International Journal of Scientific Research in Mathematical and Statistical Sciences **Review Paper**
Volume-5, Issue-6, pp.123-129, December (2018) **E-ISSN:** 2348-4519

A Nonparametric Discriminant Stepwise Algorithm for Classification to Two Populations

S. Padmanaban[1], Martin L. William[2]

[1]HRRC Field Unit, Indian Council of Medical Research, KMC Hospital, Chennai, India
[2]Department of Statistics, Loyola College, Chennai, India

Available online at: www.isroset.org

Received: 25/Nov/2018, Accepted: 12/Dec/2018, Online: 31/Dec/2018

Abstract- This paper provides a nonparametric discriminant stepwise algorithm to discriminate two multivariate populations and an optimal decision rule for classification of a member to either of the two populations. This 'two-way-stepwise algorithm' is a combination of the 'forward-stepwise' and the 'backward-stepwise' approaches recently proposed for the same classification problem by Padmanaban and William (2016a, b). As has been done in the above-referred papers, this paper relaxes the 'equal variance-covariance matrices' condition traditionally imposed and develops a discrimination-classification procedure by including variables that contribute to effective discrimination in a 'forward' manner one-by-one and excluding variables that do not contribute to effective 'discrimination' in a 'backward' manner one-by-one. The inclusion of variables in the discriminant is determined on the basis of maximum discriminating ability and exclusion is on the basis of least 'discriminating ability' as reflected in 'difference' between the distributions of the discriminant in the two populations. A decision-rule for classification or membership-prediction with a view to maximizing correct predictions is provided as done in the forward and backward approaches referred above.

The proposed algorithm is applied to develop an optimal discriminant for predicting respiratory tract disease(RD) among newborns of mothers with PPROM in the city of Chennai, India, and its performance is compared with logistic regression.

Keywords-Classification, Discriminant, Kolmogorov-Smirnov Statistic.

I. INTRODUCTION

The question of effective discrimination of objects belonging to two populations and the associated question of 'correct' classification of members to the two groups have been interesting questions for many years now. In the literature on discriminant analysis, the requirement for a non-parametric setting is that the two populations must have equal variance-covariance matrices, while this is not needed for multivariate normal populations. The inclusion of variables in the discriminant is generally carried out by comparing their means in the two populations. Further, the classification of an object to one of the two populations is made based on distances from the centroids of the two populations.

Recently, Padmanaban and William (2016 a, b) introduced two new algorithms to building discriminant models under a non-parametric setting removing the restrictions traditionally imposed. A new optimality criterion was considered and the 'optimal discriminant' was obtained using two different model-building algorithms. The 'Forward Method' or 'Variable Selection Algorithm' proposed in Padmanaban and William (2016 a) starts with one variable that is selected based on 'maximum discriminatory ability' followed by inclusion of variables one-at-a-time based on the same criterion. The 'Backward Method' or 'Variable Elimination Algorithm' proposed in Padmanaban and William (2016 b) starts with 'all' variables and eliminates the variables one-at-a-time based on reduced discriminatory ability. The aim of this paper is to develop a two-way stepwise algorithm for obtaining an optimal discriminant that would have 'good' ability to classify objects to two populations. The algorithm to be presented here combines both the 'forward' and 'backward' procedures and works step-by-step to achieve a discriminant model that would give an effective classification rule by including the significant 'discriminating' variables while at the same time maintaining parsimony of the model. Like the one-way algorithms referred above, the stepwise algorithm has a wider scope of application than the traditional discriminant analysis.

In real-life situations with multivariate data, joint normality or equality of variance-covariance matrices in the two populations is not guaranteed. It is well known that, if the datasets follow the multivariate normal law, the equality of variance-covariance matrices can be tested and if the hypothesis of equality is accepted, one can apply the traditional linear discriminant analysis. If the test results in rejection of the hypothesis of equality, quadratic discriminant function based on normal law can be used. If the variables are not jointly normal, the distribution-free Fisher's discriminant analysis is applicable, subject to the condition that the variance-covariance matrices in the two populations are equal. However, there is no known procedure to test whether this condition is satisfied. Practitioners mostly 'assume' equality and apply the traditional discriminant analysis. This gap between theoretical requirements and practical situations remained unabridged until recently when Padmanaban and William (2016 a) brought about a distribution-free approach removing the traditional restriction on variance-covariance matrices.

In the work referred above, Padmanaban and William (2016 a) developed a discrimination-classification procedure for a non-parametric setting without imposing the equality constraint on the variance-covariance matrices and a model-building algorithm for the inclusion of variables in the discriminant function. This algorithm has been termed 'forward method' in line with the term usually used in predictive modeling. Following this, Padmanaban and William (2016 b) introduced a 'backward method' which involves starting the model with all the candidate variables and eliminating the variables that reduce the discriminating capacity one-by-one until we have only those variables that are capable of effectively discriminating the two populations. In this paper, we propose a stepwise process to obtain the ideal discriminant function. This process operates in two ways, starting like the 'forward method' to allow entry of variables in the model, but re-evaluates the already-entered variables on their discriminating ability in the presence of the subsequently entering variables, like the 'backward method'. The theoretical framework for this work was developed by Padmanaban and William (2016 a). Like the 'forward' and 'backward' methods, the proposed 'stepwise' method is applicable without the restrictions that limit the traditional approaches.

The traditional theory of multivariate discriminant analysis has gone through an interesting history of development over the past many years due to the contributions made by several scholars. Different approaches to develop discriminant models consider the issue of identifying the important variables to serve as inputs into the discriminant function. A few of the early contributions on this topic include those of Chang (1983) who suggested the use principal components for discriminating a mixture of two multivariate normal populations and a paper of Bensmail and Celeux (1996) who discussed Gaussian discriminant analysis through eigenvalue decomposition. A stepwise algorithm using 'Bayesian Information Criterion' was given by Murphy *et al.* (2010) following Raftery and Dean (2006) who suggested a similar approach for model-based clustering. The above contributions apply to parametric settings and are limited in their scope of applications.

There are other extensions of discriminant analysis to non-parametric settings that exist in the literature. Hastie *et. al.* (1994) proposed a nonparametric discriminant analysis with nonlinear classifiers applicable to situations with a large number of input variables. Baudat and Anouar (2000) gave a 'kernel approach' to nonlinear discriminant analysis which is theoretically close to supporting vector machines. Nonparametric discriminant analysis with adaptation to nearest-neighbor classification was developed by Bressan and Vitria (2003). Chiang and Pell (2004) proposed combining genetic algorithms with discriminant analysis for identifying key variables. In these works, a common question of concern was on the inclusion of relevant variables that would be effective in discriminating the populations under consideration.

This paper follows the approach given by Padmanaban and William (2016 a, b) which is different from the ideas presented in the above-mentioned works on two-population discriminant analysis. Interestingly, the new approach adheres well to the basic spirit and mathematical objective of the classical discriminant analysis. We seek to develop a two-way stepwise process as an alternative method to the forward and backward methods developed by Padmanaban and William (2016 a, b) for constructing an effective discriminant model. The stepwise method proposed here seeks to obtain an ideal discriminant model by including variables that contribute most to a discriminatory capacity of the model and excluding variables already selected if found to be ineffective discriminators on re-evaluation in the light of the additional input variables included. For a discussion on the 'model performance' measure to evaluate the classification ability of the discriminant model and the decision rule for identifying the optimal cut-off point for classification, reference is made to the original paper of Padmanaban and William (2016 a).
Hence, the objectives of the present work are:

i. To present a stepwise method for discriminating two populations and an easy-to-apply procedure for classification of objects.
ii. To apply the algorithm to a biomedical phenomenon and compare its classification-performance with that of logistic regression.

This paper is organized as follows: Following this introductory section, a review of the recently introduced nonparametric discriminant procedure given by Padmanaban and William (2016) is given briefly. The new Stepwise Algorithm for building an efficient discriminant model is discussed in Section 3. As an application of the proposed methodology, the prediction of respiratory tract disease(RD) among newborns of mothers with PPROM in the city of Chennai is considered. We discuss the phenomenon of PPROM, a potential risk that pregnant woman face, and the possible factors associated with the phenomenon. The stepwise algorithm of discriminant model building is applied to predict Respiratory tract disease among newborn of200 mothers with Preterm premature rupture of membranes, who delivered babies in Institute of Social Obstetrics, Government

Kasturba Gandhi Women and Children Hospital, Chennai, India, during the twelve month period of May 2016 to April 2017. Finally, a comparison of our discriminant model with the binary logistic model for the same response variable (RD) is carried out.

II. A REVIEW OF THE RECENTLY INTRODUCED PROCEDURE

Consider two multivariate populations π_1 and $\pi_2$whose relative sizes are in the proportions p_1 and p_2. The discrimination of the members of the two classes is to be carried out based on multivariate data on a random vector, say, $X = (X_1, X_2,..., X_p)^T$. The classification or membership-prediction of objects is a pertinent issue when there is a 'significant' difference between the distributions and the 'correctness' or 'incorrectness' of classifications, therefore, becomes a matter of concern.

Let the mean-vectors of Xin the two populations be$\mu_i=E_i(X)$, i = 1, 2, and the variance-covariance matrices of Xin the two populations be Σ_i, i = 1, 2. From Padmanaban and William (2016 a), we have the following theoretical results:

(i) For a random vector X and another random object W, the relationship between the unconditional and conditional mean vectors and variance-covariance matrices is given by

$E(X) = E_W[E_{X|W} (X)]$ and $V(X) = E_W\{V_{X|W}(X)\} + V_W\{ E_{X|W}(X)\}$. . . (2.1)

(ii) The overall variance-covariance matrix of the combined population is given by

$\Sigma = p_1\Sigma_1+ p_2\Sigma_2+ p_1(1-p_1)\, \mu_{1.}\, \mu_1^T +p_2(1-p_2)\, \mu_{2.}\, \mu_2^T - p_1\, p_2(\mu_1\, \mu_2^T + \mu_2\, \mu_1^T)$. . . (2.2)

In Discriminant Analysis, the multivariate observations (X) are transformed to univariate observations (Y) by considering linear combinations of the X_i's. For any linear combination $Y = \ell^T X$, where ℓ is a px 1 vector of constants, the means of Y in the two populations are $\mu_{1Y} = \ell^T\mu_1$ and$\mu_{2Y} = \ell^T\mu_2$ and in the combined population it is given by$\mu_Y = p_1\ell^T\mu_1 + p_2\ell^T\mu_2$. And, the variance of Y in the combined population is given by $V(Y) = \ell^T\, \Sigma\, \ell$.

The linear combination that gives the maximum (squared) distance between μ_{1Y} and μ_{2Y} relative to the variation in Y in the combined population would help best in discriminating the two populations. That linear combination of the X_i's which maximizes the scaled difference is the 'optimum discriminant function' based on X. We shall call it 'X-based optimal discriminant' and it is given by

$Y = (\mu_1 - \mu_2)^T\Sigma^{-1}X$ . . . (2.3)

For reasons of parsimony, it may not be proper to construct the optimal discriminant function using the entire set of variables $X_1, X_2,..., X_p$ that the investigator considers for the study, and only a subset of the variables may be found as important and as effective discriminators. Let the subset of the variables used to build the optimal discriminant be $X_{(s)}$. Denote the mean vectors of $X_{(s)}$ in the two populations as $\mu_{1(s)}$ and $\mu_{2(s)}$ and the 'overall' variance-covariance matrix of $X_{(s)}$ as $\Sigma_{(s)}$. The $X_{(s)}$-based optimal discriminant is

$Y_{(s)}= (\mu_{1(s)} - \mu_{2(s)})^T\Sigma^{-1}_{(s)}X_{(s)}$ . . . (2.4)

Generally, the parameters are replaced by the sample estimates in practice. Computing the variable $Y_{(s)}$ for all members in both the samples, the performance of the $X_{(s)}$-based optimal discriminant is measured by the two sample Kolmogorov-Smirnov Statistic based on the $Y_{(s)}$ measurements. Denoting the (empirical) cumulative distribution functions of $Y_{(s)}$ for the two populations as $F_{1(s)}(\cdot)$ and $F_{2(s)}(\cdot)$, the performance measure is given by

$KS_{(s)} = \max_y \left(| F_{1(s)}(y) - F_{2(s)}(y) | \right)$. . . (2.5)

Given two sub-vectors $X_{(s1)}$ and $X_{(s2)}$, the optimal $X_{(s1)}$-based discriminant is said to be 'more efficient' than the optimal $X_{(s2)}$-based discriminant if $KS_{(s1)}> KS_{(s2)}$. If there exists a sub vector $X_{(s*)}$ for which $KS_{(s*)}> KS_{(s)}$ for every other sub vector $X_{(s)}$, then the corresponding optimal discriminant $Y_{(s*)}$ is the 'most efficient' discriminant.

However, the process of obtaining the 'most efficient' discriminant is computationally prohibitive in the presence of a very large number of predictor variables (i.e.) in case of the very high dimension of the underlying random vector X. For instance, with 'p' input variables, one has to build as many as 2^{p-1} models to identify the 'best' one. This is true of every model-building situation involving a large number of predictor variables. To avoid this problem, different algorithms are suggested to 'build' improved models sequentially instead of considering 'all possible' models to get the 'most efficient'.

With this view, Padmanaban and William (2016 a) introduced a 'forward model-building' algorithm to build a 'sequence' of models, starting with a single variable and 'select' variables one-by-one evaluating their ability to 'add' to the discriminatory ability of the model. Following this, Padmanaban and William (2016 b) gave a 'backward' algorithm wherein the process starts with the 'largest' model involving all the 'p' variables under consideration and eliminates the least relevant variables one-by-one to arrive at the optimal model. In the same spirit, the next section presents a stepwise algorithm to construct a 'sequence' of models starting with one variable which has the maximum discrimination capacity and adding/removing variables depending on their capacity to increase/reduce the discriminatory ability of the discriminant, ultimately leading to an efficient discriminant model. The proposed algorithm takes the view that a variable that has already entered the model needs to be re-evaluated on its continuing effectiveness in the presence of variables that enter the model in later steps.

III. THE PROPOSED STEPWISE ALGORITHM

The proposed stepwise model-building algorithm evaluates each of the candidate input variables in a sequential manner and, in each step, brings in a single variable that contributes ‘most’ in improving the discriminatory capacity of the model and re-evaluates the already entered variables in each step to remove the one that loses its discriminatory ability in the presence of the latest entered variable. It starts like the ‘forward method’ of Padmanaban and William (2016 a) and also proceeds like the ‘backward method’ of Padmanaban and William (2016 b) as the model becomes ‘larger’ with more variables included in the discriminant function. In this context, we refer to the work of Habbema and Hermans (1977) who considered variable-selection for Gaussian discriminant analysis on the basis of F-statistics and error rates. We also refer to the paper by Pfeiffer (1985) who considered smoothing factors of kernel functions for nonparametric discriminant analysis with different criteria like distances, error rates and density-ratios.

As in the classical discriminant analysis. the discriminant scores form the basis for comparing the two populations. The inclusion or ‘entry’ of a variable in any step of the process is based on maximum differentiation between the distributions of the discriminant scores in the two populations. In contrast, the exclusion or ‘exit’ of a variable in any step is based on least differentiation between the discriminant scores in the two populations. These are measured by the two-sample Kolmogorov-Smirnov (KS) statistic used for comparison of two distributions. The discriminatory capacity of a model is measured by this statistic and we seek to maximize this statistic’s value as the distributions of the discriminant score need to be ‘significantly’ different for effective discrimination and separation. The exact stepwise model building process is described below.
Let $X_1, X_2, . . . , X_p$ be the candidate input variables and denote the vector $(X_1, X_2, . . . , X_p)$ as **X**.

Step 1 (Forward): With one variable at a time, ‘p’ discriminants $Y_{(1)}, Y_{(2)},... .,Y_{(p)}$, where $Y_{(i)}$ is the discriminant based on single input variable X_i, and their corresponding scores are obtained for each individual record in the data. Let the Kolmogorov-Smirnov Statistic for $Y_{(i)}$ be denoted as $KS_{(i)}$. If

$$KS_{(i)}> KS_{(j)} \text{ for every } j \neq i$$

then among the individual variables considered on a one-at-a-time basis, X_i is the top discriminator between the two populations. The significance of this $KS_{(i)}$ statistic is evaluated and if found significant at the desired level, X_i first 'enters' the model and model building continues.

Step 2 (Forward) : With X_i having been already selected, we take one additional variable at a time and obtain (p–1) discriminants having input-pairs $(X_1,X_i),....,(X_{i-1}, X_i),(X_{i+1},X_i),....,(X_p,X_i)$. Denote the discriminants as $Y_{(i,1)}, Y_{(i,2)},..., Y_{(i,i-1)}, Y_{(i,i+1)}, ... Y_{(i,p)}$ and the corresponding Kolmogorov-Smirnov statistics as $KS_{(i,1)}, KS_{(i,2)},...,KS_{(i,i-1)}, KS_{(i,i+1)},..., KS_{(i,p)}$. If for some 'm',

$$KS_{(i,m)}> KS_{(i,j)} \text{ for every } j \neq m, \text{ and } KS_{(i,m)}> KS_{(i)},$$

then X_m enters the model as the second variable. It is to be noted that the significance of $KS_{(i,m)}$ is guaranteed because of the significance of $KS_{(i)}$ in the first step. In contrast, if

$$KS_{(i,m)}> KS_{(i,j)} \text{ for every } j \neq m, \text{ but } KS_{(i,m)} \leq KS_{(i)},$$

then X_m does not enter the model, nor any of the remaining X_j's enter, as its entry leads to reduced discriminatory ability and the model building stops with only one input variable. Clearly, no other variable can enter.

Step 3 (Forward): With X_i and X_m having been already selected, we take one additional variable at a time and obtain (p–2) discriminants having input-triplets (X_i, X_m, X_j), for $j \neq i$, m. Denote the discriminants as $Y_{(i, m, j)}$ and the corresponding Kolmogorov-Smirnov statistics as $KS_{(i, m, j)}$.
If for some 'r',

$$KS_{(i, m, r)}> KS_{(i, m, j)} \text{ for every } j \neq r, \text{ and } KS_{(i,m,r)}> KS_{(i,m)},$$

then X_r enters the model as the third variable. It is to be noted that the significance of $KS_{(i,m,r)}$ is guaranteed because of the significance of $KS_{(i,m)}$ in the previous step. In contrast, if

$$KS_{(i,m,r)}> KS_{(i,m,j)} \text{ for every } j \neq r, \text{ but } KS_{(i,m,r)}\leq KS_{(i,m)},$$

then X_r does not enter the model, nor any of the remaining X_j's enter, as its entry leads to reduced discriminatory ability – the model building stops with only two input variables. Clearly, no other variable can enter.

Step 3 (Backward): This backward part of Step 3 is intended to re-evaluate the continued relevance of the variables X_i and X_m that entered the model in the previous two steps in the light of the entry of X_r. We consider two discriminants having input variable pairs (X_m, X_r) and (X_i, X_r), that is, one by removing X_i (which entered in Step 1) and the other by removing X_m (which entered in Step 2). Denote the discriminants as $Y_{(\sim i,m,r)}$ and $Y_{(i,\sim m,r)}$ and the corresponding Kolmogorov-Smirnov statistics as $KS_{(\sim i,m,r)}$ and $KS_{(i,\sim m,r)}$. The '~' indicates the variable is removed.

If $\max\{KS_{(\sim i,m,r)}$, $KS_{(i,\sim m,r)}\}<KS_{(i,m,r)}$, then neither $X_i$nor$X_m$leaves the model because, their exit reduces the KS statistic value leading to lower discriminatory capacity than the model involving the three variables X_i, X_m, X_r.
If $KS_{(\sim i,m,r)} \geq KS_{(i,m,r)} > KS_{(i,\sim m,r)}\}$, then X_i leaves the model because the model with just $X_m$and$X_r$itselfgives a discriminatory capacity which is better or at least as good as the one which includes X_i, X_m and X_r. Similarly, if $KS_{(i,\sim m,r)} \geq KS_{(i,m,r)} > KS_{(\sim i,m,r)}$, then X_m leaves the model for the same reason as above.

If $KS_{(\sim i,m,r)} \geq KS_{(i,\sim m,r)} \geq KS_{(i,m,r)}$, then X_i (but not X_m) leaves as its exit gives the greatest improvement in KS value. In a similar vein, if $KS_{(i,\sim m,r)} \geq KS_{(\sim i,m,r)} \geq KS_{(i,m,r)}$, then X_m (but not X_i) leaves. We note here that, in this case, the exit of either X_i or X_m leads to improved KS value, but we remove only one at a time, and the one is that variable whose removal gives the greatest improvement in the KS value.

At every subsequent step, in the forward part, a new entry is allowed if there is a strictly positive addition to the KS value by its entry; and, in the backward part, an already entered variable is removed if there is no reduction in the KS value by its exit.
The process stops at a step where the forward part finds that none of the 'waiting' variables qualify for entry. When the process stops at $(k+1)^{th}$ step, the optimal discriminant function is the one obtained in the k^{th} step with the maximum KS value, leading to significant and maximum discrimination between the two populations. We denote the final subset of variables reached in this process as $X_{(S^*)}$ and the 'final' efficient discriminant as $Y_{(s^*)}$.

Note that the backward part is not applicable in the 2nd step because, the KS statistic for the model with X_i removed, namely, $KS_{(\sim i,m)}$ is nothing but $KS_{(m)}$ of Step 1 and similarly $KS_{(i,\sim m)}$ is same as $KS_{(i)}$ of step 1 and, clearly, $KS_{(i,\sim m)} > KS_{(\sim i,m)}$ so X_i cannot leave the model as its removal reduces the KS statistic value.
The classification or prediction rule, the 'explanation' to the KS statistic and also the suggestion to use the 'Reliability Function' for computing the KS Statistic are provided in the paper of Padmanaban and William (2016 a) wherein the proposal for the forward model-building process was originally proposed.

IV. PREDICTION OF RESPIRATORY TRACT DISEASE AMONG NEWBORN OF MOTHERS WITH PRETERM PREMATURE RUPTURE OF MEMBRANES

Preterm Premature rupture of Membranes(PPROM): Premature rupture of membranes(PROM) is defined as the spontaneous rupture of the amniotic membrane with a release of amniotic fluid at least one hour before the onset of labour. If the membranes rupture after 37 weeks of gestation it is called term PROM. The rupture of membranes(ROM) occurring after 28 weeks but before 37 weeks of gestation is termed as the preterm premature rupture of membrane (PPROM).
PROM occurs in approximately 10 % of all pregnancies and in 70% of the cases at term. Although there is some morbidity when PROM occurs in term pregnancies, the fundamental clinical problem is PPROM, a condition that occurs in 3 % of all pregnancies and is responsible for approximately 30% of all preterm deliveries.

Respiratory Tract Disease: Respiratory disease is a medical term that encompasses pathological conditions affecting the organs and tissues that make gas exchange possible in higher organisms, and includes conditions of the upper respiratory tract, trachea, bronchi, bronchioles, alveoli, pleura, and pleural cavity, and the nerves and muscles of breathing. Respiratory diseases range from mild and self-limiting, such as the common cold, to life-threatening entities like bacterial pneumonia, pulmonary embolism, acute asthma, and lung cancer.

The incidence of Respiratory Tract Disease is 53% among newborns of mothers with Preterm Premature rupture of membranes. In this context, we refer toa paper of Khade and Bava (2018) which studied maternal and perinatal outcome in PPROM cases.

Objective: The present study aims to relate three factors namely Mother's Age, Gestation Week (of Birth) and BMI of mother to Respiratory Tract Disease of a newborn child through the 'Stepwise' discriminant model building algorithm developed in this paper. We wish to identify the significant factors that are associated with the risk of Respiratory Tract Disease among newborns. We introduce linear as well as square terms of these variables for possible inclusion in the model.

We have data on 200 mothers with PPROM who delivered babies in the Institute of Social Obstetrics, Government Kasturba Gandhi Women and Children Hospital, Chennai, India, during the period May 2016 to April 2017. The following is a sample of the data on the variables considered as 'Potential factors' along with the birth outcome:

Record #	X_1 (AGE)	X_2 (BMI)	X_3 (GW)	X_4 (AGE^2)	X_5(BMI^2)	X_6(GW^2)	Outcome
1	27	23	32	729	529	1024	1
2	28	25	30	784	625	900	1
3	23	27	32	529	729	1024	1
4	22	21	36	484	441	1296	0
5	24	25	36	576	625	1296	0

Here, '1' denotes Respiratory Tract Disease and '0' denotes 'No RD group'.
We apply the stepwise algorithm developed in this paper and get the following results.

Step 1 (FW): The KS statistics for models with single variables are found to be:
KS$_{(X1)}$ = 0.2869, KS$_{(X2)}$ = 0.2146, **KS$_{(X3)}$ = 0.4183**, KS$_{(X4)}$ = 0.2869, KS$_{(X5)}$ = 0.2146, KS$_{(X6)}$ = 0.4183
With maximum KS, X_3 enters the model in the first step. The KS value of 0.4183 is found to be statistically significant. We can find that the KS$_{(X3)}$ = KS$_{(X6)}$ where X_6 is the square of X_3. Similar equality holds between X_1 and X_4 and, also between X_2 and X_5. That is, the linear and square terms of a factor have same discriminatory power as stand-alone variables. With maximum KS, X_3 and X_6 both qualify for entry in Step 1, but preference is to the direct variable (linear) X_3.

Step 2 (FW): The KS statistics for models including one additional variable with X_3 are found as
KS$_{(X1,X3)}$ = 0.5136, KS$_{(X2,X3)}$ = 0.459, KS$_{(X3,X4)}$ = 0.5136, KS$_{(X3,X5)}$ = 0.4509, KS$_{(X3,X6)}$ = 0.4183
X_1 enters the model in the second step. Even though, X_4 also qualifies to enter with maximum KS, the preference is to the linear term X_1.

Step 3 (FW): In this step we get
KS$_{(X1,X2,X3)}$= 0.4509, **KS$_{(X1,X3,X4)}$ = 0.5449**, KS$_{(X1,X3,X5)}$ = 0.4762, KS$_{(X1X3,X6)}$ = 0.5136
X_4 enters the model in the third step.

Step 3(BW): We check for possible removal of X_3 and X_1 which entered in previous steps. We get
KS$_{(\sim X3,\ X1,X4)}$ = 0.4195, KS$_{(X3,\sim X1,X4)}$ = 0.5136
both of which are less than the KS attained in the forward Step 3. Thus, no variable leaves the model.

Step 4 (FW): In this step we get
KS$_{(X1,X2,X3,X4)}$ = 0.4749, KS$_{(X1,X3,X4,X5)}$ = 0.4955, **KS$_{(X1,X3,X4,X6)}$ = 0.5702**.
X_6 enters in this step.

Step 4 (BW): We check for possible exit of X_1, X_3, X_4. We have
KS$_{(\sim X1,X3,X4,X6)}$ = 0.5136, KS$_{(X1,\sim X3,X4,X6)}$ = 0.5449, KS$_{(X1,X3,\sim X4,X6)}$ = 0.5136, all three being less than the KS in Step 4 (FW). So, there is no exit in this step.

Step 5 (FW): In this step we get
KS$_{(X1,X2,X3,X4,X6)}$ = 0.5003, KS$_{(X1,X3,X4,X5,X6)}$ = 0.5003, both less than the KS attained in Step 4 (FW).
So, neither of the two remaining variables X_2, X_5 enters the model.
As none of the latest KS statistics exceeds the previous maximum KS value, the stepwise process model building process stops with four variables being selected up to Step 4, in the order of X_3, $X_{1,}$ $X_{4.}$ and X_6.The 'Efficient Discriminant' obtained at the end of Step 3 of our algorithm is:

$$Y = -0.91774*Age + 5.65331*GW + 0.01798*Age^2 - 0.0911* \ldots \quad (4.1)$$

and the 'efficient cut-point' is $y_0 = 74.11286$
<u>Membership-Prediction Rule</u>: If 'y' denotes the measured value of the 'Efficient Discriminant' Y of (4.1) for a mother, then the prediction rule for her would-be born child is:

$$\text{Classify individual to:}\begin{cases} Respiratory\ Tract\ Disease\ Group & if\ \ y > 74.1128 \\ No\ Disease\ Group & if\ y\ \leq 74.1128 \end{cases}$$

We note that higher values of the discriminant ‘Y’, indicates higher risk for RD. Further, the discriminant Y is a convex function of ‘Age’ for fixed ‘GW’ and a concave function of ‘GW’ for fixed ‘Age’.
We find that the minimum risk for child’s RD is attained when the mother’s age is around 25 years. New borns of lower age mothers (between 18 and 25) and excess age mothers (above age 25) with PPROM both face a higher risk for RTD.

Also, when the ‘Gestation Week’ (the week in which the premature birth takes place) is around 31, the maximum risk for RD is reached. The range of GW is 29 to 36 and for mothers with GW less than 31 also, the risk is not substantially lower than GW 31, but stays high. The risk is decreased when the birth does not take place ‘too early’ and moves closer towards the normal GW of 40.

V. CONCLUSION

Comparison with the Logistic Regression Model:

Denoting Respiratory Tract Disease outcome' as the outcome of interest (1), we build a binary logistic regression model using the stepwise method of model building. We have the following logit equation from the model:

$$\log\left(\frac{p}{1-p}\right) = -16.2071 + 2.04551*\text{BMI} - 0.033801*\text{BMI}^2 - 0.010467*\text{GW}^2$$

where 'p' is the probability of preterm labour. The KS for this model is found to be **0.525618**which is less than the KS obtained for the 'Efficient Discriminant Model(with KS=**0.570223**). Thus, the new method performs better than the binary logistic regression method in predicting Respiratory disease of newborns to mothers with PPROM.

We note that, while logistic regression identifies two factors BMI and GW, our efficient discriminant model captures ‘Age’ and ‘GW’. It is noted that there is no proper medical evidence or research papers establishing the relationship between BMI of mothers with PPROM and Respiratory Tract Disease of Newborns. However, both models confirm that higher values of GW lead to lower risk for RTD. That is when the premature birth does not happen too early but closer to the normal 40 week period, the risk for RTD decreases. Interestingly, our discriminant model identifies the fact that, in PPROM cases, the risk for RD of the newborn decreases as the age of mother moves up to 25 years, attains a minimum around 25 and increases forage beyond 25 years.

Looking after a premature infant puts an immense burden on the economic and health care resources of a family. Risk scoring strategies to identify high-risk cases are very useful to initiate timely remedial actions for treating them prior to the rupture of the membrane. The above application of our discriminant model is a step in this direction. Identification of other risk factors for earlier detection of high-risk cases will be addressed in future studies.

Another interesting aspect of the above application is the inclusion of square terms along with linear terms which have not been attempted in non-parametric discriminant analysis. This has resulted in identifying non-linear relations of input factors to the discriminant function which could throw up interesting insights towards discovering risks for an outcome.

REFERENCES

[1].Baudat, G. and Anouar, F. (2000). Generalized Discriminant Analysis using a Kernel Approach. Neural Computation, 12 (10), 2385-2404

[2].Bensmail, H. and Celeux, G. (1996). Regularized Gaussian discriminant analysis through eigen value decomposition. J. Amer. Statist. Assoc. 91, 1743-1748

[3]. Bressan, M. and Vitria, J. (2003). Nonparametric Discriminant Analysis and Nearest Neighbor Classification. Pattern Recognition Letters. 24, 2743-2749

[4]. Catov, J.M., Bodnar, L.M., Ness, R.B., Barron, S.J. and Roberts, J.M. (2007). Inflammation and Dyslipidemia related risk of Spontaneous Preterm Birth. Am. J. Epidemiol. 166, 1312-1319

[5]. Chang, W.-C. (1983). On using Principal Components before Separating a Mixture of two Multivariate Normal Distributions. J. Roy. Statist. Soc. Ser C. 32, 267-275

[6]. Chiang, L.H. and Pell, R.J. (2004). Genetic algorithms combined with discriminant analysis for key variable identification. J. Process Control, 14, 143-155

[7]. Habbema, J.D.F. and Hermans, J. (1977). Selection of variables in discriminant analysis by F- statistic and error rate. Technometrics. 19, 487-493

[8]. Hastie, T., Tibshirani, R., and Buja, A. (1994). Flexible Discriminant Analysis by Optimal Scoring. J. Amer. Statist. Assoc. 89, 1255-1270

[9].Khade, S.A. and Bava, A, K. (2018). Preterm Premature rupture of membranes – maternal and perinatal outcome- International Journal of Reproduction, Contraception, Obstetrics and Gynecology, 7 (11), 4499-4505

[10].Mudd, L.M., Holzman, C.B., Catov, J.M., Senagore, P.K. and Evans, R.W. (2012). Maternal lipids at mid pregnancy and risk of preterm delivery. Acta Obstet. Gynecol. Scand. 91, 726-735

[11].Murphy, T.B., Dean, N. and Raftery, A.E. (2010). Variable Selection and updating in Model- Based Discriminant Analysis for High Dimensional Data with Food Authenticity Applications. The Annals of Applied Statistics, Vol.4, No.1, 396-421

[12].Padmanaban, S and Martin L. William (2016). A nonparametric discriminant variable-selection algorithm for classification to two populations International Journal of Applied Mathematics and Statistical Sciences. 5 (2),87-98

[13].Padmanaban and William (2016a).A nonparametric discriminant variable-elimination algorithm for classification to two populations. International Journal of Applied Mathematics and Statistical Sciences. 5(6),7-16.

[14]. Pfeiffer, K.P.(1985). Stepwise Variable Selection and Maximum Likelihood Estimation of Smoothing Factors of Kernel Functions for Nonparametric Discriminant Functions evaluated by Different Criteria. J. Biomed. Informatics. 18, 46-61

[15]. Raftery, A.E. and Dean, N. (2006). Variable Selection for Model-Based Clustering. J. Amer. Statist. Assoc. 101, 168-178