Machine Learning with Applications 12 (2023) 100459

Contents lists available at ScienceDirect

Machine Learning with Applications

journal homepage: www.elsevier.com/locate/mlwa

Shrinkage heteroscedastic discriminant algorithms for classifying multi-class high-dimensional data: Insights from a national health survey

Olushina Olawale Awe [a,*], Natisha Dukhi [b], Ronaldo Dias [a]

[a] *Department of Statistics (IMECC), University of Campinas, Brazil*
[b] *Human Sciences Research Council (HSRC), Cape Town, South Africa*

ARTICLE INFO

Keywords:
Discriminant analysis
High-dimensional data
Heteroscedasticity
Supervised learning
Shrinkage

ABSTRACT

In many practical data applications, there are often a large number of pre-processed heteroscedastic features. Discriminant analysis is a standard statistical learning method that is useful for classifying such multivariate features. It is well known in literature that the Linear Discriminant Analysis (LDA) is quite sub-optimal for the analysis of high-dimensional heteroscedastic data because of the inherent singularity and instability of the within-class variance. However, shrinkage discriminant analysis (SDA) and its variants often perform better due to its robustness against inherent multicollinearity and heteroscedasticity. In this article, we propose some newly modified discriminant classification algorithms based on the SDA and compare their sensitivities with those of other competing algorithms. The empirical application show that the proposed algorithms perform moderately well for datasets with high dimensions and unequal co-variance structures when applied to simulated and nutrition data with inherent heteroscedasticity and outliers. The sensitivity and precision of the algorithms for the target classes ranges from 70%–100%. The balanced accuracy of all the algorithms ranges from 50 to 75% for the three-class problem considered. Heteroscedastic discriminant algorithm performs moderately well with high sensitivity for classifying health data with high and low dimensions.

1. Introduction

Recent technologies have led to the prevalence of datasets with high dimensions and intricate multi-class structures (Zhou, Gao, Ding, & Liu, 2022). High-dimensional data have become more common in many scientific fields because new automated data collection techniques have been developed in recent times and they have elicited new statistical challenges because many datasets have a large number of features. Some data have as many features (p) as there are observations (n), ($n = p$), while some have even more features than observations ($p > n$). Such datasets pose a challenge in data analysis because classical statistical methods of analysis, such as linear regression, logistic regression and Linear discriminant analysis (LDA), cannot be appropriate for them (Thomaz, Kitani, & Gillies, 2006; Zhao, Wang, & Nie, 2018). While the penalized logistic regression and its variants can be extrapolated and used in multi-class linear classification problems, they are rarely used in practice because of the difficulty in interpretation (Gyamfi, Brusey, Hunt, & Gaura, 2017). Discriminant analysis is a popular statistical learning method used in data science to find a linear combination of features that characterizes or separates two or more classes of objects/groups (Dhamnetiya, Goel, Jha, Shalini, & Bhattacharyya, 2022). The LDA focuses on finding a feature subspace that maximizes the separability between the target groups (classes). When ($n < p$), LDA and the quadratic discriminant analysis may be seriously biased because of data heteroscedasticity in a multi-class problem. Discriminant analysis aims to find the low-dimensional representation of data such that data points within the same class will be pulled together while those between different classes are separated as far as possible (Nie, Wang, Wang, Wang, & Li, 2019; Qin, 2018). As an attractive approach to building models with more predictive power than multiple logistic regressions, supervised shrinkage discriminant approach is proposed in this study for classifying high-dimensional, multi-class heteroscedastic data. Machine learning techniques are ideal modeling tools due to their ability to model non-linear and high-dimensional data, with complex multi-domain variables (Hastie, Tibshirani, Friedman, & Friedman, 2009).

Supervised shrinkage discriminant models have the capacity to utilize high-dimensional data such that they can be used to model smaller datasets that have a great number of predictor variables with decreased overfitting. This buttresses the argument for choosing machine learning over traditional statistical methods due to the ease of application and high predictive capacity of machine learning models (Colmenarejo, 2020). Several supervised machine learning tasks in real life can be performed as multi-class classification problems. However, discriminant analysis approaches are well known to learn discriminative features

[*] Corresponding author.
E-mail addresses: oawe@unicamp.br (O.O. Awe), NDukhi@hsrc.ac.za (N. Dukhi), dias@unicamp.br (R. Dias).

https://doi.org/10.1016/j.mlwa.2023.100459
Received 26 October 2022; Received in revised form 28 February 2023; Accepted 28 February 2023
Available online 6 March 2023

in the statistical pattern recognition literature and can be easily extended to multi-class cases (Li, Zhu, & Ogihara, 2006; Nie, Wang, Wang, & Li, 2019). The use of discriminant analysis has not been fully experimented in the data mining literature, especially for multi-class high-dimensional cases (Zhao et al., 2018). In classification problems, the investigators often try to find the decision boundary, which can divide the dataset into different classes. The classification problems can also be divided on the basis of whether to use binary classifier or multi-class classifier. One of the most renowned algorithms for handling multi-class problems is the logistic regression. In logistic regression, there is no requirement about the within-group covariance matrices of the predictors, and most importantly the method is not so sensitive to outliers that arise due to either changes in system behavior, human or instrument error, or through natural deviations.

However, LDA suffers from several drawbacks, wherein, the first drawback is that conventional LDA is incompetent to deal with multi-modal data whose distribution is more complex than the Gaussian distribution. To overcome this issue, Nie, Wang, Wang, Wang, and Li (2019) presented a pairwise formulation of LDA, namely Neighborhood MinMax Projections (NMMP), which attempts to pull the considered pairwise points within the same class as close as possible and push those between different classes separate. Furthermore, LDA also requires sufficient train data to avoid the small sample Size problem (Nie, Wang, Wang, & Li, 2019). Recent literature reveals that several machine learning methods, including discriminant analysis, have been used to analyze obesity prevalence (Adhikary & Ghosh, 2022; Alkhalaf, Yu, Shen, & Deng, 2022; Cheng, Steinhardt, & Ben Miled, 2022; Hammond et al., 2019; Liu, Fang, Zhou, Dou, & Dou, 2022). Other statistical machine learning methods apart from discriminant algorithms have been used to predict obesity in literature (see for instance, Chatterjee, Jha, Kumari, & Chatterjee, 2021; Li, Hong, Hao, Chen, & Huang, 2020; Pang, Forrest, Lê-Scherban, & Masino, 2021; Ramya & Rohini, 2021; Safaei, Sundararajan, Driss, Boulila, & Shapi'i, 2021). However, literature on shrinkage-based discriminant analysis with application to obesity data is scanty in literature, hence the essence of this present study. Due to obesity complexity in adolescence, with major multi-domain factors that influences growth interactions, traditional statistical methods such as linear models display limitations and focus largely on analyses with decreased predictive power. Such models do not possess the ability to deal with high-dimensional data (Liu et al., 2022).

In this work, we explore the use of discriminant analysis for multi-class and binary classification problems. We evaluate the performance of discriminant analysis on a collection of population-based body mass index (BMI) survey dataset and investigate its usage in BMI categorization. Several studies have been conducted to explain the factors causing obesity, citing a range of factors including socioeconomic factors; food prices; the prevalence of restaurants; cigarettes and alcohol intake. Prior research only took into consideration limited factors in their models, using simple statistical analysis or simple classification and association methods (Steyn, Nel, Nantel, Kennedy, & Labadarios, 2006). More so, there is a paucity of adolescent overweight and obesity prediction using machine learning techniques, especially in Africa (Dugan, Mukhopadhyay, Carroll, & Downs, 2015; Hammond et al., 2019; Lingren et al., 2016). Results from this present study will help to identify relevant supervised discriminant machine learning algorithms that are useful for accurately classifying or predicting obesity status for policy implementation.

Hence, the overaching aim of this present work is to propose some modified shrinkage-based algorithms for characterizing multi-class heteroscedastic data with application to a national health survey data and compare their performances with the baseline mixture discriminant algorithm (MDA) and the generalized linear model (GLM) which is a supervised machine learning classification model used in predicting the probability of a target variable. It is the gold baseline standard for comparisons with other machine learning models in literature (Zhang et al., 2009). The logistic regression and multinomial regression used in this study belongs to the family of generalized linear models. After this non-exhaustive introductory section, the rest of this paper is organized as follows. In Section 2, we review the methodology of LDA and some proposed modified shrinkage-based discriminant algorithms for high-dimensional, multi-class data classification. Section 3 briefly describes some proposed algorithms. Section 4 contains some experimental results and discussion, while we conclude the study in Section 5.

2. Methodological framework

Modern multivariable data are often corrupted by heteroscedastic noise because of the heterogeneous nature of data with multivariate features (Li et al., 2020). Heteroscedastic Discriminant Analysis (HDA) is a generalized method for feature space transformation that does not impose the equal within-class covariance assumptions required by the standard LDA. HDA is considered to be the constrained maximum likelihood projection where the loglikelihood of the samples in the projected space is maximized. Heteroscedasticity (non-constant error variances) is assumed to be present in high-dimensional datasets. A modified Fisher-based method is proposed based on the maximum entropy covariance selection that overcomes both the singularity and instability of the within-class scatter matrix of the LDA, especially when the classes are more than two (Gyamfi et al., 2017). To create a discriminant, we model a multivariate Gaussian distribution over a p-dimensional input vector x for each class c_i with unequal variance as

$$N_p(x|\mu,\Sigma) = \frac{1}{(2\pi)^{\frac{p}{2}}}\frac{1}{|\Sigma|^{0.5}}\exp(-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)).$$

The main methodology and purpose of discriminant analysis is to assign an unknown subject to one of k classes on the basis of a multivariate observation $x=(x_1,\dots,x_p)^T$ where p is the number of features present in the study. Suppose the class labels y_i are defined to be integers ranging from $1,\dots,k$, assuming that there are n_k observations in class k, where μ_k and Σ_k are the corresponding mean vector and covariance matrix of the p-dimensional heteroscedastic multivariate normal distribution $x\ N_p(\mu_k,\Sigma_k)$.

Consider a linear feature space transformation

$$y = V'x \tag{1}$$

where x is a p-dimensional sample predictor vector belonging to one of the c_i, classes of response y, and V is a projection matrix. The goal of LDA is to find a projection matrix V resulting in the best possible separation of the classes in the projected space by defining two scatter matrices as follows:

The within-class scatter matrix S_i is the measure of variability of the class samples defined as

$$S_i = \sum_{x\in c_i}(x-\bar{x}_i)(x-\bar{x}_i)' \tag{2}$$

where $\bar{x}_i$ is the mean of the class c_i.

The sum of the within-class sample variance defines the total within-class variance S_w given by

$$S_w = \sum_{i=1}^{c} S_i \tag{3}$$

If n denotes the total number of samples and n_i is the number of samples belonging to each class c_i, then the grand mean is

$$\bar{x} = \frac{1}{n}\sum_{x} x \tag{4}$$

The average of individual classes is given as

$$\bar{x}_i = \frac{1}{n_i}\sum_{x\in c_i} x \tag{5}$$

The between-class sample variance S_b is given by

$$S_b = \sum_{i=1}^{c} n_i(\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})' \tag{6}$$

Using the scatter matrices above, Fisher's criterion is defined as

$$G(V) = \arg\max_{V} \frac{|V'S_bV|}{|V'S_wV|} \tag{7}$$

which is maximized to get a projection vector that provides maximum discrimination of samples in the projection space $V = [v_1, v_2, v_3, \dots, v_{c-1}]$. Where c is the class number. Taking the derivatives of the generalized Fisher's Index G(V) with respect to V in (7) and setting it equal to 0, we obtain:

$$\frac{\frac{d}{dV}G(V) = \frac{d}{dV}(V'S_bV)V'S_wV - \frac{d}{dV}(V'S_wV)V'S_bV}{(V'S_wV)^2} = 0$$

$$= \frac{(2S_bV)V'S_wV - (2S_wV)V'S_bV}{(V'S_wV)^2} = 0$$

$$\implies V'S_wV(S_bV) - V'S_bV(S_wV) = 0$$

$$\implies \frac{(V'S_wV)(S_bV)}{V'S_wV} - \frac{(V'S_bV)(S_wV)}{V'S_wV} = 0$$

$$\implies S_bV - \frac{(V'S_bV)(S_wV)}{(V'S_wV)} = 0$$

which when simplified, becomes

$$S_bV = \lambda S_wV \tag{8}$$

where V is the matrix of eigenvectors and λ is the eigenvalues of $S_w^{-1}S_b$ respectively.

3. Modified shrinkage-based discriminant algorithms

Shrinkage-based LDA is a form of regularized supervised machine learning method used to improve the estimation of co-variance matrices in situations where the number of training samples (n) is smaller than the number of features (p). In situations where the sample size of each group is less than the number of variables, it is clear that regularization and shrinkage discriminant techniques will enhance and improve parameter estimation and group classification (Thomaz et al., 2006). The reason is that the commonly used estimators for the class-specific variances or the pooled variance in LDA can become unstable and therefore reduce the classification accuracy of the discriminant classifier. This is the major motivation behind these algorithms. The shrinkage-based discriminant algorithms considered are based on modifications of (8).

Suppose we multiply both sides of Eq. (8) by S_w^{-1} to give

$$S_w^{-1}S_bV - S_w^{-1}S_wV\lambda = 0$$

$$\implies S_w^{-1}S_bV - IV\lambda = 0$$

which implies

$$(S_w^{-1}S_b)V = V\lambda \tag{9}$$

where $rank(S_b) \leq (c-1)$.

Note that the eigenvectors of $(S_w + S_b)^{-1}S_b$ are the same as that of $S_w^{-1}S_b$ (Thomaz et al., 2006). The performance of the standard LDA can be seriously sub-optimal if there are only a limited number of total training samples (n) compared to the dimension of the feature space p. We consider shrinkage-based algorithms proposed for targeting limited sample high-dimensional data by adjusting regularized discriminant rules to improve the within-variance estimation of the classes. The major motivation behind these proposed algorithms is to use the discriminating information of the null space of the within-class scatter matrix (S_w) to maximize the between class scatter matrix (S_b) when S_w is singular. The main intuition behind the proposed algorithms is to select the most discriminant features in the original sample space when S_w is non-singular (Jiang, Wang, & Leng, 2018). The general procedure is as follows:

1. Calculate the rank (r) of the within-class scatter matrix S_w.
2. If S_w is non-singular, (i.e $r = n$), then the projection matrix V is composed of the eigenvectors corresponding to the largest eigenvalues of $(S_b + S_w)^{-1}S_b$.
3. Otherwise, calculate the eigenvector matrix of the singular within-class scatter matrix S_w.
4. Suppose P is the matrix that spans the S_w null space sub-matrix of λ, then the projection matrix is composed of the eigenvectors corresponding to the largest eigenvalues of $PP'S_b(PP')$.
5. Then the resulting eigenvectors obtained through the transformation PP' are the most discriminant vectors.

3.1. Shrinkage linear discriminant algorithm (SLDA*)

The main idea behind this algorithm is to discard the null space of the between class scatter matrix (S_b) by diagonalizing it first before diagonalizing the within-class scatter matrix (S_w). This process avoids the singularity problems related with the use of the LDA methods in high-dimensional data. The key intuition behind this shrinkage discriminant algorithm is to discard the null space of S_b by diagonalizing it first before S_w as follows:

1. Diagonalize S_b i.e calculate the eigenvector matrix $P'S_bP = \lambda$.
2. Let T be the first q columns of P corresponding to the largest eigenvalues of S_b, where $q \leq rank(S_b)$
3. Calculate $D_b = T'S_bT$, which is the diagonal of the $q * q$ sub-matrix of the eigenvalues matrix λ.
4. Let $Z = TD_b^{-1/2}$ be a whitening transformation of S_b that also reduces its dimensionality. Diagonalize $Z'S_wZ$ by computing $U'(Z'S_wZ)U = D_w$ and
5. Calculate the projection matrix $D_w^{-1/2}U'Z'$.

We propose other versions of shrinkage algorithms as follows:

3.2. Shrinkage discriminant algorithm (SDA*)

This version of shrinkage discriminant analysis proceeds as follows:

1. Find the eigenvectors (θ), and the eigenvalues (λ) of S_p, where $S_p = \frac{S_w}{n-k}$. n is the sample size and k is the number of classes.
2. Obtain the average eigenvalue λ of S_p using $\bar{\lambda} = \frac{1}{n}\sum_{j=1}^{n}\lambda_j = \frac{tr(S_p)}{n}$.
3. Form a new matrix of eigenvalues based on the following largest dispersion values: $\lambda^* = diag[\max(\lambda_1, \bar{\lambda}), \dots, \max(\lambda_n, \bar{\lambda})]$.
4. Form the modified within-class variance $S_w^* = S_p^*(n-k) = (\theta\lambda^*\theta')(n-k)$.
5. The modified LDA is then constructed by replacing S_w with S_w^* in the Fisher's criterion (7).

A shrinkage regularization method is then performed as follows:

5a. Compute a pooled covariance matrix

$$S_p(\omega) = (1-\omega)S_p + \omega\bar{\lambda}I$$

where

$$S_p = \frac{S_w}{n-k}$$

and

$$\bar{\lambda} = \frac{1}{n}\sum_{j=1}^{n}\lambda_j = \frac{tr(S_p)}{n}.$$

5b. Take the shrinkage parameter ω from $0 \leq \omega \leq 1$ which could be selected to maximize the leave-one-out classification accuracy (Thomaz et al., 2006).

5c. Compute the identity matrix multiplier.

This regularization method has the effect of decreasing the larger eigenvalues λ and increasing the smaller ones which counters the bias inherent in sample-based estimation of eigenvalues.

3.3. Heteroscedastic discriminant algorithm (HDA*)

We now proceed with the generalization to HDA. If n denotes the total number of samples and n_i the number of samples belonging to the class c_i, the HDA extension is generated by introducing a modified objective function into (7) which takes into account weighted contributions of the individual classes as:

$$\prod_{i=1}^{c}(\frac{|V'S_bV|}{|V'\Sigma_iV|})^{n_i} = \frac{(|V'S_bV|)^n}{\prod_{i=1}^{c}(|V'\Sigma_iV|)^{n_i}} \tag{10}$$

where Σ_i is the covariance matrix of the class c_i given by

$$\Sigma_i = \frac{1}{n_i}S_i. \tag{11}$$

By taking the logarithm of (10), we obtain the following objective function:

$$\gamma(V) = \sum_{i=1}^{c} -n_i \log|V'\Sigma_iV| + n\log|V'S_bV| \tag{12}$$

which is maximized by using the matrix differentiation,

$$\frac{d}{dv}\gamma(V) = \sum_{i=1}^{c} -2n_i(V'\Sigma_iV)^{-1}V'\Sigma_i + n(V'S_bV)^{-1}VS_b.$$

However, this optimization procedure has no analytical solution. In order to have an analytical solution, one attempt is to diagonalize Eq. (12). For this, we have:

$$G(V) = \sum_{i=1}^{c} -n_i \log|diag(V'\Sigma_iV)| + n\log|(V'S_bV)| \tag{13}$$

which in turn directly maximizes the between-class variance.

3.4. Mixture discriminant analysis (MDA)

The MDA is a popular supervised machine learning algorithm in the DA family that can be viewed as a deeper look into the LDA classifier because it assumes many of the assumptions of LDA such as the equality of covariance matrix between groups, but it permits more variabilities in its assumptions. While the LDA classifier assumes that each class has a single Gaussian distribution, MDA assumes that each class is a mixture of Gaussian subclasses, that is, each class is comprised of a mixture of multivariate Gaussian subclasses. Because of this additional assumption of MDA, it is capable of outperforming LDA and other discriminant algorithms, especially when the distribution of the data points in the classes is complex. Hence, we use it as a standard of comparison in this work. The setup of the MDA is such that if we have n_j training data points from $j \in \{1,\dots,J\}$ population where each of the j is further divided into R_j artificial subclasses $C_{jr}, r = 1,\dots,R_J$ so that $R = \sum_j R_j$ where $n = \sum_j n_j$. Let a data point from an rth subclass from jth population (having prior probability π_j) have a multivariate Gaussian distribution with mean vector μ_{jr}, equal covariance matrix Σ and unknown mixing probability π_{jr}, $\sum_r \pi_{jr} = 1$ estimable from the data points, the mixture density for the population j is given by

$$m_j(x) = P(X = x|G = j) = |2\pi\Sigma|^{-\frac{1}{2}} \sum_{r=1}^{R_j} \pi_{jr} \exp\frac{-D(x,\mu_{jr})}{2} \tag{14}$$

where $D(x,\mu_{jr})$ is the Mahalanobis distance between x and μ_{jr} with respect to Σ. The expectation maximization algorithm is used to estimate π_{jr}, μ_{jr} and Σ (obtained by combining all the classes), and the posterior probability that is used to classify a data point into the jth class when it has the maximum probability given by

$$P(G = j|X = x) \sim \pi_j Prob(x|j) \sim \pi_j \sum_{r=1}^{R_j} \pi_{jr} \exp\frac{-D(x,\mu_{jr})}{2}. \tag{15}$$

More notes on the theoretical framework and applications of MDA are contained in Friedman (1989) and Hastie and Tibshirani (1990).

3.5. Model evaluation metrics

There are several model evaluation metrics that are commonly used in machine learning, such as accuracy, precision, sensitivity, specificity, and F1-score. The performance evaluation metrics used in this study are described as follows:

- Sensitivity: the ability of a test to correctly identify patients with a disease. It is calculated as
$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{16}$$
where TP is true positive and FN is false negative.
- Specificity: the ability of a test to correctly identify people without the disease. It is calculated as
$$\text{Specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}} \tag{17}$$
where TN is true negative, FP is false positive.
- Negative Predictive Value: It is the ratio of subjects truly diagnosed as negative to all those who had negative test results (including patients who were incorrectly diagnosed as healthy).
- Prevalence is the number of cases in a defined population at a single point in time.
- Accuracy measures how accurate the algorithm is. It is obtained as
$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FN} + \text{FP}} \tag{18}$$
- Balanced Accuracy (BAC): It is calculated as:
$$\text{BAC} = \frac{(\text{Sensitivity} + \text{Specificity})}{2} = \tag{19}$$
- F1-Score: This is the weighted average of precision and recall (= sensitivity). It was adopted in this study because it is often considered more useful than the accuracy because it combines both precision and sensitivity. It can be obtained by the formula:
$$\text{F1-Score} = 2 * \frac{\text{Precision} * \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}} \tag{20}$$
To minimize both false positive and false negative outcomes at the same time, precision and sensitivity can be summarized by using the F1-score, where precision is defined as:
$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}.$$

4. Experimental analyses

4.1. Simulation experiment

We perform a simulation experiment to determine the performance of the proposed algorithms. The simulated data is composed of a sample infected with outliers for a multi-class problem as inspired by the works of Li, Jiang, Chen, Xu, and Yu (1999) and Stage, Carter, and Nora (2004). The simulated dataset follows a multivariate normal distribution with variables $(x_1, x{-}2, x_3, x_4, x_5)$ with both multi-level (a1) and binary (b1) variables with different mean vectors mu, correlation matrix (V) and variance–covariance matrix Σ as follows:

$$\mu = E(x) = (0.75, 2.75, 50.55, 0.48, 40.12)' \tag{21}$$

$$\Sigma = E(x-\mu)(x-\mu)' = diag(0.44, 1.09, 8.51, 0.47, 6.12)' \tag{22}$$

$$V = (c(1, .152, 12.096, .043, .109, .152, 1, .400, \dots, .382, 35.103, 1), 5, 5) \tag{23}$$

A plot of the simulated data considered is shown in Fig. 1. We implement the algorithms for simulated data with both multi-class (k = 3) and binary (k = 2) variables. We then report the sensitivities of the algorithms as shown in Fig. 3. The result shows that HDA outperforms all other algorithms in terms of sensitivity of the model with multi-class response.

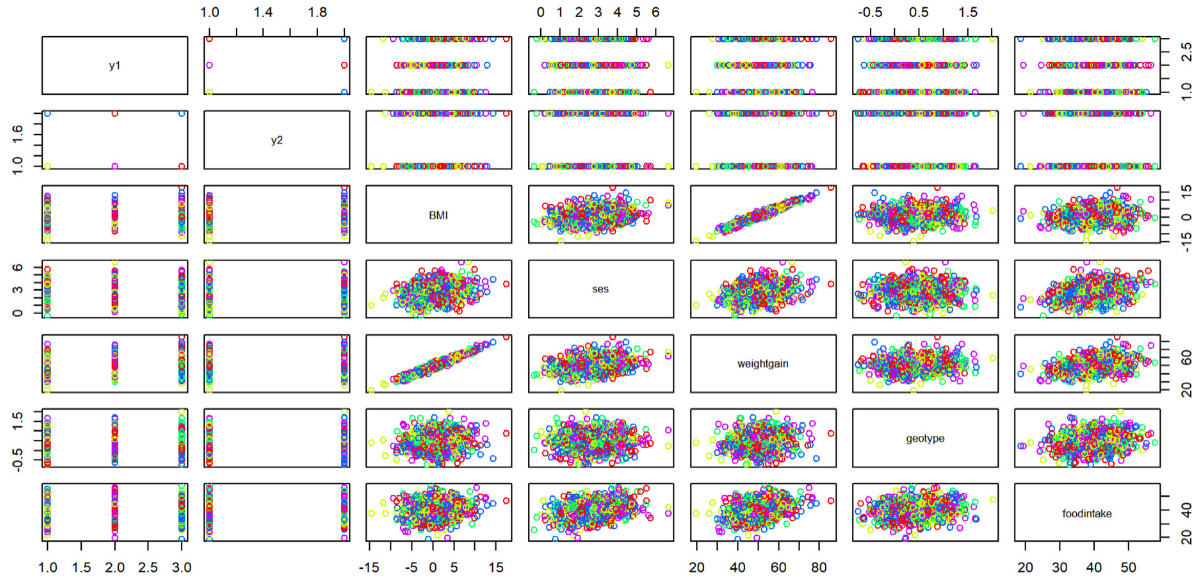

Fig. 1. Simulated data.

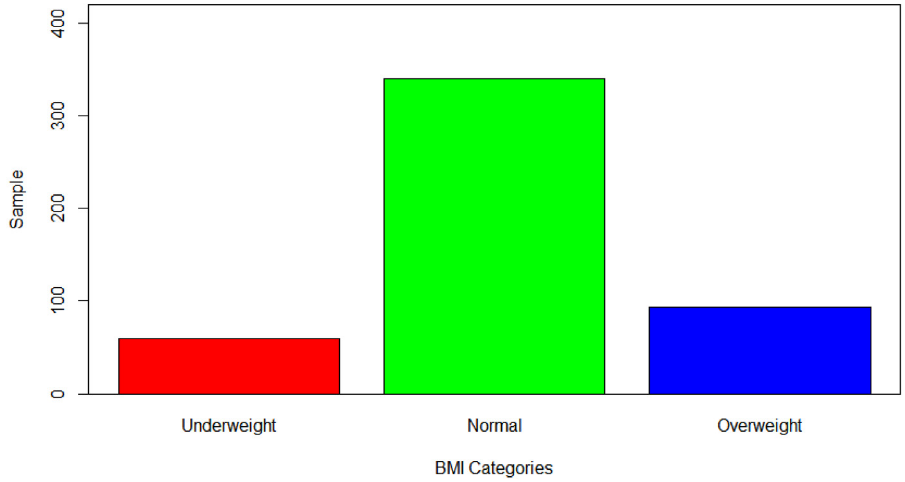

Fig. 2. Outcome variable: BMI.

4.2. Application and insights from SANHANES data

In this section, we report experimental results based on insights from a national health survey data in order to evaluate the performance of shrinkage-based methods. In particular, we apply the algorithms to a population-based nutrition data. The data was extracted from the South African National Health and Nutrition Examination Survey (SANHANES), a cross-sectional national household survey conducted in 2011/12. The survey investigated the health and nutritional status of South African adolescents. In many real life applications, there may be a need to classify a given object under one of a number of distinct classes based on a set of features or variables that describes them. A typical problem is the task considered in this study of classifying body mass index into one of a number of multi-class health states, namely underweight, normal and overweight (see Fig. 2).

4.3. Covariates

In the SANHANES data considered, the primary outcome variable is the body mass index (BMI) categorization of each respondent. Weights and heights were measured using the method adopted by Nieman and Lee (2019). The target variable assigns the patients into three BMI classes namely Underweight, Normal and Overweight. BMI was calculated for all participants as weight (in kg) divided by the square of height (in m; kg/m2). The recommended Centers for Disease Control (CDC) BMI-for-age (indicated as a percentile) cut-offs were used to classify participants as underweight, normal and overweight. All the covariates were categorized into the following domains: socioeconomic, demographic, behavioral risk factors, dietary variables, and family history. The socioeconomic covariates comprised dwelling type, household health insurance, household engagement in meat or poultry agriculture, household food insecurity, household income, household wealth index and having accessed healthcare from a healthcare provider in the past two years. The demographic variables comprised age, province, race group and household locality type. Locality type included urban informal, urban formal, rural informal (typically traditional tribal areas) and rural formal (typically farm areas). The behavioral risk factors include physical activity, weight loss attempts, weight gain attempts, current smoking, high alcohol consumption, and psychological distress. Current smoking was based on self-report and high-risk alcohol use was measured by the AUDIT-C, a 3-item alcohol screening tool.

The dietary variables include nutrition knowledge, dietary diversity, consumption of fat, sugar and fruit and vegetables, red meat consumption, daily milk consumption, daily snack consumption, and preference for fat spreads. The nutrition knowledge score was derived from nine questions assessing knowledge on fiber, fat, sugar and fruit and scores of 0–3 were considered low, 4–6 as medium and 7–9 as high knowledge. A dietary diversity score was calculated using participants recall of the foods and drinks they had the previous 24 hours where a score below four is considered to be low (Steyn et al., 2006). Sugar consumption was assessed by four items on weekly consumption of sweetened beverages, confectionery and sweet snacks. Fruit and vegetable consumption was assessed by four items on the weekly consumption of fresh fruit and vegetables. Ten items assessed weekly fat consumption. The sum scores for each of fat, sugar and fruit and vegetables were categorized into low, moderate and high based on the data distributions of their sum scores. Red meat consumption assessed consumption of red meat with fat removed. Based on the number of meals and snacks consumed per day, three or more meals with snacks in-between were considered high snack frequency. Preference of fat spreads referred to the amount of butter, fat or margarine usually spread on bread or crackers. Blood pressure referred to systolic and diastolic blood pressure, which were measured during the physical examination processes of the survey.

4.4. Results

We utilize the confusion matrix to determine the performance of the proposed discriminant algorithms for classifying incidence of obesity in South Africa using the evaluation indices of balanced accuracy, F1-Score, sensitivity and specificity. The balanced classification accuracy measures the proportion of cases correctly classified by the algorithms, while sensitivity measures the fraction of positive cases that are correctly classified as positive. Specificity, on the other hand, measures the fraction of negative cases that are classified as negative. The algorithm with the highest sensitivity and balanced accuracy or F1-Score is usually considered as the best predictive model for classification. Ten-fold cross-validation with five repeats were executed on the data, using a 70–30 divide for the training and testing set. Each model is automatically tuned and is evaluated using five repeats of 10-fold cross validation. Once the models are trained and an optimal parameter configuration found for each, the accuracy results from each of the best models are collected. Each winning model has 50 results (i.e 5 repeats of 10-fold cross validation). The objective of comparing results is to compare the accuracy distributions between the models. All analyses were done with the *R* Software.

Each machine learning algorithm has several features that processes data in different ways. However, a considerable amount of research in literature have been devoted to modified versions of the shrinkage discriminant algorithms which are capable of handling small samples and high dimensional problems in the presence of heteroscedasticity. Most often, the data that is fed into these algorithms are also different depending on various scenarios and experiments. However, in many practical situations, there is no one model that works best for every data. The assumptions that a great model works well for all problems may not be true because such model may not fit another data well. It is therefore pertinent and peculiar in machine learning to attempt various models or algorithms in order to discover the one that performs best for a specific problem (Dukhi, Sewpaul, Sekgala, & Awe, 2021). In this work, we consider two cases of experiments in order to assess the performance of the proposed shrinkage algorithms. In the first experiment (Case 1), we consider a high-dimensional data (HDD) multi-class case where n = 50, p = 62 and k (number of classes) = 3. In the second case (Case 2), we consider a low-dimensional data (LDD) binary case where n = 671, p = 30 and k = 2.

Table 1
Results from Shrinkage Discriminant Algorithm (SDA*).

Metric	Underweight	Normal	Overweight
Sensitivity (%)	0.00	0.89	0.14
Specificity (%)	1.00	0.22	0.82
Neg. Pred. Value	0.89	0.67	0.60
Prevalence	0.11	0.50	0.39
Detection Rate	0.00	0.44	0.06
BAC (%)	0.50	0.56	0.48

Table 2
Results from Heteroscedastic Discriminant Algorithm (HDA*).

Metric	Underweight	Normal	Overweight
Sensitivity (%)	0.00	1.00	0.29
Specificity (%)	1.00	0.33	0.91
Neg. Pred. Value	0.89	1.00	0.67
Prevalence (%)	0.11	0.50	0.39
Detection Rate	0.00	0.50	0.11
Bal. Accuracy (%)	0.50	0.67	0.60

Table 3
Model comparisons: Sensitivity, Specificity and F1-Score for high-dimensional case.

Model	Sensitivity (%)	Specificity (%)	F1-Score (%)
SDA*	0.941	0.214	0.871
HDA*	1.000	0.000	0.878
GLM	0.927	0.238	0.867
SLDA*	0.967	0.024	0.864
MDA	0.861	0.214	0.828

4.4.1. Case 1: High-dimensional multi-class case

In this instance, we consider the number of respondents to be n = 50 patients with 62 features recorded for each respondent in the survey, which are the explanatory variables (covariates). This makes the data a high-dimensional data. First, we show the table of results of each of the proposed shrinkage and heteroscedastic discriminant algorithms (SDA* and HDA*) for classifying a multi-class high-dimensional problem in Tables 1 and 2. Results in Table 1 shows that SDA* performs well for the multi-class case with a sensitivity of 89% for the normal class with 56% balanced accuracy and a negative predictive value of 67%. Table 2 shows that the HDA* performs well the classification of the normal weight with 100% sensitivity, 100% negative predictive value and 67% balanced accuracy.

We then compare the performance metrics of these shrinkage algorithms with the generalized linear model (GLM), which is often considered as the gold standard approach for modeling public health outcomes such as overweight and obesity (Zhang et al., 2009). Figs. 4–6 shows that the high dimensional discriminant algorithm (HDA*) and the shrinkage discriminant algorithm (SDA*) have the highest accuracies. These models are closely followed by the GLM and SLDA in terms of accuracy. Table 3 shows the results of the comparison of the performance of the shrinkage algorithms with the GLM and MDA for the high-dimensional data.

Fig. 2 shows that SDA* and HDA* have the highest accuracies followed by the GLM and SLDA. In terms of sensitivity, HDA* depicts the highest sensitivity followed by SLDA*, SDA* and GLM for the high-dimensional case (Fig. 3). In Fig. 4, the HDA* also depicts the highest performance in terms of F1-Score, followed by SDA*, GLM and SLDA*. MDA performed lowest in terms of sensitivity and F1-Score.

4.4.2. Case 2: Low-dimensional binary case

In Case 2, which is the low-dimensional case, the HDA is also the highest in terms of accuracy, followed by SLDA and the GLM (see Fig. 7). In Fig. 8, the sensitivity of the HDA is also the highest, followed by the SLDA. The GLM has the lowest sensitivity. In terms of the F1 Score (Fig. 9), the HDA is the best performing model, followed by the SLDA, while the GLM has the lowest F1 Score. Also, the balanced

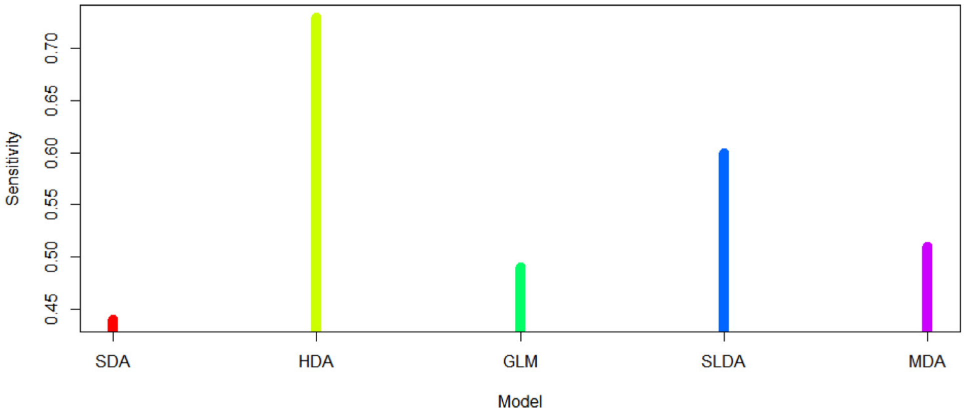

Fig. 3. Comparison of model sensitivities for simulated multi-class outlier infested data.

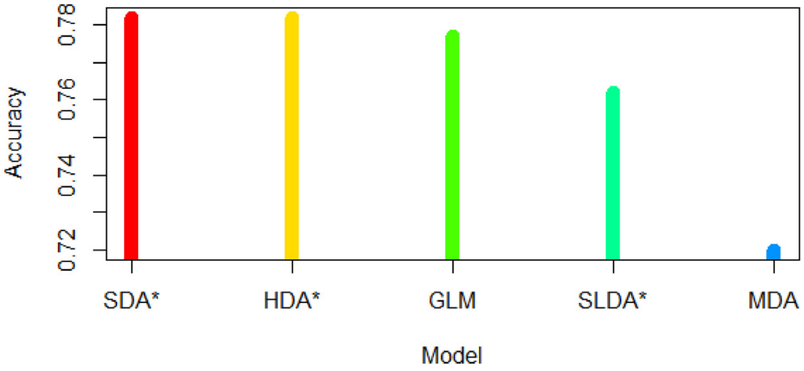

Fig. 4. Case 1 (HDD): Model comparison of Accuracy.

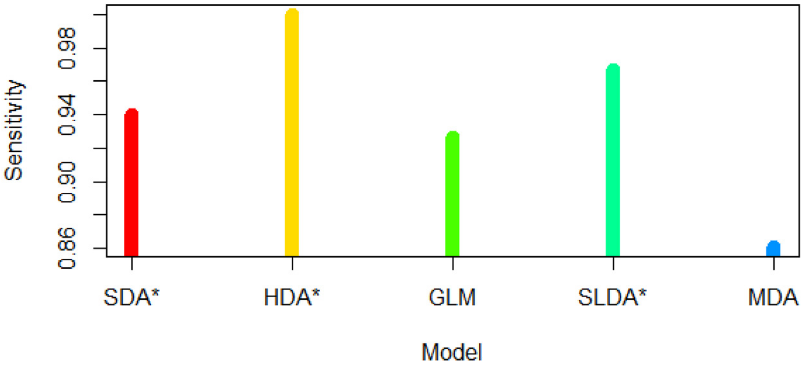

Fig. 5. Case 1(HDD): Model comparison of Sensitivity.

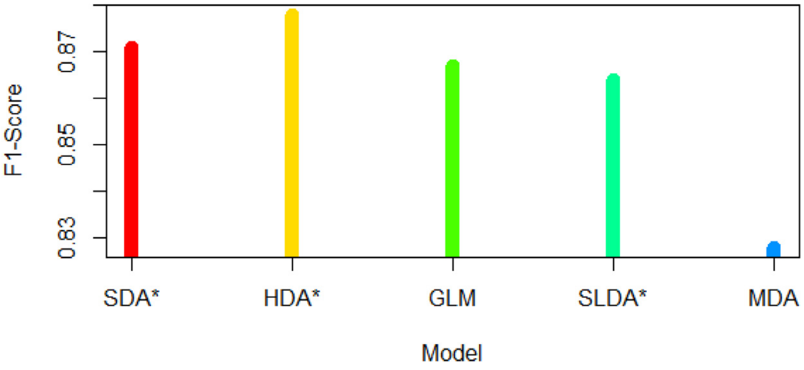

Fig. 6. Case 1 (HDD): Model comparison of F1-Score.

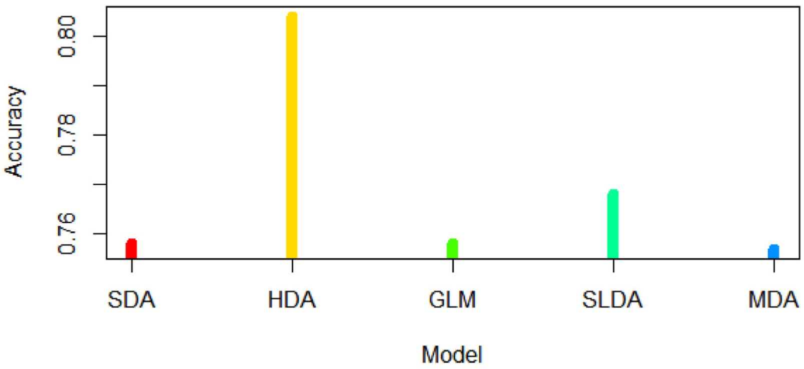

Fig. 7. Case 2 (LDD): Model comparison of Accuracy.

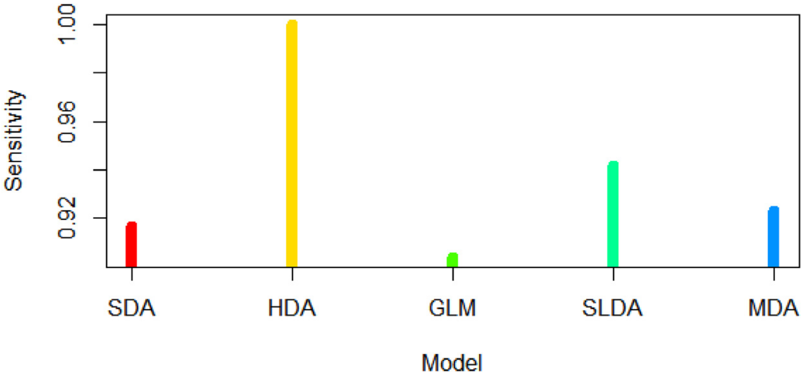

Fig. 8. Case 2 (LDD): Model comparison of Sensitivity.

Table 4
Model comparisons: Sensitivity, Specificity and F1-Score for low-dimensional case.

Model	Sensitivity (%)	Specificity (%)	F1-Score (%)
SDA	0.917	0.105	0.859
HDA	1.000	0.000	0.891
GLM	0.904	0.158	0.857
SLDA	0.942	0.053	0.867
MDA	0.923	0.079	0.859

accuracies of all the algorithms range mostly from 50 to 75% for all classes.

It is evident that in all the model experiments, the HDA, SDA and SLDA outperforms others, for both multi-class and binary cases. They performed better that the GLM and MDA in all cases. The HDA has been shown to be a model-based generalization of linear discriminant analysis (LDA) derived in the maximum-likelihood framework to handle heteroscedastic-unequal variance-classifier models.

Table 4 shows the results of the comparison of the performance of the shrinkage algorithms with the GLM and MDA for the low-dimensional data. HDA has the highest sensitivity, while GLM has the lowest sensitivity. Feature importance was performed with the best performing model to discover the most significant features affecting obesity among adolescents in South Africa. Fig. 10 shows that feature importance using HDA depicts that Gender (Sex), location (geotype), weight gain attempt and family history of diabetes are the most significant factors in predicting obesity status among adolescents in South Africa. It has been reported that overweight and obesity in adults are increasing in South Africa and it is contributing substantially to deaths and disability from non-communicable diseases. In our study, Sex (Gender) is the feature with the highest importance to obesity and overweight as seen in Fig. 10. Compared to men, women suffer more from obesity, which has had adverse affects on their health (Nglazi & Ataguba, 2022).

4.5. Discussion

In literature, obesity is strongly associated with multiple risk factors. It is significantly contributing to an increased risk of chronic diseases globally (Alkhalaf et al., 2022). There are various challenges to better understand the association between the risk factors of obesity and the occurrence of obesity. The traditional regression approach limits analysis to only a small number of predictors and imposes the popular assumptions of normality, independence and linearity. Machine Learning (ML) methods have been used as an alternative approach that provides useful information on data analysis on obesity in this study. The novelty in this present work is that we have proposed modifications to the original shrinkage-based algorithms for high-dimensional heteroscedastic data to further improve their discriminating performance. This is done by adjusting some regularized discriminant rules to improve the within-variance estimation of the classes. These algorithms performs considerably well with the heteroscedastic discriminant algorithm outperforming other algorithms for classifying obesity patients. The heteroscedastic discriminant algorithm (HDA) outperforms other models in terms of sensitivity, accuracy and precision for classifying obesity status among adolescents in South Africa. Feature importance, as shown in Fig. 10, using HDA depicts that Gender (Sex), location (geotype), weight gain attempt and family history of diabetes are the most significant factors for predicting obesity status among adolescents in South Africa. Identifying these risk factors can better inform health authorities in designing or adjusting existing policies for managing and controlling chronic and non-communicable diseases in relation to risk factors associated with obesity.

Moreover, applying ML methods on publicly available health data, such as the SANHANES is a promising strategy to fill the gap for a more robust understanding of the associations of multiple risk factors for predicting public health outcomes. The result contained in this research is a typical example of an intelligent systems application for disease classification and prediction. The present empirical study in

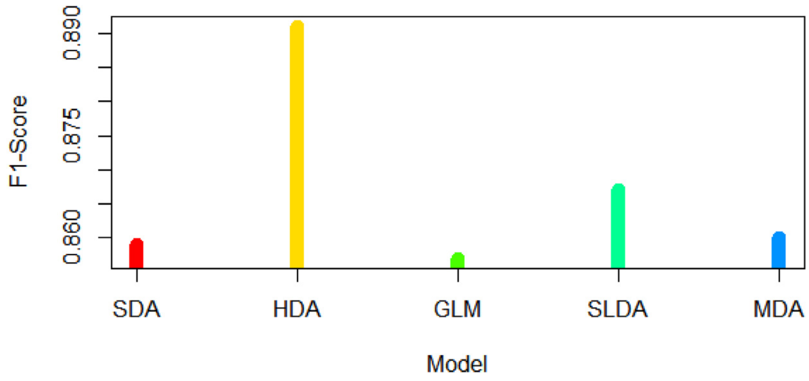

Fig. 9. Case 2 (LDD): Model comparison of F1-Score.

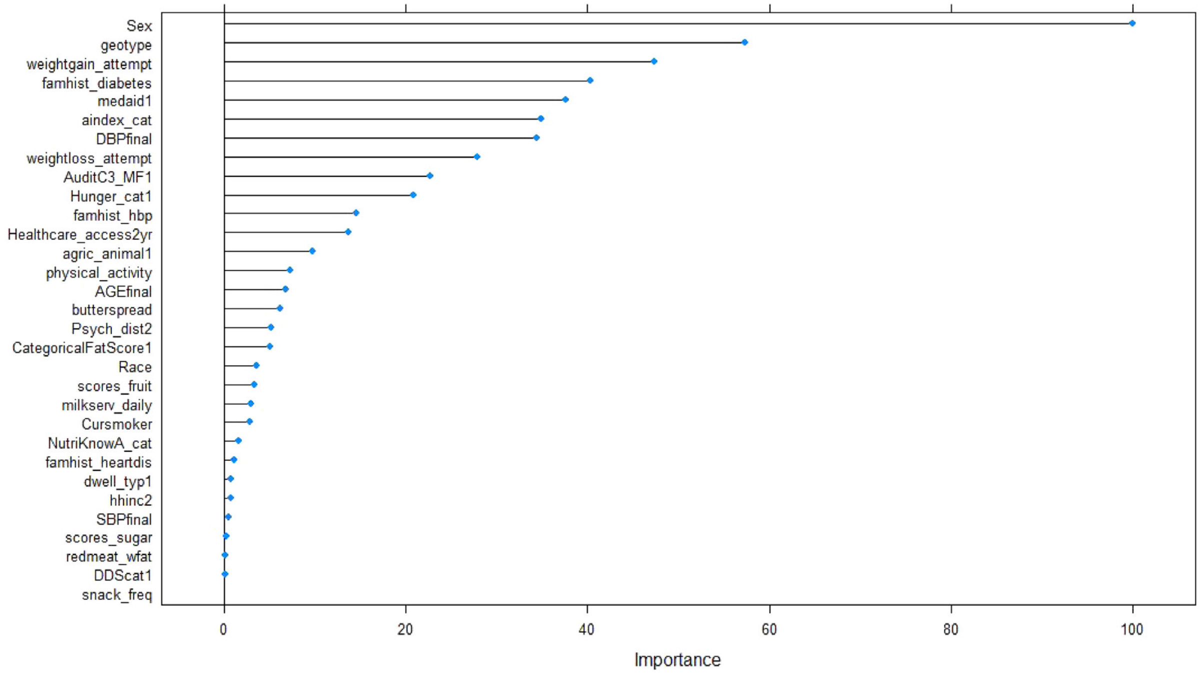

Fig. 10. Variable importance with HDA.

this article show that our proposed algorithms perform moderately well for nutritional datasets with moderate dimensions and unequal co-variance structures. Our future work will focus further on some high-dimensional simulation studies and applications to classification problems involving spatially correlated data in several other fields.

5. Conclusion

Discriminant analysis has been shown in this study to be a standard statistical learning tool for modern data analysis of high and low-dimensional data with multivariate heteroscedastic features. In this work, we have proposed some newly modified statistical learning classification algorithms based on the shrinkage discriminant analysis. We have shown an estimation consistency property of these supervised methods, and compared their performances with a few other competitors like the mixture discriminant algorithm and the generalized linear model for both multi-class and binary scenarios. Our empirical study shows that the proposed machine learning algorithms perform moderately well for datasets with moderate dimensions and unequal co-variance structures when applied to a nutrition (BMI) data. The heteroscedastic discriminant algorithm showed better results on the accuracy, specificity, precision, and F1-Score metrics. Our experiments suggest that shrinkage heteroscedastic discriminant analysis provides a fast, efficient and moderately accurate alternative for general multi-class classification problems with high and low dimensions. Therefore, it is vital that to address the prevention of adolescent obesity, modern predictive models should be developed to identify individuals who are likely at great risk. Such models focus on high-risk populations, while taking on a personalized and cost-effective approach in weight reducing policy interventions.

Applying supervised machine learning methods to public health data can help to improve predictions and find a rich structure among publicly available data in order to increase understanding of complex problems in public health, including risk factors for obesity. The ML methods and applications used in this study could inform the design of more appropriate health policies and programs to address several non-communicable diseases, notably in predicting obesity incidence and prevalence as well as reducing severity and cost of treating overweight and obesity-related conditions which eventually could improve the health and well-being of the populace. Apart from that, the discriminant methods shown in this study could be utilized to identify the most significant risk factors for predicting obesity status. In particular, shrinkage heteroscedastic discriminant algorithms can be applied to publicly available national datasets, such as the SANHANES data.

Finally, machine learning models showed good performance for BMI classification when survey data related to diet and eating habits were used. The algorithms proposed demonstrated that machine learning is a powerful tool that can be used in health research to make policy decisions for timely treatment of people at risk of obesity.

CRediT authorship contribution statement

Olushina Olawale Awe: Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Software, Validation, Writing – original draft, Writing – review & editing. **Natisha Dukhi:** Conceptualization, Investigation, Project administration, Funding acquisition, Validation, Review & editing. **Ronaldo Dias:** Conceptualization, Funding acquisition, Resources, Supervision, Writing – original draft, Writing – review & editing, Quality check.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The authors acknowledge two anonymous reviewers who helped to improve the quality of this article. This work is based on a research support by the National Research Foundation, South Africa (Ref number BRIC190307422517) awarded to Natisha Dukhi. In addition, O. Olawale Awe immensely appreciate FAPESP, Brazil for the financial support (No. 21/05136-4) granted him during the period of this research.

References

Adhikary, S., & Ghosh, A. (2022). E-BMI: A gait based smart remote BMI monitoring framework implementing edge computing and incremental machine learning. *Smart Health*, *24*, Article 100277.

Alkhalaf, M., Yu, P., Shen, J., & Deng, C. (2022). A review of the application of machine learning in adult obesity studies. *Applied Computing and Intelligence*, *2*(1), 32–48.

Chatterjee, K., Jha, U., Kumari, P., & Chatterjee, D. (2021). Early prediction of childhood obesity using machine learning techniques. In *Advances in communication and computational technology* (pp. 1431–1440). Springer.

Cheng, E. R., Steinhardt, R., & Ben Miled, Z. (2022). Predicting childhood obesity using machine learning: Practical considerations. *BioMedInformatics*, *2*(1), 184–203.

Colmenarejo, G. (2020). Machine learning models to predict childhood and adolescent obesity: A review. *Nutrients*, *12*(8), 2466.

Dhamnetiya, D., Goel, M. K., Jha, R. P., Shalini, S., & Bhattacharyya, K. (2022). How to perform discriminant analysis in medical research? Explained with illustrations. *Journal of Laboratory Physicians*.

Dugan, T. M., Mukhopadhyay, S., Carroll, A., & Downs, S. (2015). Machine learning techniques for prediction of early childhood obesity. *Applied Clinical Informatics*, *6*(03), 506–520.

Dukhi, N., Sewpaul, R., Sekgala, M. D., & Awe, O. O. (2021). Artificial intelligence approach for analyzing anaemia prevalence in children and adolescents in BRICS countries: A review. *Current Research in Nutrition and Food Science Journal*, *9*(1), 01–10.

Friedman, J. H. (1989). Regularized discriminant analysis. *Journal of the American Statistical Association*, *84*(405), 165–175.

Gyamfi, K. S., Brusey, J., Hunt, A., & Gaura, E. (2017). Linear classifier design under heteroscedasticity in linear discriminant analysis. *Expert Systems with Applications*, *79*, 44–52.

Hammond, R., Athanasiadou, R., Curado, S., Aphinyanaphongs, Y., Abrams, C., Messito, M. J., et al. (2019). Predicting childhood obesity using electronic health records and publicly available data. *PLoS One*, *14*(4), Article e0215571.

Hastie, T. J., & Tibshirani, R. J. (1990). *Generalized additive models* (p. 335). Chapman and Hall.

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

Jiang, B., Wang, X., & Leng, C. (2018). A direct approach for sparse quadratic discriminant analysis. *Journal of Machine Learning Research*, *19*(1), 1098–1134.

Li, Z., Hong, X., Hao, K., Chen, L., & Huang, B. (2020). Gaussian process regression with heteroscedastic noises—A machine-learning predictive variance approach. *Chemical Engineering Research and Design*, *157*, 162–173.

Li, Y., Jiang, J. -H., Chen, Z. -P., Xu, C. -J., & Yu, R. -Q. (1999). Robust linear discriminant analysis for chemical pattern recognition. *Journal of Chemometrics: A Journal of the Chemometrics Society*, *13*(1), 3–13.

Li, T., Zhu, S., & Ogihara, M. (2006). Using discriminant analysis for multi-class classification: An experimental investigation. *Knowledge and Information Systems*, *10*(4), 453–472.

Lingren, T., Thaker, V., Brady, C., Namjou, B., Kennebeck, S., Bickel, J., et al. (2016). Developing an algorithm to detect early childhood obesity in two tertiary pediatric medical centers. *Applied Clinical Informatics*, *7*(03), 693–706.

Liu, W., Fang, X., Zhou, Y., Dou, L., & Dou, T. (2022). Machine learning-based investigation of the relationship between gut microbiome and obesity status. *Microbes and Infection*, *24*(2), Article 104892.

Nglazi, M. D., & Ataguba, J. E. -O. (2022). Overweight and obesity in non-pregnant women of childbearing age in South Africa: Subgroup regression analyses of survey data from 1998 to 2017. *BMC Public Health*, *22*(1), 1–18.

Nie, F., Wang, Z., Wang, R., & Li, X. (2019). Submanifold-preserving discriminant analysis with an auto-optimized graph. *IEEE Transactions on Cybernetics*, *50*(8), 3682–3695.

Nie, F., Wang, Z., Wang, R., Wang, Z., & Li, X. (2019). Towards robust discriminative projections learning via non-greedy L1-norm MinMax. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *43*(6), 2086–2100.

Nieman, D. C., & Lee, R. (2019). *Nutritional assessment*. United States of America: McGraw-Hill Education.

Pang, X., Forrest, C. B., Lê-Scherban, F., & Masino, A. J. (2021). Prediction of early childhood obesity with machine learning and electronic health record data. *International Journal of Medical Informatics*, *150*, Article 104454.

Qin, Y. (2018). A review of quadratic discriminant analysis for high-dimensional data. *Wiley Interdisciplinary Reviews: Computational Statistics*, *10*(4), Article e1434.

Ramya, A., & Rohini, K. (2021). Comparative evaluation of machine learning classifiers with obesity dataset. In *2021 international conference on computing sciences* (pp. 38–41). IEEE.

Safaei, M., Sundararajan, E. A., Driss, M., Boulila, W., & Shapi'i, A. (2021). A systematic literature review on obesity: Understanding the causes & consequences of obesity and reviewing various machine learning approaches used to predict obesity. *Computers in Biology and Medicine*, *136*, Article 104754.

Stage, F. K., Carter, H. C., & Nora, A. (2004). Path analysis: An introduction and analysis of a decade of research. *The Journal of Educational Research*, *98*(1), 5–13.

Steyn, N. P., Nel, J. H., Nantel, G., Kennedy, G., & Labadarios, D. (2006). Food variety and dietary diversity scores in children: Are they good indicators of dietary adequacy? *Public Health Nutrition*, *9*(5), 644–650.

Thomaz, C. E., Kitani, E. C., & Gillies, D. F. (2006). A maximum uncertainty LDA-based approach for limited sample size problems—with application to face recognition. *Journal of the Brazilian Computer Society*, *12*(2), 7–18.

Zhang, S., Tjortjis, C., Zeng, X., Qiao, H., Buchan, I., & Keane, J. (2009). Comparing data mining methods with logistic regression in childhood obesity prediction. *Information Systems Frontiers*, *11*(4), 449–460.

Zhao, H., Wang, Z., & Nie, F. (2018). A new formulation of linear discriminant analysis for robust dimensionality reduction. *IEEE Transactions on Knowledge and Data Engineering*, *31*(4), 629–640.

Zhou, R., Gao, W., Ding, D., & Liu, W. (2022). Supervised dimensionality reduction technology of generalized discriminant component analysis and its kernelization forms. *Pattern Recognition*, *124*, Article 108450.