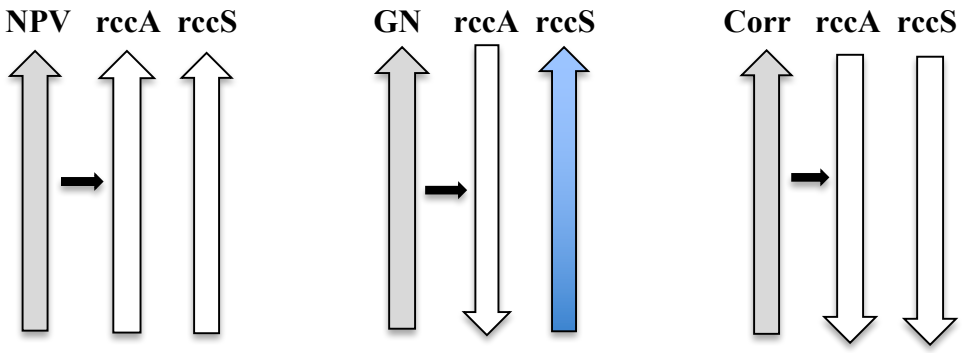

Figure 28

Reactions of rccA and rccS to Increases in NPV, GN, and Corr

For a more detailed comparison of data conditions and methods, ordered mean rccA and rccS values with their conditions are presented in Appendixes B and C.

Chapter Four

Discussion

This chapter summarizes the primary findings, provides an integration of results of this with the literature, addresses the limitations of this study, and provides recommendations for applied researchers and for future study.

Primary Findings Summary

In this study, performances of group membership techniques were assessed. CART was found to have an overall better performance for predicting group membership than both LDA and LR. While two different measures of outcomes were used to evaluate the performances of the methods, CART still showed higher performance rates in most of the controlled data conditions for this study. However, in certain instances of combined conditions (a higher number of predictor variables, group number, and low correlations), LR showed better performance rates than CART. In fact, change in certain conditions (NPV, GN, Corr) led to faster performance improvement for LR than for CART. Therefore, when the study data conditions include having a higher number of predictor variables (10 or more) and number of groups (three, four, or more) in addition to low correlations among predictor variables, superiority of LR might be expected. While the focus of this study was a performance comparison of the classification methods LDA,

LR, and CART under controlled conditions, influences of the controlled conditions were also examined. A discussion regarding the influence of the conditions follows.

All controlled conditions in this study had an influence on prediction accuracy. Moreover, based on partial eta squares, group size ratio was the most influential factor for the prediction of all groups (rccA). The second most influential factor was the group number. Following this was method, number of predictor variables, and level of correlation in that order. Dropping the group size ratio for rccS, group number was the most influential factor. Number of predictor variables was the second most influential factor and correlation was the third most influential factor. Moreover, the method effect was smaller than medium in size for prediction of small group classification accuracy, but the partial eta squared was very close to a medium effect. Therefore, the importance of the method was different for rccA and rccS in terms of the importance rank. But, this should be considered with the reminder that one factor (GSR) was not included for the outcome measure rccS. This could potentially affect the rank order of condition importance. Moreover, this generalization might be specific to the design of this study and in some other designs the rank order of conditions may be different.

Showing the highest influence among all the conditions for rccA, group size ratio is an important element to discuss. The prediction accuracy for the balanced and imbalanced cases were meaningfully different in favor of the imbalanced case. In general, specifically for the binary case that includes imbalanced data, the focus should be prediction accuracy of the smaller group rather than prediction accuracy of all groups. For example, assuming imbalanced data with a group size ratio of 10:90, without applying any statistical procedures, if the researcher makes a decision that all the

observations belong to the larger group, s/he makes a prediction with 90% accuracy for the whole group, but 0% percent accuracy for the smaller group. Thus, accuracy of the smaller group is important to be able to evaluate grouping factors in classification studies. This study showed that the performance difference between the methods for imbalanced and balanced cases were noticeably different from each other. For instance, in the imbalanced case CART performed about 3% better than LR on rccA, but the difference was about 10% for the balanced case.

The factor, group number, had different implications for prediction of all groups and prediction of the smallest group. While increasing group numbers yielded lower classification accuracy for all groups, prediction accuracy for the smaller group increased with higher group numbers. While increasing group numbers from two to four yielded a 17.8% decrease in prediction accuracy of all groups, it led to a 44% increase in prediction accuracy for the smallest group in terms of sample size. Therefore, a great impact of number of groups on small group prediction accuracy was noted.

Increasing the number of predictor variables also yielded higher prediction rates for all groups and for the smallest group. The interaction of method and number of predictor variables was smaller than medium in effect size. On average, the contribution of each additional predictor variable by increasing the number or predictor variables from two to ten for prediction of all groups was about 1.25% for LDA and LR, and about 0.8% for CART. On the other hand, contribution of each additional predictor variable for prediction of the smallest group with LDA was around 2.3%, with LR around 2%, and with CART was around 1.4%, on average. Thus, it was concluded that the influence of additional predictor variables is greater for prediction accuracy of the smallest group.

Prediction accuracy increased more for LDA and LR with additional predictor variables than for CART.

Correlation levels between predictors had also a meaningful effect on prediction accuracy. As expected, at the low level of correlation the prediction accuracies were higher. The rationale is that correlated variables contain similar information, so that the contribution of additional correlated variables is limited while a less correlated variable has the potential to contribute more unique information. The change between correlation levels from .5 to .2 led a higher percentage improvement in prediction accuracy of the small group than prediction accuracy of all groups.

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.

With the rule of thumb used in this study for a medium effect, some three-way and two-way interactions were found to have meaningful effects on prediction accuracies for all groups and these results were reported in Chapter Three. On the other hand, no interactions were found to have meaningful effect on the prediction accuracy of small group prediction. However, while the criterion for a meaningful effect was set as a medium effect size, effects between small and medium could be evaluated. If one wishes to have smaller rule of thumb for a meaningful effect, the interactions method x NPV x GN, NPV x GN, Corr x GN, Corr x NPV, and method x GN could be evaluated as meaningful.

To address the research questions, in this study significant and meaningful effects of the studied conditions (correlation, number of predictor variables, groups numbers, and

group size ratio) were observed on rccA and rccS. Moreover, meaningful and significant interactions of the conditions were observed and reported. While in most of the cases (GN=2,3; NPN<10), CART performed better than the other two methods, in some conditions (GN=4, NPV=10), LR performed better than the other two methods.

Implications for the Literature

Classification techniques and particularly group membership techniques have been useful tools in a variety of research areas. Moreover, while in social sciences and education, applications of traditional classification techniques such as LDA and LR are very common, applications of newer techniques such as CART have been limited.

As many techniques have been developed to predict group membership, interest about which techniques result in better prediction arose. In many applied research studies which focus on prediction of categories, application to several classification techniques at the same time is common. While some of these studies also provided comparisons between accuracies of the techniques, the generalization of the comparison results could not go beyond the content of the research. Therefore, instead of using real data from content areas, some researchers simulated data to compare the effectiveness of techniques. An important advantage of simulated data is that the researcher can control conditions. Thus, using simulated data, some studies compared performances of methods under controlled conditions such as sample size (Bolin & Finch, 2014; Finch et al., 2014; Holden et al., 2011), effect size (Finch & Schneider, 2006; Holden et al., 2011), distribution of variables (Harrell & Lee, 1985; Pai et al., 2012a, b), group size ratio (Lei & Koehly, 2003), and homogeneity of variance-covariance (Fan & Wang, 1998; Kiang, 2003). Under these conditions, though with some conflicting results, in general

comparable performances of LR and LDA were reported (Dey & Austin, 1993; Hess et al., 2011). Moreover, studies showing higher performance of LR than LDA (Baron, 1991) or studies showing better performances of LDA than LR (Williams, 1999) are available in the literature. On the other hand, in simulation studies, CART was generally found to perform better than LDA and LR (Finch et al., 2014; Holden et al., 2011).

In addition to conditions which were studied well, this study evaluated the performances of the methods under different fundamental conditions. Consistent with the existent studies on the performance evaluation of the methods, CART showed an overall higher performance than LDA and LR. However, a new result emerged from this study that the condition with a high number of predictor variables, group number, and low correlations, LR may perform better than CART and LDA. Moreover, greater improvement in prediction accuracy for LDA under certain conditions was observed.

In the previous studies, when comparing LDA with LR, there were studies showing that in general LR performs better than LDA or the two are comparable. Especially in the case where the assumptions for LDA were satisfied, researchers expect similar performances of these two methods (Hastie et al., 2013). The results from the current study also showed that the performances of these two methods were comparable in the case when data are multivariate normal. On the other hand, it should be noted that even though the difference was not large, in almost every case LR performed better, which was also consistent with the extant literature. In the studies which compared CART with LDA or LR, the general conclusion of the studies showed superiority of CART and the current study also showed that in most of the controlled conditions CART performed better than the two other methods. On the other hand, in some controlled

conditions LR had a better performance than CART. As mentioned above, the present study showed under which data conditions the methods have comparable performance and in which the performances differed.

Consistent with previous research, this study found that performance of the methods for overall prediction accuracy decreased with an increase in group number (Finch & Schneider, 2007; Pohar et al.,2004). Moreover, this study found that prediction accuracy for the groups which were the smallest was higher when there were more groups. While few studies investigated the effect of number of predictor variables, the results of this study agreed with Finch and Schneider (2007) that additional predictor variables increase the accuracy of group membership prediction.

Group size ratio was also found to have a meaningful effect on prediction accuracy and prediction accuracy for the whole group can be expected to increase by increasing inequality in groups' sample size proportions. Moreover, a large difference between prediction accuracy for the smaller group and for the whole group was found as most of smallest group prediction accuracies in this study were less than 50% and most of the whole group prediction accuracies were higher than 50%.

Similar to previous findings, correlation had an effect on classification accuracy (Kiang, 2003). Moreover, findings of this study resonate with the comment Pai et al. (2012) made regarding ineffectiveness of multicollinear variables, as at higher levels of correlation the contributions of additional variables were smaller. As the highest level of correlations for this study was .5, at even higher levels less or trivial contributions may be expected. This study also found that the effect of correlation was less for CART than LR and LDA.

While most of the existing studies evaluated performances of the methods based on prediction accuracy for all groups, in this study prediction accuracies for smallest group were also evaluated and different findings for the two outcome measures were obtained. For the instances of highly imbalanced data, very high classification accuracies for all groups and very low prediction accuracies for smaller groups were observed. Therefore, this controversial situation should be noted when evaluating performance of methods, particularly for imbalanced data.

Limitations

Using a fixed standardized mean difference as the degree of consecutive group separation for different group numbers was a limitation of this study. For instance, for the binary case, the difference between two groups was .5 in terms of the standardized group difference, but for the case when there were four groups, the difference between the largest and smallest group in size was 1.5. Therefore, for the case when there were more groups, the group separation between the largest group and the smallest group were higher than for the binary case. Therefore, the result for rccS, which was that rccS was higher for the case with more groups may not be generalizable to all conditions since degree of group separation is an important factor in obtaining higher rccA and rccS. Moreover, since the group differences were fixed, the effect of differential variable importance on group separation was not included in this study as level of variable correlation was set to be equal for all variables. Finally, different levels of group separation such as standardized mean differences less than or more than .5 (medium effect size) were not included in this study.

Another limitation of this study was regarding the ratios for imbalanced data. While in this study, the percentage of the smallest group in terms of sample size was 10%, smaller or higher ratios were not used in this study. Moreover, with multiple group numbers, the ratios were fixed and other scenarios with different ratios were not included. Furthermore, while negative correlations between predictor variables are common in application, negative correlations were not included in this study.

Due to the complexity of having many controlled conditions, the data were simulated under an assumption of multivariate normality for each category, and this is another limitation of this study. Moreover, conditions such as having categorical predictor variables, multimodality, different sample sizes for all groups, and heterogeneity of variance-covariance matrices were other conditions not included in this study.

Recommendations for Applied Researchers

Recommendations of this study for applied researchers can be categorized into two themes. The first theme is regarding choice of optimal method for different conditions and the second is how to increase prediction accuracy.

It is recommended that, in general, practitioners apply CART rather than LDA and LR for their data analysis when the number of predictor variables is less than 10, the number of groups is less than four, and medium or higher level of correlations are found between predictor variables. The methods may perform similarly in the case where there are two groups in the dependent variable and the group size ratio is highly imbalanced, so CART, LDA, or LR would be appropriate. On the other hand, under the conditions with 10 or more predictor variables, 3, 4, or more groups, and generally low correlations

between predictor variables, LR might be a better alternative. LDA showed the worst performance under almost every controlled data condition and it had the lowest overall performance, it is not recommended to apply LDA unless the researchers have a particular rationale for its use.

If researchers would like to increase the whole groups' prediction accuracy, having more variables and fewer groups is suggested. If there is a concern about prediction accuracy for the small group, including more predictor variables with low correlations is recommended. Moreover, additional group numbers might increase the smaller group prediction accuracy. (On the other hand, increasing the number of the groups for the case when there are more than four groups may not increase prediction accuracy for small group as these cases were not investigated in this study.) Moreover, balancing techniques such as increasing the sample size of small group with repetitions or applying propensity score analysis techniques might increase accuracy of small group prediction. Finally, as in the previous studies indicated, before applying classification methods, decisions regarding prior probabilities and the cut score for LDA and LR are required.

In general, it is suggested that applied researchers use CART in cases when there are two or three groups and there are fewer than 10 predictor variables for better results for rccA and rccS. On the other hand, when there four or more groups and more than 10 predictor variables with low correlations, LR might be a better alternative.

It is also recommended that courses on regression or multivariate statistics include CART in their content coverage as it was found that in many data conditions it performed

better than the classical group membership methods (LDA, LR) traditionally taught in statistics courses.

Recommendations for Future Study

For better evaluation of prediction of group membership phenomena, applied researchers need to know which methods performs better under different conditions and which conditions influence the accuracies of group membership prediction. While effects of some conditions are well studied, some conditions and their interactions with other conditions have not been studied widely. Therefore, some conditions which have not been studied widely such as predictor variable correlation, number of predictor variables, group number, and group size ratio and their interactions were evaluated in this study for whole group and small group prediction accuracies. However, the interplay of correlation, group number, number of predictor variables, and group size ratio with some other important data conditions such as predictor variables' distributions, sample size, effect size, and homogeneity of variance-covariance matrices should be studied in future research projects.

As this research was limited to a multivariate normal distribution for each category, a fixed sample size, and a fixed degree of group separation, effect of the conditions controlled in this study can be evaluated under different sample sizes, different versions of non-normal data, and different degrees of group separation. Moreover, future studies can use different levels of the data conditions evaluated in this study. Particularly, the number of predictors of 10 or more, the case when group number is more than four, and lower levels of correlation can be examined to investigate if LDA or LR perform better than CART. In this study, an increased number of predictor variables and group

numbers at low correlation levels resulted in the superiority of LR or comparable performance between LR and CART.

To the effect of imbalanced data on prediction accuracy, different group size ratios other than the ones evaluated in this study may be explored in a future study. Additionally, the effect of different levels of correlations and negative correlations and the cases with a mixture of positive and negative correlations should be studied.

As one limitation of this study was that as group number increased, group separation between the smaller and larger groups increased, the findings regarding increasing group numbers resulting in a decrease in rccA and decrease in rccS can be tested for the cases which fix the degree of group separation between the largest and smallest groups for different numbers of groups.

As found in this study and previous studies, small group predictions rates are smaller than larger group prediction rates. Therefore, improvement of statistical techniques or procedures which increase prediction of small group are encourage for researchers who work on the methodological development of statistical techniques. Moreover, application of data balancing techniques such as propensity score analysis and improvement of classification methods for imbalanced data after applying balancing techniques with simulated data can be explored in a future study.

Due to limitations of LDA, categorical variables were not included in this study. A future study can test the effectiveness of LR, CART, loglinear analysis, and other classification techniques which are not limited to continuous variables. Moreover, the dependent variable in this study was categorical; a future study can test performances of the methods for ordinal dependent variables.

Some methods might show better performances in specific content areas. Thus, researchers in the content areas can study which classification techniques perform better in their area. Specifically, it is encouraged that careful analyses of imbalanced data be conducted, as in some fields the smallest group is the focus of interest. For instance, in the social sciences, recognition and understanding of underrepresented populations might be more challenging than of populations with higher representation. Therefore, in a group membership study, the focus should be the prediction accuracy of underrepresented groups rather than all groups. Similar examples in health sciences such as diagnosing an illness can be given.

While this study just investigated LDA, LR, and CART, other classification methods such as neural networks, random forests, C5.0, boosting, generalized additive models, kth nearest neighbor, quadratic and discriminant analysis, etc. can be evaluated under the same controlled conditions in a future study.

While this study evaluated prediction accuracy, and the methods LDA, LR, and CART are also used to make model estimations, a future study should evaluate the efficiency of coefficient estimation of the methods under similar conditions.

Finally, some future studies can use measures other than rccA and rccS to evaluate performance of the methods. For example, Pohar et al. (2004) used indexes they named as C, B, and Q indexes to compare performances of LR and LDA. Moreover, due to complexity of simulated data with multiple iterations and the cases of multiple groups, receiver operating characteristic curves (ROC curves) were not applied in this study to evaluate method performance and a future study can improve the methodologies to apply ROC curves in such situations.

References

Agresti, A. (2002). *Categorical data analysis* (2nd ed. ed.). Hoboken, NJ: Wiley-Interscience.

Ashikaga, T., & Chang, P. C. (1981). Robustness of Fisher's linear discriminant function under two-component mixed normal models. *Journal of the American Statistical Association*, *76*(375), 676-680. Retrieved from jstor.org.

Anderson, T. W. (1984). *An introduction to multivariate statistical analysis* (2nd ed.). New York, NY: Wiley.

Arabie, P., & De Soete, G. (1996). *Clustering and classification*. Singapore: World Scientific.

Arminger, G., Enache, D., & Bonne, T. (1997). Analyzing credit risk data: A comparison of logistic discrimination, classification tree analysis and freeforward networks. *Computational Statistics*, *12*, 293-310.

Bailey, K. D. (1994). In NetLibrary I. (Ed.), *Typologies and taxonomies: An introduction to classification techniques*. Thousand Oaks, CA: Sage Publications.

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.

Berk, R. A. (2016). *Statistical learning from a regression perspective*. New York: Springer.

Betebenner, D. W., Shang, Y., Xiang, Y., Zhao, Y., & Yue, X. (2008). The impact of performance level misclassification on the accuracy and precision of percent at performance level measures. *Journal of Educational Measurement*, *45*(2), 119-137.

Bidelman, W. P. (1957). Spectral classification of stars noted on objective prism plates. P*ublications of the Astronomical Society of the Pacific*, *69*(409), 326-332.

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

Breckenridge, J. N. (2000). Validating cluster analysis: Consistent replication and symmetry. *Multivariate Behavioral Research, 35*(2), 261-285. doi:10.1207/S15327906MBR3502_5

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

Britzke, E. R., Duchamp, J. E., Murray, K. L., Swihart, R. K., & Robbins, L. W. (2011). Acoustic identification of bats in the eastern United States: a comparison of parametric and nonparametric methods. *The Journal of Wildlife Management*, *75*(3), 660-667.

Catts, H. W., Fey, M. E., Zhang, X., & Tomblin, J. B. (2001). Estimating the risk of future reading difficulties in kindergarten children: A research-based model and its clinical implementation. *Language, Speech, and Hearing Services in Schools*, *32*(1), 38-50. doi:10.1044/0161-1461(2001/004)

Chhikara, R. S., & McKeon, J. (1984). Linear discriminant analysis with misallocation in training samples. *Journal of the American Statistical Association*, *79*(388), 899-906.

Clark, D. A., Beck, A. T., & Beck, J. S. (1994). Symptom differences in major depression, dysthymia, panic disorder, and generalized anxiety disorder. *The American Journal of Psychiatry 151(2),205-209*. doi: http://dx.doi.org/10.1176/ajp.151.2.205

Clayton, K., Blumberg, F., & Auld, D. P. (2010). The relationship between motivation, learning strategies and choice of environment whether traditional or including an online component. *British Journal of Educational Technology*, *41*(3), 349-364. https://doi.org/10.1111/j.1467-8535.2009.00993.x

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

Colombet, I., Ruelland, A., Chatellier, G., Gueyffier, F., Degoulet, P., & Jaulent, M. C. (2000). Models to predict cardiovascular risk: comparison of CART, multilayer perceptron and logistic regression. In *Proceedings of the AMIA Symposium* (p. 156). American Medical Informatics Association.

Cook, B. G., Li, D., & Heinrich, K. M. (2015). Obesity, physical activity, and sedentary behavior of youth with learning disabilities and ADHD. *Journal of Learning Disabilities*, *48*(6), 563-576.

Cox, D. R., & Snell, E. J. (1989). *Analysis of binary data* (Vol. 32). Boca Raton, FL: CRC Press.

Craen, S. D., Commandeur, J. J. F., Frank, L. E., & Heiser, W. J. (2006). Effects of group size and lack of sphericity on the recovery of clusters in K-means cluster analysis. *Multivariate Behavioral Research, 41*(2), 127-145. doi:10.1207/s15327906mbr4102_2

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

Dunn, M. W. (2007). Diagnosing Reading Disability: Reading Recovery as a Component of a Response-to-Intervention Assessment Method. *Learning Disabilities: A Contemporary Journal*, *5*(2), 31-47.

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

Eisenbeis, R. A. (1977). Pitfalls in the application of discriminant analysis in business, finance, and economics. *The Journal of Finance*, *32*(3), 875-900. doi: https://doi.org/10.1111/j.1540-6261.1977.tb01995.x

Efron, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association, 70*(352), 892-898. doi:10.1080/01621459.1975.10480319

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.

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

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Human Genetics*, *7*(2), 179-188.

Flowers, C. P., & Robinson, B. (2002). A structural and discriminant analysis of the Work Addiction Risk Test. *Educational and Psychological Measurement*, *62*(3), 517-526. doi: https://doi.org/10.1177/00164402062003008

Friedman, J. H. (1989). Regularized discriminant analysis. *Journal of the American Statistical Association, 84*(405), 165-175. doi:10.1080/01621459.1989.10478752

Ghasemi, A., & Zahediasl, S. (2012). Normality tests for statistical analysis: a guide for non-statisticians. *International Journal of Endocrinology and Metabolism*, *10*(2), 486. doi: 10.5812/ijem.3505

Glaser, B. A., Calhoun, G. B., & Petrocelli, J. V. (2002). Personality characteristics of male juvenile offenders by adjudicated offenses as indicated by the MMPI–A. *Criminal Justice and Behavior*, *29*(2), 183-201.

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

Haase, R. F., Waechter, D. M., & Solomon, G. S. (1982). How significant is a significant difference? Average effect size of research in counseling psychology. *Journal of Counseling Psychology*, *29*(1), 58.

Harrell, F. E., & Lee, K. L. (1985). A comparison of the discrimination of discriminant analysis and logistic regression under multivariate normality. *Biostatistics: Statistics in Biomedical, Public Health and Environmental Sciences*, *1*(1), 333-343.

Harrell, F. E., Lee, K. L., & Mark, D. B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, *15*(4), 361-387. doi: https://doi.org/10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4

Hastie, T. (2009). In Tibshirani R., Friedman, J H (Jerome H) (Eds.), *The elements of statistical learning: Data mining, inference, and prediction* (Second edition, corrected 7th printing. ed.) New York, NY: 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

Holden, J. (2009). The effects of misclassified training data on the classification accuracy of supervised and unsupervised classification techniques (Doctoral dissertation, Indiana University). (UMI No. 3358919)

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

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, *15*(3), 651-674. doi: https://doi.org/10.1198/106186006X133933

Hosmer, D.W. and Lemeshow, S. (1989): *Applied logistic regression.* New York, NY: Wiley.

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

Huberty, C. J. (1975). In Curry A. R. (Ed.), *Linear versus quadratic multivariate classification*. S.l.]: S.l.: Distributed by ERIC Clearinghouse.

Huberty, C. J, & Lowman, L. L. (2000). Group overlap as a basis for effect size. *Educational and Psychological Measurement, 60, 543-563*. doi: https://doi.org/10.1177/0013164400604004

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

IBM Corp. (2013). IBM SPSS Statistics for Windows, Version 22.0. Armonk, NY: IBM Corp. Retrieved from https://www.ibm.com/us-en/marketplace/statistical-analysis-and-reporting

Jo, H., Han, I., & Lee, H. (1997). Bankruptcy prediction using case-based reasoning, neural networks, and discriminant analysis. *Expert Systems with Applications*, *13*(2), 97-108. doi: https://doi.org/10.1016/S0957-4174(97)00011-0

Johnson, R. A., & Wichern, D. W. (1988). *Applied multivariate statistical analysis* (2nd ed.) Englewood Cliffs, NJ: Prentice Hall.

Kapantzoglou, M., Restrepo, M. A., & Thompson, M. S. (2012). Dynamic assessment of word learning skills: Identifying language impairment in bilingual children. *Language, Speech, and Hearing Services in Schools*, *43*(1), 81-96.

Keogh, B. K. (2005). Revisiting classification and identification. *Learning Disability Quarterly*, *28*(2), 100-102. doi: https://doi.org/10.2307/1593603

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

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (Fourth edition. ed.) New York, NY: The Guilford Press.

Krzanowski, W. J. (1977). The performance of Fisher's linear discriminant function under non-optimal conditions. *Technometrics*, 19(2), 191-200.

Kohavi, R. (1995). *The power of decision tables* (pp. 174-189, Working paper). Berlin: European Conference on Machine Learning.

Kurt, I., Ture, M., & Kurum, A. T. (2008). Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. *Expert Systems with Applications*, *34*(1), 366-374.

Lachenbruch, P. A. (1966). Discriminant analysis when the initial samples are misclassified. *Technometrics 8*(1), 657–662. doi: 10.2307/1266637

Lachenbruch, P. A. (1974). Discriminant analysis when the initial samples are misclassified II: non-random misclassification models. *Technometrics 16*(1), 419–424. doi: 10.1080/00401706.1974.10489211

LeBlanc, M., & Tibshirani, R. (1996). Combining estimates in regression and classification. *Journal of the American Statistical Association*, *91*(436), 1641-1650. doi: https://doi.org/10.1080/01621459.1996.10476733

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.

Lei, Y., Nollen, N., Ahluwahlia, J. S., Yu, Q., & Mayo, M. S. (2015). An application in identifying high-risk populations in alternative tobacco product use utilizing logistic regression and CART: a heuristic comparison. *BMC Public Health*, *15*(341), 1-9. doi: https://doi.org/10.1186/s12889-015-1582

Lillvist, A. (2010). Observations of social competence of children in need of special support based on traditional disability categories versus a functional approach. *Early Child Development and Care*, *180*(9), 1129-1142. doi: https://doi.org/10.1080/03004430902830297

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 1*(3), 86-92. doi:10.1027/1614-2241.1.3.86

Mammarella, I. C., Lucangeli, D., & Cornoldi, C. (2010). Spatial working memory and arithmetic deficits in children with nonverbal learning difficulties. *Journal of Learning Disabilities*, *43*(5), 455-468. doi: https://doi.org/10.1177/0022219409355482

Manel, S., Dias, J. M., & Ormerod, S. J. (1999). Comparing discriminant analysis, neural networks and logistic regression for predicting species distributions: a case study with a Himalayan river bird. *Ecological Modelling*, *120*(2-3), 337-347. doi: https://doi.org/10.1016/S0304-3800(99)00113-1

Maroco, J., Silva, D., Rodrigues, A., Guerreiro, M., Santana, I., & de Mendonça, A. (2011). Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. *BMC Research Notes*, *4*(1), 299. doi: https://doi.org/10.1186/1756-0500-4-299

McLachlan, G. J. (1972). Asymptotic results for discriminant analysis when initial samples are misclassified. *Technometrics, 14*(1), 415–422. doi: 10.1080/00401706.1972.10488926

McClachlan, G. J. (1992). *Discriminant analysis and statistical pattern recognition*. New York: John Wiley.

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

Meshbane, A., & Morris, J. D. (1996). Predictive discriminant analysis versus logistic regression in two-group classification problems. Paper presented at the annual meeting of the American Educational Research Association, New York.

Meyers, L. S., Gamst, G., & Guarino, A. J. (2016). *Applied multivariate research: Design and interpretation*. Thousand Oaks, CA: Sage Publications.

Myers, J., Well, A., Lorch Jr, R. (2013). *Research design and statistical analysis*. New York: Routledge.

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). *Applied linear statistical models* (Vol. 4, p. 318). Chicago, IL: Irwin.

Ozasa, K. (2008). The effect of misclassification on evaluating the effectiveness of influenza vaccines. *Vaccine 26*(1), 6462–6465. doi: 10.1016/j.vaccine.2008.06.039

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.

Pai, D. R., Lawrence, K. D., Klimberg, R. K., & Lawrence, S. M. (2012). Experimental comparison of parametric, non-parametric, and hybrid multi-group classification. *Expert Systems with Applications*, *39*(10), 8593-8603.

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.

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.

Press, S. J., & Wilson, S. (1978). Choosing between logistic regression and discriminant analysis. *Journal of the American Statistical Association, 73*(364), 699-705. doi:10.1080/01621459.1978.10480080

Quinlan, J. R. (2014). *C4. 5: programs for machine learning*. San Mateo, CA: Morgan Kaufman Publishers.

Phelps, M. C., Merkle, E. C. (2008). *Classification and regression trees as alternatives to regression.* In Proceedings: 4th Annual Symposium: Graduate Research and Scholarly Projects. Wichita, KS: Wichita State University, p.77-78.

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

Ratner, B. (2017). *Statistical and machine-learning data mining* (Third edition. ed.) Boca Raton, FL: CRC Press.

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.

Remus, W., & Wong, C. (1982). An evaluation of five models for the admission decision. *College Student Journal, 16*(1), 53-59.

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.

Russell, J. A. (2008). A discriminant analysis of the factors associated with the career plans of string music educators. *Journal of Research in Music Education*, *56*(3), 204-219.

Ripley, B. D. (1994). Neural networks and related methods for classification. *Journal of the Royal Statistical Society: Series B (Methodological), 3*(1), 409-456.

Rodriguez, G., Nobili, F., Rocca, G., De Carli, F., Gianelli, M. V., & Rosadini, G. (1998). Quantitative electroencephalography and regional cerebral blood flow: discriminant analysis between Alzheimer's patients and healthy controls. *Dementia and Geriatric Cognitive Disorders*, *9*(5), 274-283.

Sainani, K. L. (2014). Explanatory versus predictive modeling. *PM&R*, *6*(9), 841-844.

Schumacher, M., Rossner, R., & Vach, W. (1996). Neural networks and logistic regression: Part I. *Computational Statistics: Data Analysis*, *21*(1), 661-682.

So, T. S. H. (2003). *Comparisons of Linear Probability Model, Linear Discriminant Function, Logistic Regression, and K-means Clustering in Two-group Prediction* (Doctoral dissertation, Indiana University). (UMI No. 3111935)

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.

Stevens, J. (1996). *Applied multivariate statistics for the social sciences*. Mahwah, NJ: Lawrence Erlbaum.

Suh, S., Suh, J., & Houston, I. (2007). Predictors of categorical at-risk high school dropouts. *Journal of counseling & development*, *85*(2), 196-203.

Swain, S., & Sarangi, S. S. (2013). Study of Various Classification Algorithms using Data Mining. *International Journal of Advanced Research in*, *2*(2), 110-114.

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

Tatsuoka, M. M., & Lohnes, P. R. (1988). *Multivariate analysis: Techniques for educational and psychological research*. Macmillan Publishing Co, Inc.

Timofeev, R. (2004). Classification and regression trees (CART) theory and applications (Master's thesis, Humboldt University). (UMI No. 188778).

Udris, E. M., Au, D. H., McDonell, M. B., Chen, L., Martin, D. C., Tierney, W. M., & Fihn, S. D. (2001). Comparing methods to identify general internal medicine clinic patients with chronic heart failure. *American Heart Journal*, *142*(6), 1003-1009. doi: DOI: https://doi.org/10.1067/mhj.2001.119130

Valentin, L., Hagen, B., Tingulstad, S., & Eik-Nes, S. (2001). Comparison of 'pattern recognition'and logistic regression models for discrimination between benign and malignant pelvic masses: a prospective cross validation. *Ultrasound in Obstetrics & Gynecology*, *18*(4), 357-365.

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.

Wilson, R. L., & Hardgrave, B. C. (1995). Predicting graduate student success in an MBA program: Regression versus classification. *Educational and Psychological Measurement*, *55*(2), 186-195.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Cambridge, MA: Morgan Kaufmann.

Witten, I.H., Frank, E., Trigg, L., Hall, M., Holmes, G. & Cunningham, S.J. (1999). Weka: Practical machine learning tools and techniques with Java implementations. (Working paper 99/11). Hamilton, New Zealand: University of Waikato, Department of Computer Science.

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.

Zigler, E., & Phillips, L. (1961). Psychiatric diagnosis: A critique. *The Journal of Abnormal and Social Psychology, 63*(3), 607-618.

Appendix A

Simulation Code for Some Conditions of This Study

Note: Due to the length of the simulation process, only the code for several conditions is presented here. A complete version of the code is available upon request from the author.

```
##################################################

#Condition 8
#Corr=.2(#1), NPV = 2(#1), GN = 3(#2), GSR = imbalanced(#1), Method = LR(#2)

require(MASS)
require (mvtnorm)
require(nnet)
require(rpart)
require(rpart.plot)

set.seed(1982)
iter <-1000
rates <- matrix(nrow=iter, ncol=3)

for (i in 1:iter)
{
 mu1=c(-.8,-.8)
 mu2=c(-.3,-.3)
 mu3=c(.2,.2)

 sigma=matrix(c( 1,.111,
                 .111, 1),2,2)

 pvar1 = mvrnorm(20, mu1, sigma)
 pvar2 = mvrnorm(40, mu2, sigma)
 pvar3 = mvrnorm(140, mu3, sigma)

 group1 <- c(rep("Group 1",1*20))
 group2 <- c(rep("Group 2",1*40))
 group3 <- c(rep("Group 3",1*140))
```

```
 outcome1 <-data.frame (group1, pvar1)
 outcome2 <-data.frame (group2, pvar2)
 outcome3 <-data.frame (group3, pvar3)

 colnames(outcome1)[1] <- "group"
 colnames(outcome2)[1] <- "group"
 colnames(outcome3)[1] <- "group"

 nsdataset <- rbind( outcome1, outcome2, outcome3)
 g <- runif(nrow(nsdataset))
 dataset <-nsdataset[order(g),]

 #head(dataset)
 #tail (dataset)
 #summary(dataset)
 newdataset <- dataset[c(2:3)]
 #cor(newdataset)
 a <- mean(cor(newdataset))
 meancorr <- ((4*a) - 2)/2
 #print(meancorr)

 #Logistic Regression

 mymodel <- multinom(dataset$group~ dataset$X1+dataset$X2)

 #summary(mymodel)
 #predict(mymodel, dataset)
 #predict(mymodel, dataset, type="prob")

 cm <- table (predict(mymodel),dataset$group)
 rccA <-sum(diag(cm))/sum(cm)
 rccS <-cm[1,1]/20 #use this for imbalanced cases

 rates[i,1] <-rccA
 rates[i,2] <-rccS
 rates[i,3] <-meancorr
}

#Dataset for Factorial ANOVA

Cond <- 08
Method <- 2
Corr <- 1
NPV <- 1
```

```
GN <- 2
GSR <- 1

C08 <-data.frame(Cond, Method, Corr, NPV, GN, GSR, rates)
C08 <-rename(C08, c(X1="rccA", X2="rccS", X3="Corr")
################################################

#Condition 88
#Corr = .5(#2), NPV = 5(#2), GN = 4(#3), GSR = balanced(#2), Method =
LDA(#1)

require(MASS)
require (mvtnorm)
require(nnet)
require(rpart)
require(rpart.plot)

set.seed(1982)
iter <-1000
rates <- matrix(nrow=iter, ncol=3)

for (i in 1:iter)
{
 mu1=c(-.75,-.75, -.75, -.75, -.75)
 mu2=c(-.25,-.25, -.25, -.25, -.25)
 mu3=c(.25,.25, .25, .25, .25)
 mu4=c(.75,.75, .75, .75, .75)

 sigma=matrix(c( 1,.345,.345,.345,.345,
                 .345, 1,.345,.345,.345,
                 .345,.345,1,.345,.345,
                 .345,.345,.345,1,.345,
                 .345,.345,.345,.345,1),5,5)

 pvar1 = mvrnorm(50, mu1, sigma)
 pvar2 = mvrnorm(50, mu2, sigma)
 pvar3 = mvrnorm(50, mu3, sigma)
 pvar4 = mvrnorm(50, mu4, sigma)

 group1 <- c(rep("Group 1",1*50))
 group2 <- c(rep("Group 2",1*50))
 group3 <- c(rep("Group 3",1*50))
 group4 <- c(rep("Group 4",1*50))

 outcome1 <-data.frame (group1, pvar1)
```

```
 outcome2 <-data.frame (group2, pvar2)
 outcome3 <-data.frame (group3, pvar3)
 outcome4 <-data.frame (group4, pvar4)

 colnames(outcome1)[1] <- "group"
 colnames(outcome2)[1] <- "group"
 colnames(outcome3)[1] <- "group"
 colnames(outcome4)[1] <- "group"

 nsdataset <- rbind( outcome1, outcome2, outcome3, outcome4)
 g <- runif(nrow(nsdataset))
 dataset <-nsdataset[order(g),]

 #head(dataset)
 #tail (dataset)
 #summary(dataset)
 newdataset <- dataset[c(2:6)]
 #cor(newdataset)
 a <- mean(cor(newdataset))
 meancorr <- ((25*a) - 5)/20
 #print(meancorr)

 #Linear Discriminant Analysis

 mymodel <- lda(dataset$group ~
dataset$X1+dataset$X2+dataset$X3+dataset$X4 +dataset$X5, prior
=c(.1,.15,.2,.55))

 #summary(mymodel)
 #predict(mymodel, dataset)
 #predict(mymodel, dataset, type="prob")

 cm <- table (predict(mymodel)$class,dataset$group)
 rccA <-sum(diag(cm))/sum(cm)

 rates[i,1] <-rccA
 rates[i,2] <-rccS
 rates[i,3] <-meancorr
}
#Dataset for Factorial ANOVA

Cond <- 88
Method <- 1
Corr <- 2
NPV <- 2
```

```
GN <- 3
GSR <- 2

C88 <-data.frame(Cond, Method, Corr, NPV, GN, GSR, rates)
C88 <-rename(C88,c(X1="rccA", X2="rccS", X3="Corr"))

################################################

#Condition 96
#Corr=.5(#2), NPV = 10(#3), GN = 2(#1), GSR = balanced(#2), Method =
CART(#3)

require(MASS)
require (mvtnorm)
require(nnet)
require(rpart)
require(rpart.plot)

set.seed(1982)
iter <-1000
rates <- matrix(nrow=iter, ncol=3)

for (i in 1:iter)
{
 mu1=c(-.25,-.25, -.25, -.25, -.25,-.25, -.25, -.25, -.25, -.25)
 mu2=c(.25,.25,.25,.25,.25,.25,.25,.25,.25,.25)


 sigma=matrix(c( 1,.47,.47,.47,.47,.47,.47,.47,.47,.47,
                 .47, 1,.47,.47,.47,.47,.47,.47,.47,.47,
                 .47,.47,1,.47,.47,.47,.47,.47,.47,.47,
                 .47,.47,.47,1,.47,.47,.47,.47,.47,.47,
                 .47,.47,.47,.47,1,.47,.47,.47,.47,.47,
                 .47,.47,.47,.47,.47,1,.47,.47,.47,.47,
                 .47,.47,.47,.47,.47,.47,1,.47,.47,.47,
                 .47,.47,.47,.47,.47,.47,.47,1,.47,.47,
                 .47,.47,.47,.47,.47,.47,.47,.47,1,.47,
                 .47,.47,.47,.47,.47,.47,.47,.47, .47,1),10,10)

 pvar1 = mvrnorm(100, mu1, sigma)
 pvar2 = mvrnorm(100, mu2, sigma)

 group1 <- c(rep("Group 1",1*100))
 group2 <- c(rep("Group 2",1*100))
```

```
 outcome1 <-data.frame (group1, pvar1)
 outcome2 <-data.frame (group2, pvar2)


 colnames(outcome1)[1] <- "group"
 colnames(outcome2)[1] <- "group"

 nsdataset <- rbind(outcome1, outcome2)
 g <- runif(nrow(nsdataset))
 dataset <-nsdataset[order(g),]

 #head(dataset)
 #tail (dataset)
 #summary(dataset)
 newdataset <- dataset[c(2:11)]
 #cor(newdataset)
 a <- mean(cor(newdataset))
 meancorr <- ((100*a) - 10)/90
 #print(meancorr)

 #CART
 mymodel <-
rpart(dataset$group~dataset$X1+dataset$X2+dataset$X3+dataset$X4
+dataset$X5+dataset$X6+dataset$X7+dataset$X8+dataset$X9+dataset$X10,
data=dataset, method="class")

 #summary(mymodel)
 cr<- predict(mymodel, dataset, type="class")
 #predict(mymodel, dataset, type="prob")

 cm <- table (cr,dataset$group)
 rccA <-sum(diag(cm))/sum(cm)

 rates[i,1] <-rccA
 rates[i,2] <-rccS
 rates[i,3] <-meancorr
}
#Dataset for Factorial ANOVA

Cond <- 96
Method <- 3
Corr <- 2
NPV <- 3
GN <- 1
```

```
GSR <- 2

C96 <-data.frame(Cond, Method, Corr, NPV, GN, GSR, rates)
C96 <-rename(C96,c(X1="rccA", X2="rccS", X3="Corr"))
```

Appendix B

List of Data Conditions by Ordered Mean rccA Values

Rank	Method	Corr	NPV	GN	GSR	Mean rccA	Condition Number
1	CART	0.2	#10	#2	Imbalanced	.913	39
2	CART	0.5	#10	#2	Imbalanced	.910	93
3	CART	0.2	#5	#2	Imbalanced	.908	21
4	CART	0.5	#5	#2	Imbalanced	.907	75
5	LR	0.2	#10	#2	Imbalanced	.906	38
6	LDA	0.2	#10	#2	Imbalanced	.906	37
7	CART	0.2	#2	#2	Imbalanced	.903	3
8	CART	0.5	#2	#2	Imbalanced	.903	57
9	LR	0.5	#10	#2	Imbalanced	.902	92
10	LDA	0.5	#10	#2	Imbalanced	.902	91
11	LDA	0.2	#5	#2	Imbalanced	.902	19
12	LR	0.2	#5	#2	Imbalanced	.902	20
13	LR	0.5	#5	#2	Imbalanced	.901	74
14	LDA	0.5	#5	#2	Imbalanced	.901	73
15	LR	0.2	#2	#2	Imbalanced	.900	2
16	LDA	0.2	#2	#2	Imbalanced	.900	1
17	LR	0.5	#2	#2	Imbalanced	.900	56
18	LDA	0.5	#2	#2	Imbalanced	.900	55
19	LR	0.2	#10	#4	Balanced	.833	53
20	CART	0.2	#10	#2	Balanced	.811	42
21	LR	0.2	#10	#4	Imbalanced	.808	50
22	CART	0.2	#10	#3	Imbalanced	.806	45
23	LDA	0.2	#10	#4	Balanced	.799	52
24	CART	0.5	#10	#2	Balanced	.795	96
25	LDA	0.2	#10	#4	Imbalanced	.790	49
26	CART	0.5	#10	#3	Imbalanced	.787	99

27	CART	0.2	#5	#2	Balanced	.784	24
28	LR	0.2	#10	#3	Imbalanced	.773	44
29	CART	0.5	#5	#2	Balanced	.772	78
30	CART	0.2	#5	#3	Imbalanced	.772	27
31	CART	0.5	#5	#3	Imbalanced	.772	81
32	LDA	0.2	#10	#3	Imbalanced	.770	43
33	CART	0.2	#10	#4	Imbalanced	.762	51
34	CART	0.2	#2	#3	Imbalanced	.747	9
35	CART	0.5	#2	#3	Imbalanced	.747	63
36	LR	0.2	#5	#3	Imbalanced	.741	26
37	CART	0.2	#2	#2	Balanced	.740	6
38	CART	0.2	#10	#3	Balanced	.740	48
39	LDA	0.2	#5	#3	Imbalanced	.740	25
40	CART	0.5	#2	#2	Balanced	.735	60
41	LDA	0.5	#10	#3	Imbalanced	.733	97
42	LR	0.5	#10	#3	Imbalanced	.731	98
43	CART	0.2	#5	#4	Imbalanced	.728	33
44	CART	0.2	#10	#4	Balanced	.720	54
45	LR	0.5	#5	#3	Imbalanced	.718	80
46	LDA	0.5	#5	#3	Imbalanced	.718	79
47	LR	0.2	#10	#2	Balanced	.718	41
48	CART	0.5	#10	#4	Imbalanced	.717	105
49	LDA	0.2	#10	#2	Balanced	.715	40
50	LR	0.2	#2	#3	Imbalanced	.715	8
51	LDA	0.2	#2	#3	Imbalanced	.714	7
52	LDA	0.5	#2	#3	Imbalanced	.710	61
53	LR	0.5	#2	#3	Imbalanced	.709	62
54	CART	0.2	#5	#3	Balanced	.703	30
55	CART	0.5	#10	#3	Balanced	.701	102
56	CART	0.5	#5	#4	Imbalanced	.695	87

57	LR	0.2	#5	#4	Imbalanced	.694	32
58	LDA	0.2	#5	#4	Imbalanced	.690	31
59	LR	0.2	#5	#2	Balanced	.682	23
60	LDA	0.2	#5	#2	Balanced	.682	22
61	LR	0.2	#10	#3	Balanced	.681	47
62	CART	0.5	#5	#3	Balanced	.673	84
63	CART	0.2	#2	#4	Imbalanced	.670	15
64	CART	0.2	#5	#4	Balanced	.670	36
65	LR	0.5	#10	#2	Balanced	.663	95
66	LDA	0.5	#10	#2	Balanced	.661	94
67	CART	0.5	#2	#4	Imbalanced	.655	69
68	CART	0.5	#10	#4	Balanced	.655	108
69	LR	0.5	#10	#4	Imbalanced	.644	104
70	LR	0.5	#5	#2	Balanced	.643	77
71	LDA	0.5	#10	#4	Imbalanced	.643	103
72	LDA	0.5	#5	#2	Balanced	.642	76
73	LR	0.2	#2	#2	Balanced	.633	5
74	LDA	0.2	#2	#2	Balanced	.633	4
75	CART	0.2	#2	#3	Balanced	.626	12
76	CART	0.5	#2	#3	Balanced	.626	66
77	LR	0.2	#5	#4	Balanced	.625	35
78	LDA	0.5	#2	#2	Balanced	.619	58
79	LR	0.5	#2	#2	Balanced	.619	59
80	CART	0.5	#5	#4	Balanced	.618	90
81	LDA	0.5	#5	#4	Imbalanced	.617	85
82	LR	0.5	#5	#4	Imbalanced	.617	86
83	LDA	0.2	#10	#3	Balanced	.609	46
84	LR	0.2	#2	#4	Imbalanced	.607	14
85	LDA	0.2	#2	#4	Imbalanced	.606	13
86	LR	0.2	#5	#3	Balanced	.606	29

87	LR	0.5	#2	#4	Imbalanced	.590	68
88	LDA	0.5	#2	#4	Imbalanced	.589	67
89	LDA	0.2	#5	#4	Balanced	.587	34
90	CART	0.2	#2	#4	Balanced	.580	18
91	LR	0.5	#10	#3	Balanced	.567	101
92	CART	0.5	#2	#4	Balanced	.562	72
93	LR	0.5	#10	#4	Balanced	.536	107
94	LR	0.5	#5	#3	Balanced	.536	83
95	LR	0.2	#2	#3	Balanced	.521	11
96	LDA	0.2	#5	#3	Balanced	.518	28
97	LR	0.5	#2	#3	Balanced	.498	65
98	LR	0.5	#5	#4	Balanced	.492	89
99	LDA	0.5	#10	#4	Balanced	.478	106
100	LR	0.2	#2	#4	Balanced	.473	17
101	LDA	0.5	#10	#3	Balanced	.456	100
102	LR	0.5	#2	#4	Balanced	.440	71
103	LDA	0.5	#5	#4	Balanced	.431	88
104	LDA	0.5	#5	#3	Balanced	.426	82
105	LDA	0.2	#2	#4	Balanced	.414	16
106	LDA	0.2	#2	#3	Balanced	.412	10
107	LDA	0.5	#2	#3	Balanced	.386	64
108	LDA	0.5	#2	#4	Balanced	.374	70

Appendix C

List of Data Conditions by Ordered Mean rccS Values

Rank	Method	Corr	NPV	GN	Mean rccS	Condition Number
1	LR	0.2	#10	#4	.828	50
2	LDA	0.2	#10	#4	.781	49
3	LR	0.2	#5	#4	.642	32
4	CART	0.2	#10	#4	.636	51
5	LDA	0.2	#5	#4	.623	31
6	CART	0.2	#5	#4	.601	33
7	CART	0.5	#10	#4	.575	105
8	LR	0.2	#10	#3	.561	44
9	CART	0.5	#5	#4	.545	87
10	LR	0.5	#10	#4	.538	104
11	LDA	0.2	#10	#3	.533	43
12	CART	0.2	#10	#3	.525	45
13	CART	0.2	#2	#4	.520	15
14	LDA	0.5	#10	#4	.516	103
15	CART	0.5	#2	#4	.466	69
16	LR	0.5	#5	#4	.462	86
17	LDA	0.5	#5	#4	.450	85
18	CART	0.5	#10	#3	.420	99
19	LR	0.2	#2	#4	.418	14
20	LR	0.2	#5	#3	.417	26
21	LDA	0.2	#2	#4	.404	13
22	LDA	0.2	#5	#3	.392	25
23	CART	0.2	#5	#3	.382	27
24	CART	0.5	#5	#3	.382	81
25	LR	0.5	#2	#4	.354	68

26	LDA	0.5	#2	#4	.341	67
27	LDA	0.5	#10	#3	.330	97
28	CART	0.2	#10	#2	.321	39
29	LR	0.5	#10	#3	.320	98
30	CART	0.2	#2	#3	.297	9
31	CART	0.5	#2	#3	.297	63
32	CART	0.5	#10	#2	.261	93
33	LR	0.5	#5	#3	.240	80
34	LDA	0.5	#5	#3	.227	79
35	CART	0.2	#5	#2	.224	21
36	LR	0.2	#2	#3	.209	8
37	LDA	0.2	#2	#3	.202	7
38	CART	0.5	#5	#2	.188	75
39	LR	0.5	#2	#3	.152	62
40	LDA	0.5	#2	#3	.146	61
41	LR	0.2	#10	#2	.133	38
42	LDA	0.2	#10	#2	.125	37
43	CART	0.2	#2	#2	.111	3
44	CART	0.5	#2	#2	.094	57
45	LR	0.5	#10	#2	.065	92
46	LDA	0.5	#10	#2	.062	91
47	LR	0.2	#5	#2	.055	20
48	LDA	0.2	#5	#2	.053	19
49	LR	0.5	#5	#2	.024	74
50	LDA	0.5	#5	#2	.023	73
51	LR	0.2	#2	#2	.016	2
52	LDA	0.2	#2	#2	.015	1
53	LDA	0.5	#2	#2	.010	55
54	LR	0.5	#2	#2	.010	56