UNIVERSIDAD NACIONAL AGRARIA

LA MOLINA

ESCUELA DE POSGRADO

MAESTRÍA EN ESTADÍSTICA APLICADA

"COMPARACIÓN DEL ANÁLISIS DISCRIMINANTE NO MÉTRICO, ÁRBOLES DE CLASIFICACIÓN CHAID Y LA REGRESIÓN LOGÍSTICA MULTINOMIAL"

Presentada por:

RUBÉN ELVIS SUCARI SUCARI

TESIS PARA OPTAR EL GRADO DE

MAGISTER SCIENTIAE EN ESTADÍSTICA

APLICADA

Lima-Perú

2018

UNIVERSIDAD NACIONAL AGRARIA LA MOLINA

ESCUELA DE POSGRADO

MAESTRÍA EN ESTADÍSTICA APLICADA

"COMPARACIÓN DEL ANÁLISIS DISCRIMINANTE NO MÉTRICO, ÁRBOLES DE CLASIFICACIÓN CHAID Y LA REGRESIÓN LOGÍSTICA MULTINOMIAL"

TESIS PARA OPTAR EL GRADO DE
MAGISTER SCIENTIAE

Presentada por:

RUBÉN ELVIS SUCARI SUCARI

Sustentada y aprobada ante el siguiente jurado:

Mg.Sc.Fernando Miranda Villagómez	Mg.Sc.Jaime Carlos Porras Cerrón
PRESIDENTE	**PATROCINADOR**
Mg.Sc.Carlos López de Castilla Vásquez	Mg.Raphael Valencia Chacón
MIEMBRO	**MIEMBRO**

Dedicado a Dios, a mi madre y
a mis hermanos.

AGRADECIMIENTO

A mi madre Marcela Sucari, por la motivación constante para superarme.

A mis hermanos por su fé inquebrantable en mi trabajo.

A mi asesor por su confianza en la calidad de mi investigación.

A los miembros del jurado por dedicar su tiempo para revisar este texto.

RESUMEN

En la presente tesis se desarrolló el método de clasificación llamado Análisis Discriminante No Métrico, y se comparó su desempeño con el Árbol de Clasificación CHAID y la Regresión Logística Multinomial, los cuales también son métodos que no necesitan la condición de normalidad multivariada, linealidad ni varianza homogénea para las variables independientes. Esta comparación de desempeño fue evaluado mediante la Validación Cruzada. Para la realización del estudio comparativo de estos clasificadores se utilizó conjuntos de datos que son proporcionados por la Universidad de California Irving (UCI).

Se concluye que la Regresión Logística Multinomial tiene mejor desempeño en la clasificación de datos teniendo en cuenta la tasa de clasificación promedio y el tiempo de procesamiento.

Palabras clave: Análisis Discriminante Lineal, Análisis Discriminante No Métrico**,** Árboles de Clasificación CHAID, Regresión Logística Multinomial, Validación Cruzada.

ABSTRACT

In this thesis a method was developed called Non-Metric Discriminant Analysis, and its performance was compared with the Classification Tree CHAID and Multinomial Logistic Regression, which are also non-parametric methods. This performance comparison was evaluated using Cross Validation. To perform the comparative study of these classifiers we used data sets that are provided by the University of California Irving (UCI).It is concluded that the Multinomial Logistic Regression performs better in the classification of data taking into account the average classification rate and processing time.

Keywords: Linear Discriminant Analysis, Discriminant Analysis No Metric Trees CHAID Classification, Multinomial Logistic Regression, Cross Validation.

ÍNDICE GENERAL

ÍNDICE DE CUADROS

ÍNDICE DE FIGURAS

ÍNDICE DE ANEXOS

I. INTRODUCCIÓN

En diversas áreas como la industria, investigación de mercados, administración, salud, meteorología, finanzas y otros se presenta el problema de clasificar un nuevo individuo (al que se evalúa ciertas características cualitativas o cuantitativas) en alguna de las clases (o categorías) preestablecidas. A este tipo de estudio se le conoce como Análisis de Clasificación. Tal es el caso que ocurre con los analistas de riesgo de un banco que deben aprobar o denegar las solicitudes de préstamos a sus clientes.

Las técnicas de clasificación son herramientas importantes que fueron desarrolladas dentro de la Estadística con el fin de clasificar nuevos elementos. Algunas técnicas de clasificación son: Árbol de Clasificación CHAID, la Regresión Logística Multinomial, Análisis Discriminante Lineal, entre otras. Sin embargo esta última técnica, requiere el cumplimiento de ciertos supuestos como son: la Normalidad Multivariada, la Homogeneidad de matrices de covarianza y la no multicolinealidad. Es decir, los métodos de clasificación pueden ser divididos en paramétricos cuando las variables provienen de una distribución normal multivariada con igual variancia dentro de cada grupo (homocedasticidad) y no paramétricos cuando no se requiere el supuesto de normalidad y homocedasticidad entre ellos: el método de los k vecinos más cercanos; el basado en núcleos Kernel; los árboles de decisión y las redes neuronales artificiales.

En esta tesis se desarrolló el método clasificación llamado Análisis Discriminante No Métrico, y se comparó su desempeño con el Árbol de Clasificación CHAID y la Regresión Logística Multinomial, los cuales son métodos que no exigen el requerimiento de los supuestos de normalidad y homocedasticidad. Esta comparación de desempeño fue evaluado mediante la Validación Cruzada. No existen muchos trabajos referentes a este tema, sólo se ha publicado una comparación de Análisis Discriminante No Métrico contra Regresión Logística para el caso de dos poblaciones, Usuga (2006).

Por tal motivo, el objetivo de este trabajo de investigación será comparar el Análisis Discriminante No Métrico con el Árbol de Clasificación CHAID y La Regresión Logística Multinomial utilizando conjuntos de datos que están a disposición para la comunidad científica y son proporcionados por la Universidad de California Irving (UCI), cuya dirección en internet es: http://www.ics.uci.edu/~mlearn/MLSummary.html con el fin de determinar cuál de las técnicas es la mejor técnica para clasificar tomando en cuenta la tasa de clasificación errónea y el tiempo de procesamiento.

Por último, la presente investigación presentará el desarrollo teórico de las técnicas de Análisis Discriminante No Métrico y el Árbol de Clasificación CHAID son métodos tan buenos para clasificar individuos como la Regresión Logística Multinomial.

II. REVISIÓN DE LITERATURA

2.1 TERMINOLOGÍA GENERAL

En esta parte se presentan algunos conceptos importantes que están relacionados con el estudio de los métodos de clasificación. Estos conceptos serán usados a lo largo del documento.

2.1.1 TERMINOLOGÍA BÁSICA DE LA CLASIFICACIÓN

a. Clasificación

Sea $(x_1, x_2, x_3, \ldots\ldots x_p)$ un vector $\boldsymbol{p}$ variado. Barajas y Morales (2009) mencionan que el proceso de asignar una observación $\boldsymbol{p}$ variada predictora en uno de varios grupos preestablecidos, se denomina clasificación. El objetivo básico es construir un modelo estadístico que tome los datos de las p variables para resumirla en un indicador con el cual se puede clasificar los datos de manera correcta en uno de los grupos. En la literatura estadística se pueden encontrar varios métodos de clasificación como: Análisis Discriminante Lineal (ADL), Regresión Logística (RL), Regresión Logística Multinomial, (RLM), Análisis Discriminante no Métrico (ADNM), Redes Neuronales, Árboles de Clasificación, etc.

La clasificación es la discriminación de nuevos individuos a un grupo ya predefinido. Para lograr esta clasificación se hace de un conjunto de datos que será dividido en dos partes: una parte que servirá para construir (entrenar un modelo estadístico) y el resto de los datos será para evaluar la eficiencia del modelo con los datos sin clasificar según Benny y Linoff (2004).

b. Eficiencia relativa asintótica

La variabilidad es en esencia inherente a la estadística, su razón y su objeto. Por ello la variabilidad medida por medio de la varianza, se convierte en un criterio de examen de estadísticas ya que evidentemente es más preciso aquel estimador que tenga menor varianza, ya que tiene la capacidad de producir estimadores más concentrados.

Definición 1

La eficiencia relativa de $T_n^{(2)}$=$T_2(X_1,\ldots,X_n)$ respecto a $T_n^{(1)}$=$T_1(X_1,\ldots,X_n)$ estimadores insesgados para la imagen de θ bajo una función r, basado en una muestra aleatoria $X_1,\ldots,X_n$ de una población de densidad f_X(x, θ), corresponde al cociente

$$\frac{V_\theta[T_n^{(1)}]}{V_\theta[T_n^{(2)}]}$$

Siendo la eficiencia relativa un elemento de comparación entre dos estimadores, pueden involucrarse elementos adiciones para enriquecer las mencionada comparación, como el tamaño de muestra.

Suponiendo que $T_n^{(1)}$ y $T_m^{(2)}$ sean dos estimadores para la imagen de θ bajo una función r, tales que $T_n^{(1)}$~$N\left(r(\theta),\frac{\sigma_1^2(\theta)}{n}\right)$ y $T_m^{(2)}$~$N\left(r(\theta),\frac{\sigma_2^2(\theta)}{m}\right)$ asumiendo que $\sigma_1^2(\theta)<\sigma_2^2(\theta)$, la eficiencia relativa de $T_m^{(2)}$ respecto de $T_n^{(1)}$ corresponde a

$$\frac{\sigma_1^2(\theta)/n}{\sigma_2^2(\theta)/m}.$$

En estos términos $T_m^{(2)}$ será tan eficiente como $T_n^{(1)}$ en la medida que la citada eficiencia tenga un valor igual a uno; caso en el cual $\frac{\sigma_1^2(\theta)}{\sigma_2^2(\theta)}=\frac{n}{m}$. Teniendo en cuenta que $\sigma_1^2(\theta)<\sigma_2^2(\theta)$, entonces $\frac{n}{m}<1$. Si en gracia a esta consideración el valor del cociente $\frac{\sigma_1^2(\theta)}{\sigma_2^2(\theta)}$ se asume en 0.9, quiere decir que $T_m^{(2)}$ requiere una muestra de un tamaño cercano al 11.11% mayor que el tamaño de la muestra n calculado con base en el estimador $T_n^{(1)}$ para tener igual desempeño, o igualmente que a $T_n^{(1)}$ sólo le basta contar con el 90% del tamaño de muestra calculado para $T_m^{(2)}$.

Definición 2

La eficiencia relativa asintótica de $T_n^{(2)}$ respecto a $T_n^{(1)}$, siendo $T_n^{(1)}$ y $T_n^{(2)}$ estimadores de consistencia asintóticamente normal para la imagen de θ bajo una función r con varianza $\sigma_1^2(\theta)$ y $\sigma_2^2(\theta)$ respectivamente es el cociente.

$$\frac{\sigma_1^2(\theta)}{\sigma_2^2(\theta)}$$

Mayorga. J., (2004). Inferencia estadística, Bogotá, Colombia: Universidad Nacional de Colombia

2.2 ANÁLISIS DISCRIMINANTE LINEAL (ADL)

El Análisis Discriminante Lineal es una técnica multivariante que permite asignar o clasificar nuevos elementos dentro de grupos previamente reconocidos o definidos.

En el ADL se tienen g grupos ya establecidos el cual cuenta con un conjunto de observaciones pertenecientes a cada uno de los grupos. Sean $x_{ij} \in R^p$ las observaciones donde j=1,…,g, i=1,…,n_j tal que n_j representa el número de observaciones que pertenecen al grupo j. Este conjunto de n observaciones [1] se denomina conjunto de entrenamiento y permite la construcción de la función discriminante.

Sean $\bar{x_j}$ y S_j los vectores de medias y las matrices de covarianzas de cada uno de los **g** grupos y sea $\bar{x}$ el vector de medias global del conjunto de entrenamiento. Usando estos elementos se pueden definir dos matrices B y W que representan la variabilidad entre los grupos y la variabilidad dentro de los grupos respectivamente y que están dadas por:

$$B = \sum_{j=1}^{g} n_j \left(\bar{x_j} - \bar{x}\right)\left(\bar{x_j} - \bar{x}\right)', \quad (1.1)$$

$$W = \sum_{j=1}^{g} (n_j - 1) S_j \quad (1.2)$$

donde B y W son dos matrices que representan cada una la variabilidad entre los grupos y la variabilidad dentro de los grupos respectivamente. El objetivo de la técnica ADL es encontrar un vector $a \in R^p$ de tal manera que se maximice el cociente Λ definido por (1.3). Así se encuentra un hiperplano que genera la máxima diferencia entre la variabilidad intergrupal e intragrupal.

$$\Lambda = \frac{a'Ba}{a'Wa} \quad (1.3)$$

Los valores de *a* que maximizan Λ se pueden encontrar por medio de los vectores propios $\hat{e_1}, \hat{e_2}, \ldots, \hat{e_s}$ asociados con los valores propios positivos [2] $\hat{\lambda_1} > \hat{\lambda_2} > \ldots > \hat{\lambda_s}$, de $W^{-1}B$.

De esta manera si $\hat{a} = \hat{e_1}$ entonces $\hat{a}$ se le denomina primer discriminante lineal (DL_1), si $\hat{a} = \hat{e_2}$ entonces $\hat{a}$ se le denomina segundo discriminante lineal (DL_2) y así sucesivamente hasta $\hat{a} = \hat{e_s}$, en cuyo caso $\hat{a}$ se denomina s-ésimo discriminante lineal (DL_s).

---

1. En total n=$\sum_{j=1}^{g} n_j$
2. Donde s = min {p, g − 1}

La regla para clasificar una nueva observación x en uno de los grupos basada en el primer discriminante lineal consiste en asignar x al grupo j si cumple que: $\sum_{i=1}^{r}\left[\hat{a}'_i\left(x-\bar{x}_j\right)\right]^2$es mínimo[3].

2.3 REGRESIÓN LOGÍSTICA

La Regresión Logística es una técnica estadística que desde su primera aplicación a las ciencias de la salud en 1967, se ha extendido de manera vertiginosa, siendo en los últimos años la técnica de análisis estadístico multivariante más utilizada, el modelo de regresión logística que se llama así porque la función que la define es una curva logística. El modelo de regresión logística el cual tiene una variable dependiente binomial (o multinomial) es un modelo que permita estudiar si dicha variable discreta depende o no, de otra u otras variables.

a) Regresión Logística Binaria En estadística aplicada es muy frecuente tener que calcular la probabilidad de que ocurra o de que no ocurra determinado suceso. Cuando una variable sólo tiene dos posibilidades (o sucesos) de ocurrir se dice que es dicotómico. De los dos sucesos, se denomina suceso de interés al que se desea conocer su probabilidad, en diversas circunstancias, es el principal objetivo de un trabajo de investigación.

La Regresión Logística será adecuada para estudiar la presencia o ausencia de una característica o resultado según los valores de un conjunto de covariables.

Se denotará como Y a la variable respuesta del tipo binario y P (Y) a la probabilidad de que ocurra el suceso, y a X_1, X_2,…, X_p, a las variables que pueden influir en dicha probabilidad.

Se tiene una muestra de tamaño n= $n_1 + n_2$, con n_1 observaciones de la clase C_1 y n_2 observaciones de la clase C_2.

3. Donde r≤s.

La variable predictora Y se define como 0 y 1 para cada clase.

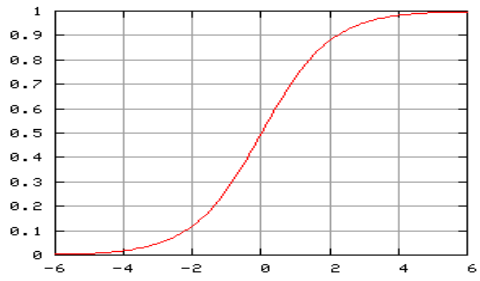


Figura 1: Gráfica de la Regresión Logística Binaria

Este modelo comúnmente presenta una forma de "S", limitada en el eje de las ordenadas entre los valores 0 y 1. El modelo descrito se denomina "**Función Logística**".

2.3.1 SUPOSICIÓN DEL MODELO DE REGRESIÓN LOGÍSTICA

Consideremos que sólo tenemos dos clases es decir que nuestro conjunto de datos consiste de una muestra de tamaño $n=n_1+n_2$, n_1 observaciones son de la clase n_1 son de la clase C_1 y n_2 son de la clase C_2 para cada observación x_j se introduce una variable binaria y que vale 1 si ella es de la clase C_1 y vale 0 si la observación pertenece a la clase C_2. La variable y tiene una probabilidad a priori π_1 de que y es 1.

Sea $f(x/C_i)$ (i=1,2) la función de densidad del vector aleatorio p-dimensional x en la clase C_i, en el modelo logístico se asume que:

$$\log\left(\frac{f(x/C_1)}{f(x/C_2)}\right) = \alpha + \beta' x$$

donde β es un vector de p parámetros y α representa el intercepto

2.3.2 ODDS RATIO

La regresión logística puede utilizarse como método para la estimación riesgo relativo (odds ratio OR). Los odds ratio o coeficiente de posibilidades es una relación entre dos probabilidades. Es la razón entre la probabilidad de que se produzca un suceso y la probabilidad de que no se produzca ese suceso.

Sea p=P(Y=1/x) la probabilidad a posteriori de que Y sea igual a 1 para un valor observado de x.

Se define la razón de apuestas (odds ratio) como:

$$\frac{p}{1-p} = \frac{\frac{P\{Y=1\}f(x/y=1)}{f(x)}}{\frac{P\{Y=0\}f(x/y=0)}{f(x)}} = \frac{\pi_1 f(x/C_1)}{\pi_2 f(x/C_2)} \qquad (1.4)$$

donde π_i representa la probabilidad a priori de que Y pertenezca a la clase C_i

2.3.3 TRANSFORMACIÓN LOGIT

Tomando logaritmo a (1.4) se tiene

$$\log\left(\frac{p}{1-p}\right) = \log\left(\frac{\pi_1}{\pi_2}\right) + \log\frac{f(x/C_1)}{f(x/C_2)}$$

Luego con la suposición se tiene que:

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta' x$$

Este modelo es útil en situaciones prácticas de investigación en que la variable respuesta puede tomar solo dos valores, por ejemplo: desaprobado o aprobado; e interesa conocer la probabilidad de que un alumno este desaprobado en función de su perfil de variables predictivas o factores de riesgo.

La utilidad del modelo se basa en que muchas veces, el perfil de variables predictivas puede estar formado por características cualitativas y cuantitativas; y se pretende hacer participar a todas en una sola ecuación conjunta que explique como la probabilidad de alcanzar una respuesta depende de todas y cada una de las variables predictivas.

Equivalentemente

$$p = \frac{\exp(\alpha + \beta' x)}{1 + \exp(\alpha + \beta' x)}$$

2.3.4 ESTIMACIÓN DE LOS COEFICIENTES DE LA REGRESIÓN LOGÍSTICA

Método de Máxima Verosimilitud

- La estimación de los coeficientes de la regresión logística se puede realizar mediante diversos métodos, pero el más utilizado es el de Máxima Verosimilitud.

- Dada una observación x, las probabilidades de que esta pertenezca a las clases C_1 y C_2 son:

  $P(C_1/x)=\frac{\exp(\alpha+\beta' x)}{1+\exp(\alpha+\beta' x)}$

$$P(C_2/x) = 1 - P(C_1/x) = \frac{1}{1+\exp(\alpha+\beta' x)}$$

- Considerando una muestra de tamaño $n=n_1+n_2$ y un parámetro binomial p igual a la función de verosimilitud es de la forma:

$$L(\alpha,\beta) = \prod_{i=1}^{n_1} \frac{\exp(\alpha+x_i'\beta)}{1+\exp(\alpha+x_i'\beta)} \cdot \prod_{j=n_1+1}^{n} \frac{1}{1+\exp(\alpha+x_j'\beta)}$$

La solución de la ecuación de verosimilitud es con métodos numéricos.

2.3.5 CONTRASTE DE SIGNIFICACIÓN DE LOS COEFICIENTES

Si se quiere contrastar si una variable, o un grupo de variables de la ecuación es significativa se utilizan diferentes estadísticos:

Prueba individual

- Test de Wald

$H_o: \beta_j = 0$

$H_1: \beta_j \neq 0$

Estadistico de Wald : $\left(\frac{\hat{\beta}-\beta}{DT(\hat{\beta})}\right)^2 = \left(\frac{\hat{\beta}}{DT(\hat{\beta})}\right)^2 \sim \chi^2(1\ gl)$

Donde DT $(\hat{\beta})$ es el error estándar del coeficiente $\hat{\beta}$.

Prueba Conjunta

2.3.6 CONTRASTE DE SIGNIFICACIÓN PARA TODOS LOS β_j

Para poder validar el modelo se realizará la evaluación de la significancia del modelo pero a partir de los coeficientes en conjunto, es decir determinar si las variables independientes son significativas o no, se plantea las siguientes hipótesis:

$H_0 : \beta_2 = \beta_3 = \ldots = \beta_k = 0$

H_1 : Al menos un $\beta_j \neq 0$

Se utiliza el estadístico de razón de verosimilitud $RV_0 = -2\left[\ln L_0 - \ln L\right]$,

donde $\ln L$ es el logaritmo de la función de verosimilitud que se ha obtenido al estimar el modelo completo, mientras que $\ln L_0$ es el logaritmo de la función de verosimilitud al estimar el modelo con sólo el término independiente. Se cumple que $RV_0 = lnL_0 - lnL \sim \chi^2_{(k-1,\alpha)}$

cuando el tamaño de muestra (n) tiende al infinito.

Donde:

$\chi^2_{(k-1,\alpha)}$ es un valor crítico.

k = número de variables

α = nivel de significación

Si: $RV_0 > \chi^2_{(k-1,\alpha)}$ entonces se rechaza H_0, por lo tanto una o más de las variables independientes consideradas en el modelo son significativos.

2.3.7 BONDAD DE AJUSTE

- Prueba de Hosmer-Lemeshow

  La prueba de Hosmer-Lemeshow evalúa un aspecto de la validez del modelo: la calibración (grado en que la probabilidad predicha coincide con la observada).

  Para evaluar la bondad de ajuste del modelo se construye una tabla de contingencia, dividiendo la muestra en aproximadamente 10 grupos iguales a partir de las probabilidades estimadas, para comparar las frecuencias observadas con las esperadas en cada uno de estos grupos a través de la prueba χ^2 con j-2 grados de libertad, en donde j es el número de grupos formados.

  Se calcula los deciles de las probabilidades estimadas $\widehat{P_i}$; i=1,…,n y D_1,…, D_9 que son los deciles observados divididos en 10 grupos dados por:

  $A_j = \left\{i \in \{1, \dots, n\} / \widehat{P_i} \in [D_{j-1}, D_j]\right\},\ j = 1, \dots, 10$

  Donde : $D_0 = 0$, $D_{10} = 1$

  Sean:

  $$n_j = número\ de\ casos\ en\ A_j\ ;\ j = 1, \dots, 10$$

  $$o_j = número\ de\ y_i = 1\ en\ A_j\ ;\ j = 1, \dots, 10$$

  $\bar{P}\, j = \frac{1}{n_j} \sum_{i \in A_j} \widehat{P_i}$; j=1,…,10.

Las hipótesis a contrastar son:

H_o: El modelo de regresión Logistica se ajusta a los datos.

H_1: El modelo de regresión Logistica no se ajusta a los datos.

Estadístico de prueba es:

$$\chi^2 = \sum_{j=1}^{10} \frac{(o_j - n_j \bar{P}\, j)^2}{\bar{P}\, j n_j (1 - n_j)} \sim \chi^2_{(j-2,\alpha)}$$

Decisión: si $X^2 \geq \chi^2_{(j-2,\alpha)}$ rechazamos H_o y concluimos que el modelo no es adecuado a un nivel de significancia de α.

b) Regresión Logística Multinomial (RLM) La técnica RLM es un modelo lineal generalizado que consiste en la estimación de la probabilidad de que una observación pertenezca a cada uno de los grupos, dados valores de las p variables que conforman la observación.

El modelo compara G−1 categorías contra una categoría de referencia.

Dadas n observaciones $(\boldsymbol{y_i}, \boldsymbol{x_i})$ donde $\boldsymbol{x_i}$ es un vector con p variables y $\boldsymbol{y_i}$ es una v.a. independiente Multinomial con valores 1, 2,..., G, la cual indica el grupo al cual pertenece cada observación, la probabilidad condicional de pertenencia de $\boldsymbol{x_i}$ a cada grupo está dada por:

$$P(y = j/x_i) = \frac{e^{\alpha_{1j} + \beta'_{1j} x_i}}{1 + \sum_{k=2}^{G} e^{\alpha_{1j} + \beta'_{1k} x_i}} \qquad (1.5)$$

Donde $\alpha_{11} = 0$ y $\beta_{11} = 0$.

La regla de clasificación consiste en que dada una nueva observación con p variables se calcula la probabilidad de que ésta observación pertenezca a cada uno de los G grupos y luego se asigna al grupo que presento la mayor probabilidad.

La ventaja de RLM es que no requiere supuestos distribucionales como ADL y por lo tanto se puede aplicar a distribuciones multivariadas con variables cuantitativas y/o cualitativas.

2.4 ÁRBOL DE CLASIFICACIÓN

Se define un Árbol de Clasificación como una estructura en forma de diagramas de construcciones lógicas en las que las ramas representan conjunto de decisión. Estas decisiones generan sucesivas reglas para la clasificación de un conjunto de datos en subgrupos disjuntos y exhaustivos. Las ramificaciones se realizan en forma recursiva hasta que se cumplen ciertos criterios de parada.

El objetivo de estos métodos es obtener individuos más homogéneos con respecto a la variable que se desea discriminar dentro de cada subgrupo y heterogéneos entre los subgrupos. Para la construcción del árbol se requiere información de variables explicativas o predictoras a partir de las cuales se va a realizar la discriminación de la población en subgrupos.

El programa AID (Automatic Interaction Detection) de Sonquist, Baker y Morgan (1971), representa uno de los métodos de ajuste de los datos basados en modelos de Árboles de Clasificación. AID está basado en un algoritmo recursivo con sucesivas particiones de los datos originales en otros subgrupos menores y más homogéneos mediante secuencias binarias de particiones. Posteriormente surgió un sistema recursivo binario similar denominado CART (Classification And Regression Tree, Arboles de clasificación y regresión) desarrollado por Breiman en 1984. Un algoritmo recursivo de clasificación no binario, denominado CHAID (Chi Square Automatic Interaction Detection, Detección de Interacción Automática de Chi Cuadrada) fue desarrollado por Kass en 1980. Recientemente se han propuesto distintos métodos: FIRM propuesto por Hawkins, una simbiosis de construcción de árboles n-arios y análisis discriminante propuesto por Loh y Vanichsetakul y otra alternativa conocida como MARS (Multivariate Adaptive Regression Splines) propuesto por Friedman en 1991.

Entre las ventajas de esta técnica no paramétrica de clasificación están las siguientes:

- Las reglas de asignación son legibles y por tanto la interpretación de resultados es directa e intuitiva.
- Es una técnica no paramétrica que tiene en cuenta las interacciones que pueden existir entre los datos.
- Es robusta frente a datos atípicos o individuos mal etiquetados.

- Es válida para variables explicativas de naturaleza: continua, nominal u ordinal.
- Los Árboles requieren grandes masa de datos para asegurarse que la cantidad de observaciones de los nodos terminales es significativa.

Por el contrario este método de clasificación de los datos tienen las siguientes desventajas:

- Las reglas de asignación son fuertes y bastante sensibles a ligeras perturbaciones de los datos.
- Dificultad para elegir el árbol "óptimo".

a) El algoritmo CHAID

El algoritmo CHAID divide en grupos los registros que presenten la misma probabilidad de resultado, basándose en los valores de las variables independientes. El algoritmo parte de un nodo raíz y se va bifurcando en nodos descendientes hasta llegar a los nodos hoja, donde finaliza la ramificación.

La ramificación puede ser binaria, ternaria, etc. y viene determinada por la prueba Chi-cuadrado. Esta prueba se lleva a cabo mediante una tabulación cruzada entre el resultado y cada una de las variables independientes. El resultado es la probabilidad de que la hipótesis nula sea correcta, estas probabilidades se clasifican, y si el mejor (el valor más pequeño) se encuentra bajo un umbral determinado, se realiza una ramificación del nodo raíz en esa ubicación.

Pasos para elaborar un Árbol de Clasificación mediante el algoritmo CHAID

Pasos del Algoritmo CHAID en el cual se desea clasificar la variable Y y se tiene como variables explicativas $X_1, X_2, \ldots, X_k$:

1. Calcular la distribución de la variable respuesta Y en el nodo raíz.
2. Para cada variable explicativa X_I (I =1,2,…,k), hay que encontrar el par de categorías que tienen menores diferencias significativas respecto a la distribución de Y dentro del nodo. Es decir, aquel que tiene el mayor p-valor. Para calcular dicho p-valor depende del tipo de variables que estemos tratando en cada momento.
    a. La relación entre la variable explicativa X_I y la variable respuesta Y dentro del nodo se representa mediante una tabla de contingencia. Se consideran todas las

sub-tablas de contingencia posibles que se puedan formar con dos categorías de la variable explicativa.

b. El algoritmo identifica el par de categorías de X_I con mayor p-valor (p_I) asociado y lo compara con el nivel α predeterminado, normalmente $\alpha_{union} = 0.05$. Si el valor p-valor p_I es mayor que este valor α_{union} se agrupan dichas categorías. Se repite el literal a. considerando el par de categorías agrupadas como una única para calcular las sub-tablas de contingencia. En el caso de no obtener superar el valor de la α_{union} no se realiza ninguna agrupación de las categorías y se pasa al numeral 3.

c. De nuevo se selecciona el par de categorías con mayor p-valor y se compara con el valor α_{union}. Si es superior se vuelve a agrupar y se vuelve a calcular las sub-tablas de contingencia. El proceso termina en el caso en el cual el p-valor es inferior a α_{union} o se llega a dos categorías.

d. El algoritmo calcula un ajustado p-valor empleando las categorías agrupadas obtenidas de X_I y la categoría Y usando el ajuste de Bonferroni.

3. Los pasos del literal a. al d. se repiten de nuevo con el resto de variables explicativas.

4. El paso final es dividir el nodo basado en la variable explicativa, con las categorías agrupadas, con el menor p-valor ajustado si el valor es menor que el prefijado $\alpha_{separacion}$. En el caso de obtener un valor superior dicho nodo no se ramifica y será un nodo terminal.

5. Se continúa ramificando el árbol hasta que se satisfaga el criterio de parada.

2.4.1 DEFINICIÓN DE LAS VARIABLES

Para definir la variable dependiente se basa en la "unidad de medida" que puede ser de tipo Continua, Nominal u Ordinal, usando para cada una de ellas un estadístico distinto, si la variable Y es continua se utiliza la prueba F de Fisher, si la variable Y es nominal se usa la prueba chi-cuadrado de Pearson y finalmente si la variable Y es ordinal se usa la prueba de razón de verosimilitud así tenemos las siguientes variables.

2.4.2 VARIABLE DEPENDIENTE CONTINUA

Si la variable dependiente Y es continua, realizar un ANOVA para la prueba F, si las medias de Y para las diferentes categorías de X son las mismas. El ANOVA de la prueba F calcula los estadísticos y se derivan los $p - values$ como:

$$F = \frac{\sum_{i=1}^{I} \sum_{n \epsilon D} \frac{w_n f_n I(x_n = 1)(\bar{y}_i - \bar{y})^2}{I - 1}}{\sum_{i=1}^{I} \sum_{n \epsilon D} \frac{w_n f_n I(x_n = 1)(\bar{y}_i - \bar{y})^2}{N_f - 1}}$$

$$p = P_r\big(F(I - 1, N_f - I) > F\big)$$

Donde

$$\bar{y}_i = \frac{\sum_{n \epsilon D} W_n f_n y_n I(x_n = i)}{\sum_{n \epsilon D} W_n f_n I(x_n = i)}, \quad \bar{y}_i = \frac{\sum_{n \epsilon D} W_n f_n y_n}{\sum_{n \epsilon D} W_n f_n}, \quad N_f = \sum_{n \epsilon D} f_n$$

y $F\big(I - 1, N_f - I\big)$ es la variable aleatoria de la Distribución F con grados de libertad $I - 1$ y $N_f - I$.

2.4.3 VARIABLE DEPENDIENTE NOMINAL

Si la variable dependiente Y es de tipo nominal, se prueba la hipótesis nula de independencia de X con Y. Se realizará una tabla de contingencia que es un cuadro formado en clases, mediante la variable Y en las columnas y las variables X en las filas. Se calcula las frecuencias de las clases bajo la hipótesis nula. Las frecuencias observadas son calculadas por el estadístico Pearson Chi – Cuadrado de Pearson o estadístico de probabilidad del ratio. El $p - value$ se calcula de acuerdo a los dos estadísticos.

$$\chi^2 = \sum_{j=1}^{J} \sum_{i=1}^{I} \frac{\big(n_{ij} - \hat{m}_{ij}\big)^2}{\hat{m}_{ij}}$$

$$G^2 = 2 \sum_{j=1}^{J} \sum_{i=1}^{I} n_{ij} \ln\big(n_{ij} - \hat{m}_{ij}\big)$$

Donde $n_{ij} = \sum_{n \epsilon D} f_n I(x_n = i \wedge y_n = j)$ y $\hat{m}$ es el estimador, el $p - value$ correspondiene es dado por $p = (\chi_d^2 > \chi^2)$ para la prueba Chi – Cuadrado de Pearson o $p = \big(\chi_d^2 > G^2\big)$ para la prueba de probabilidad del Ratio, donde χ_d^2 es la distribución Chi – Cuadrado de Pearson con grados de libertad $d = (J - 1)(I - 1)$.

2.4.4 VARIABLE DEPENDIENTE ORDINAL

Si la variable dependiente Yes categórica ordinal, la hipótesis nula de independencia de X e Y es probada con el modelo de efectos filas propuestas por Goodman (1979). Dos frecuencias esperadas de celdas $\hat{m}_{ij}$ y $\hat{\hat{m}}_{ij}$ son estimados. El estadístico de prueba y el $p - value$ son:

$$H^2 = 2\sum_{j=1}^{J}\sum_{i=1}^{I} n_{ij} \ln\left(\hat{\hat{m}}_{ij}/\hat{m}_{ij}\right)$$

$$p = Pr\left(\chi^2_{I-1} > H^2\right)$$

2.4.5 SELECCIÓN DE VARIABLES – REGRESIÓN LOGÍSTICA

La regresión Logística es una técnica multivariada por medio de la cual se analizan las relaciones de asociación entre una variable dependiente categórica dicotómica Y y una o varias variables independientes o predictoras X cuantitativas o categóricas. La función está dada por:

$$p_i = \frac{e^{\beta_0+\beta_i x_i}}{1 + e^{\beta_0+\beta_i x_i}}$$

donde p_i será un valor entre cero y uno, siendo esta la probabilidad de que el evento $Y = 1$ ocurra.

Métodos de Selección de Variables

Existen varios métodos para construir el modelo de regresión, es decir, para seleccionar de entre todas las variables que introduciremos en el modelo, cuáles son las que necesitamos para explicarlo. El modelo de regresión se puede construir utilizando las siguientes técnicas:

- **Técnica de pasos hacia adelante (Forward):** consiste en ir introduciendo las variables en el modelo únicamente si cumple una serie de condiciones hasta que no se pueda introducir ninguna más, hasta que ninguna cumpla la condición impuesta.
- **Técnica de pasos hacia atrás (Backward):** se introducen en el modelo todas las variables y se van suprimiendo si se cumplen una serie de condiciones definidas de priori hasta que no se puedan eliminar más, es decir ninguna variable cumpla la condición impuesta.
- **Técnica por pasos (Stepwise):** combina los dos métodos anteriores, adelante y atrás introduciendo o eliminando variables del modelo si cumplen una serie de condiciones

definidas a priori hasta que ninguna variable satisfaga ninguna de las condiciones expuestas de entrada o salida del modelo.

- **Técnica de introducir todas las variables obligatoriamente (Enter):** esta última técnica de selección de variables para construir el modelo de regresión, produce que el proceso de selección de variables sea manual, partiendo de un modelo inicial, en el que se obliga a que entren todas las variables seleccionadas, se va evaluando que variable es la que menos participa en él y se elimina volviendo a construir un modelo de regresión aplicando la misma técnica, pero excluyendo la variable seleccionada y aplicando el mismo proceso de selección. Este proceso se repite reiteradamente hasta que se considere que el modelo obtenido es el que mejor se ajusta a las condiciones impuestas y que no se pueda eliminar ninguna variable más de las que lo componen.

2.4.6 EVALUACIÓN DE LAS FUNCIONES DE CLASIFICACIÓN

Validación cruzada

Es una técnica que se aplica al conjunto de datos inicial y se usa para evaluar de forma más o menos exacta la calidad de un modelo de clasificación o de regresión. A continuación los tipos de validación cruzada:

Validación hold-out

La validación hold-out no se considera una validación cruzada como tal, ya que los datos nunca se cruzan. El conjunto de datos inicial se divide aleatoriamente en dos subconjuntos; el primero (D_a) se usa para la fase de entrenamiento del modelo y el segundo (D_p) es útil para evaluar su calidad. Normalmente se usa menos de la tercera parte del conjunto inicial para formar parte de D_p.

Validación cruzada en k-pasos

En la validación cruzada en k-pasos, el conjunto de datos inicial se particiona en k subconjuntos. En cada paso, se asigna a D_p un subconjunto de datos distinto y a D_a los $k - 1$ subconjuntos restantes. El proceso de validación cruzada se repite k veces, tomando en cada de ellas un subconjunto de datos de validación distinto y el resto como datos de entrenamiento.

Por lo general se usan K=10 partes y es llamado “10 fold cross validation”.

Validación cruzada leave-one-out

Este tipo de validación es un caso particular de la validación cruzada en k-pasos, en el que k coincide con el número de casos n del conjunto de datos. Estaríamos ante el caso más extremo, ya que cada modelo estaría entrenado con $n-1$ casos y probado con el caso restante

no usado. Este es el método de validación que devuelve resultados más exactos, pero su gran inconveniente es la necesidad de entrenar tantos modelos como casos tuviera el conjunto de datos inicial y el coste computacional necesario para llevar a cabo el proceso puede llegar a ser muy elevado.

2.5 ANÁLISIS DISCRIMINANTE NO MÉTRICO (ADNM)

Raveh (1983,1989) propuso un procedimiento denominado Análisis Discriminante No Métrico (ADNM) basado en la maximización de un índice de separación entre dos grupos. Guttman (1988) generalizó el índice propuesto por Raveh para múltiples grupos y lo llamó disco (discrimination coefficient).

El ADNM tiene como objetivo determinar una función discriminante de tal manera que se maximice el cociente entre la variabilidad entre grupos con la variabilidad dentro de grupos. Si se tienen G grupos p-variados X(1),X(2),...,X(G) cada uno con n_1,,...,n_G observaciones[4], el elemento denotado por $X_i(j)$ corresponde a la observación i del grupo j.

Al conjunto formado por las observaciones de todos los grupos se denomina Conjunto de Entrenamiento. Sea η un vector p-dimensional, y $z_i(g)$ la variable aleatoria definida así: $z_i(g)= \eta' X_i\ (g)$ que representa el Score de la i-ésima observación del grupo g-ésimo dado por η. El índice disco entre G grupos está dado por:

$$\text{disco} = \frac{\sum_{g=1}^{G}\sum_{h=1}^{G} n_g h_h |\bar{z}(g) - \bar{z}(h)|}{\sum_{g=1}^{G}\sum_{h=1}^{G}\sum_{i=1}^{n_g}\sum_{j=1}^{n_h} |z_i(g) - z_j(h)|}, \qquad (1.6)$$

donde $\bar{z}(g)$ representa el promedio de los scores para las observaciones del grupo g.

El numerador en (1.6) se puede escribir como

$$\sum_{g=1}^{G}\sum_{h=1}^{G}\left|\sum_{i=1}^{n_g}\sum_{j=1}^{n_h}\left[z_i(g) - z_j(h)\right]\right|,$$

4 En total n observaciones, $n=\sum_{i=1}^{G} n_i$

donde el valor absoluto corresponde a una medida de la separación entre los grupos h y g. El denominador en (1.6) contiene el elemento

$$\sum_{i=1}^{n_g}\sum_{j=1}^{n_h}\left[z_i(g)-z_j(h)\right],$$

que representa la variación total entre los grupos h y g, cuantificado mediante desviaciones absolutas.

En virtud de la siguiente desigualdad

$$\sum_{g=1}^{G}\sum_{h=1}^{G}\left|\sum_{i=1}^{n_g}\sum_{j=1}^{n_h}\left[z_i(g)-z_j(h)\right]\right| \leq \sum_{g=1}^{G}\sum_{h=1}^{G}\sum_{i=1}^{n_g}\sum_{j=1}^{n_h}\left|z_i(g)-z_j(h)\right|,$$

se obtiene que la ecuación (1.6) satisface $0 \leq \text{disco} \leq 1$.

El coeficiente disco es igual a cero si y solamente si todos los grupos tienen la misma media, y es igual a 1 si no existe superposición entre los scores de ningún par de grupos.

En virtud de $z_i(g)=\eta' X_i\,(g)$, la ecuación (1.6) puede escribirse de la siguiente manera:

$$\text{disco} = \frac{\sum_{g=1}^{G}\sum_{h=1}^{G} n_g n_h \left|\eta'[\bar{X}(g)-\bar{X}(h)]\right|}{\sum_{g=1}^{G}\sum_{h=1}^{G}\sum_{i=1}^{n_g}\sum_{j=1}^{n_h}\left|\eta'\left[X_i(g)-X_j(h)\right]\right|}, \qquad (1.7)$$

El ADNM propuesto por Raveh (1989) consiste en la búsqueda de η tal que maximice el coeficiente disco dado en la ecuación (1.5).

El disco en (1.7) puede ser escrito en forma matricial, para esto se definen dos matrices B_{gh} y $V_{gh}(i,j)$ ambas de orden p×p de la siguiente manera:

$$B_{gh} = [\bar{X}(g)-\bar{X}(h)][\bar{X}(g)-\bar{X}(h)]', \quad (1.8)$$

$$V_{gh}(i,j) = \left[X_i(g)-X_j(h)\right]\left[X_i(g)-X_j(h)\right]'. \quad (1.9)$$

Usando la identidad |a|= $a^2/|a|$ con a≠0 se tiene que:

$$|n'[\overline{X}(g) - \overline{X}(h)]| = \frac{\eta' B_{gh}(i,j)\eta}{\sqrt{\eta' B_{gh}(i,j)\eta}} \quad \text{si } B_{gh} \neq 0.$$

Además

$$|n'[X_i(g) - X_i(h)]| = \frac{\eta' V_{gh}(i,j)\eta}{\sqrt{\eta' V_{gh}(i,j)\eta}} \text{ si } V_{gh}(i,j) \neq 0.$$

De esta manera disco en (1.7) puede ser representado en notación matricial como una función del vector η así:

$$\text{disco}(\eta) = \frac{u(\eta)}{l(\eta)} = \frac{\eta' B(\eta)\eta}{\eta' V(\eta)\eta}. \quad (1.10)$$

donde B(η) y V(η) son matrices simétricas de orden p×p que sólo dependen del parámetro η de la siguiente manera:

$$B(\eta) = \sum_{g=1}^{G}\sum_{h=1}^{G} \frac{\eta_g \eta_h B_{gh}}{\sqrt{\eta' B_{gh}\eta}}. \quad (1.11)$$

y

$$V(\eta) = \sum_{g=1}^{G}\sum_{h=1}^{G}\sum_{i=1}^{n_g}\sum_{j=1}^{n_h} \frac{V_{gh}(i,j)}{\sqrt{\eta' V_{gh}(i,j)\eta}}. \quad (1.12)$$

Para maximizar el disco en (1.7) con respecto a η, Choulakian y Almhana(2001) propusieron el siguiente algoritmo:

Algoritmo

1. Se comienza con $\eta_o = \theta^*$ siendo θ^* el eigenvector obtenido mediante Análisis Discriminante Lineal (DL_1).

2. Calcular $\eta_{k+1} = \eta_k\ [1-2\times\text{disco}(\eta_k)]+2\times V(\eta_k)^{-1}B(\eta_k)\eta_k$, para k =0,1,2,...

3. El proceso se detiene cuando $|\text{disco}(\eta_{k+1}) - \text{disco}(\eta_k)| \leq \varepsilon$ donde ε es un valor real positivo definido con anterioridad, por ejemplo, $\varepsilon = 10^{-5}$.

4. El valor óptimo de la función discriminante η se obtiene haciendo $\eta = \eta_k$.

La convergencia del algoritmo anterior fue demostrado por Choulakian y Almhana (2001).

Luego de encontrar el valor óptimo de η éste puede ser utilizado para clasificar nuevas observaciones en uno de los G grupos. Para realizar la clasificación se determinan G−1 puntos de corte (PC) de la siguiente manera:
Se toman los n discos scores $z_i(g)$ con i = 1, 2, . . . , n_g y con g = 1, 2, . . . ,G; sin pérdida de generalidad, se puede suponer que los primeros n_1 scores son menores que los segundos n_2 y así sucesivamente. El punto de corte (PC) que separa los grupos 1 y 2 es igual al percentil $100(n_1/n)$% de los n scores ordenados, el segundo (PC) que separa los grupos 2 y 3 es igual al percentil $100((n_1 + n_2)/n)$% de los n scores ordenados, de manera similar se obtienen los G − 3 (PC) restantes.

Si los n scores de los G grupos no se superponen la anterior regla indica que el punto de corte (PC) entre el g-ésimo grupo y el (g + 1)-ésimo grupo seria:

$$\frac{\max_{i=1,\ldots,n_g} z_i(g) + \min_{i=1,\ldots,n_g} z(g+1)}{2},$$

lo cual asegura que la tasa de clasificación errónea de las observaciones en el conjunto de entrenamiento sea cero.

2.5.1 Observaciones del ADL y ADNM

Raveh (1989) realizó estudios de simulación bastante extensos y mostró que el análisis discriminante lineal es sólo ligeramente mejor que el análisis discriminante no métrico en el caso de distribuciones multinormales con matrices de covarianza iguales o con diferencias moderadas entre covarianzas. Para las distribuciones normales con matrices de covarianza muy diferentes y para las distribuciones lognormal y chi cuadrado (ambas sesgadas), el análisis discriminante no métrico produce menos clasificaciones erróneas que el análisis discriminante lineal.

2.5.2 Prueba de la convergencia del algoritmo

Primero calculamos el gradiente de disco (η) en (1.10) y proporcionamos una prueba de la convergencia del algoritmo. Los detalles de la prueba se encuentra en el anexo 4.

Proposición 1. El gradiente del disco es

$$\nabla disco(\eta) = \frac{[B(\eta)\eta - disco(\eta)V(\eta)\eta]}{l(\eta)}. \quad (1.13)$$

Donde $B(\eta)$, $V(\eta)$ y $l(\eta)$ son dados en (1.11), (1.12) y (1.10) respectivamente.

Corolario 1. Una condición necesaria que η^* maximiza disco (η) es que η^* es un autovector del problema autovalor-autovector no lineal

$$V(\eta^*)^{-1}B(\eta^*)\eta^* = disco(\eta^*)\eta^*. \quad (1.14)$$

La k-ésima iteración del método clásico del ascenso más generalizado es dada por

$$\eta_{k+1} = \eta_k + s_k V(\eta_k)^{-1}\nabla disco(\eta_k). \quad (1.15)$$

donde $V(\eta_k)^{-1}$ es el modificador de la matriz definida positiva para el k-esimo paso, $\nabla disco(\eta_k)$ es el gradiente del disco calculado en $\eta = \eta_k$, y s_k es un escalar positivo, elegido tal que el $disco(\eta_k + s_k V(\eta_k)^{-1}\nabla disco(\eta_k)) > disco(\eta_k)$. Ver por ejemplo Murray (1972a).

Observación. Si hacemos $s_k = 2l(\eta_k)$, entonces (1.15) será equivalente a nuestro algoritmo descrito anteriormente.

Sea

$$g_\alpha(\eta) = \eta + \alpha V(\eta)^{-1}[B(\eta)\eta - disco(\eta)V(\eta)\eta]. \quad (1.16)$$

dónde α es una constante. Nuestro algoritmo corresponde a la iteración vectorial $\eta_{k+1} = g_2(\eta_k)$

Para demostrar su convergencia, mostraremos que la función vectorial $g_2(\eta)$ es un mapeo de contracción. Una condición adecuada para un mapeo de contracción es que la norma-2 del Jacobiano de $g_2(\eta)$ es estrictamente menor que uno, es decir, $\|\nabla g_2(\eta)\|_2 < 1$. Véase, por ejemplo, Hager (1988), o Phillips y Taylor (1996). La norma 2 de una matriz A de orden m x n está definido por

$$(\|A\|_2)^2 = \max[y'(A'A)y: \ \|y\|_2 = 1], \qquad (1.17)$$

donde y ∈ R^n y $(\|y\|_2)^2 = y'y$. La ecuación (1.13) muestra que si A y A' tienen los mismos valores propios, entonces $\|A\|_2$= es el máximo de los valores absolutos de los valores propios de A.

Teorema 1. (a) $\nabla g_{\alpha}(\eta) = I[1 - \alpha disco(\eta)] - \alpha\eta\nabla disco(\eta)'$; donde I es la matriz de identidad; (b) $\|\nabla g_{\alpha}(\eta)\|_2 = |1 - \alpha disco(\eta)|$; (c) Para 0 < $disco(\eta)$ <1 y para 0 < $\alpha \leq 2$, $\|\nabla g_{\alpha}(\eta)\|_2 < 1$; y consecuentemente $max_{\alpha}\|\nabla g_{\alpha}(\eta)\|_2$ se alcanza en $\alpha = 2$.

El teorema muestra que el tamaño de paso, $s_k = 2l(\eta_k)$, es el valor mayor para que el proceso iterativo (1.15) converja.

2.6 COMPARACIONES ENTRE ADL Y RL

Efron (1975) comparó las dos técnicas para el caso de dos grupos con igual matriz de covarianzas y encontró que la eficiencia relativa asintótica (ERA) de RL con respecto a ADL está entre un medio y dos tercios. Crawley (1979) comparó RL con ADL para muestras pequeñas con dos grupos y encontró que para el caso de matrices de covarianza iguales ADL tiene un mejor desempeño que RL en el proceso de clasificación, para el caso de matrices de covarianzas diferentes RL tuvo ligeramente un mejor desempeño y para el caso de dos poblaciones distribuidas no normal RL tuvo un desempeño muy superior a ADL. Harrell & Lee (1985) realizaron una comparación entre las técnicas para el caso de dos grupos considerando normalidad con matrices de covarianzas iguales, tamaños de muestra de 50 y 130 con seis distancias de Mahalanobis entre los vectores de medias de las dos poblaciones que variaron entre los valores de 0.94 y 4.68; en este estudio se encontró que el desempeño de ADL fue mejor que RL pero que las diferencias no eran significativas. Pohar et al. (2004) llevaron a cabo un estudio donde compararon ADL y RL por medio de simulación. Para comparar los desempeños de cada una de las técnicas utilizaron el índice típico de tasa de clasificación errónea y los índices A, B, C y Q propuestos por Harrell & Lee (1985). Estos indices son los mejores y más eficientes criterios para las comparaciones y nos dicen qué tan bien los modelos discriminan entre los grupos y / o qué tan buena es la predicción.

La visión teórica y las experiencias con simulaciones revelaron que algunos índices son más y otros menos apropiados en diferentes suposiciones. Esta investigación se centró en tres medidas de precisión predictiva, los índices B, C y Q. El índice C es puramente una medida de discriminación (la discriminación se refiere a la capacidad de un modelo para discriminar o separar valores de Y) y no una medida de exactitud de la predicción. Un índice A y C de valor 1 indica la discriminación perfecta; un índice C de 0,5 indica una predicción aleatoria.

Los índices B y Q se utilizaron para evaluar la precisión de la predicción de resultados. El índice B mide un promedio de la diferencia cuadrada entre un valor estimado y un valor real. Los valores del índice B están en el intervalo [0,1], donde 1 indica predicción perfecta. En el caso de la predicción aleatoria en dos grupos de igual tamaño, el valor del índice B es 0,75. El índice Q es similar al índice B y también es una medida de precisión predictiva. Una puntuación de 1 del índice Q indica predicción perfecta. Un índice Q de 0 indica predicciones aleatorias y valores menores que 0 indican peores predicciones aleatorias.

La comparación se inició en un escenario en el cual se cumplían los supuestos de ADL y luego realizaron cambios en los tamaños de muestra, matriz de covarianzas y distancia de Mahalanobis entre las medias de los grupos simulados. Se encontró que los desempeños de ADL y RL fueron muy cercanos, siempre y cuando los supuestos de normalidad no sean afectados fuertemente, y presentaron lineamientos para identificar este tipo de situaciones; adicionalmente, discutieron las situaciones donde es inapropiado utilizar ADL para clasificación.

2.7 COMPARACIONES ENTRE ADL Y RLM

Shelley y Donner (1987) llevaron a cabo un estudio para medir la eficiencia relativa asintótica (ERA) de Regresión Logística Multinomial (RLM) comparada con análisis discriminante para el caso de poblaciones distribuidas normal multivariada con igual matriz de varianzas y covarianzas.

Los casos que estudiaron fueron con dos, tres y cuatro grupos. Además, tuvieron en cuenta varias separaciones entre los grupos y estudiaron el efecto de la colinealidad entre los vectores del modelo de regresión logística sobre la ERA. Los autores encontraron que para el caso de vectores de clasificación colineales la ERA cambió de 50% a 65% para dos grupos y de 35% a 95% para el caso de cuatro grupos cuando la distancia entre el grupo de referencia y los demás estuvo en 3.0 a 3.5. Para el caso de vectores ortogonales se encontró que ERA decae rápidamente a medida que aparecen más grupos en el proceso de clasificación.

2.8 COMPARACIONES ENTRE ADL Y ADNM

Raveh (1989) llevó a cabo un estudio de simulación donde comparó ADL y ADNM para el caso de dos grupos; se consideraron tres escenarios o tipos de distribuciones de probabilidad multivariada para cada grupo: normal multivariada, log-normal y chi-cuadrada; en cada uno

de estos escenarios se consideraron distribuciones bivariadas y trivariadas. El tamaño de muestra fue siempre de 50 observaciones para cada uno de los dos grupos. El objetivo básico del estudio fue comparar el desempeño de las dos técnicas usando la tasa de clasificación errónea para el conjunto de entrenamiento y para un nuevo conjunto de validación obtenidos de la misma distribución. Para el caso de dos grupos provenientes de una distribución normal multivariada (2 ó 3 variables) se encontró que cuando hay igualdad entre las matrices de covarianzas ADL tiene tasas de clasificación erróneas como máximo 1 % mejores que las de ADNM. Se encontró también que a medida que las matrices de covarianza difieren entre sí, la ventaja de ADL disminuye hasta el punto que ADNM obtiene menor tasa de clasificación errónea para el caso extremo de matrices de covarianza. Para el caso de grupos provenientes de una distribución log-normal se encontró que ADNM es muy superior que ADL; el desempeño de ADNM estuvo por encima de ADL en 16 % para conjuntos de entrenamiento y 14 % para conjuntos de validación. Para este mismo caso se halló que, a medida que los parámetros de la distribución log-normal para cada grupo difieren, el desempeño de ADNM mejora sobre el de ADL. Para el caso de grupos provenientes de una distribución chi-cuadrado se encontró que ADL tuvo un desempeño similar a ADNM; las diferencias entre las tasas de clasificación erróneas fueron 1 % a favor de ADNM; se observó también que la ventaja de ADNM sobre ADL se incrementaba ligeramente a medida que disminuían los grados de libertad de la distribución. Choulakian & Almhana (2001) realizaron una comparación entre ADL y ADNM usando tres conjuntos de datos: poultry data, encontrado en Raveh (1983), conformado por diez grupos con cuatro variables; wolf skull data, encontrado en Morrison (1990), conformado por cuatro grupos y nueve variables, y feelings data, encontrado en Hand (1989), conformado por cuatro grupos y veinticinco variables. En cada una de estas tres aplicaciones se construyó la función discriminante para el ADNM con base en la función discriminante de ADL y se encontró que ADNM clasifica mejor el conjunto de entrenamiento, también se halló un aumento en el coeficiente de discriminación (Disco) para cada una de las aplicaciones: para la primera aplicación el valor del Disco encontrado por Raveh de 0.9915 paso a 0.9935 esto con el algoritmo empleado por Choulakian & Almhana, para la segunda aplicación el valor del Disco encontrado por Raveh Disco cambió de 0.987 a 1 esto con el algoritmo empleado por Choulakian & Almhana y para la última aplicación ocurrió lo mismo el valor del Disco de Raveh pasó de 0.9615 a 0.9832, esto debido a que el punto de partida del algoritmo tiene un gran poder discriminatorio.

III. MATERIALES Y MÉTODOS

3.1 MATERIALES:

Los materiales y equipos de los que se hizo uso en la presente tesis:

1. Una laptop marca Toshiba con un procesador Intel ® core^{TM}CPU @2.5 GHz 2.5 GHz, con 12 GB de memoria RAM y un sistema operativo Windows 10 de 64 bits.
2. Un usb de 32 GB de velocidad 3.0 en la modalidad readyboost para aumentar la memoria caché.
3. El programa estadístico R versión 3.3.3.
4. Los paquetes de R:CHAID, nnet.
5. El programa del Análisis Discriminante No Métrico (ADNM) del algoritmo propuesto por Choulakian y Almhana.

3.2 DESCRIPCIÓN DE LOS CONJUNTO DE DATOS

En este trabajo de investigación se utilizaron conjuntos de datos de la Universidad de California Irving (UCI) que a continuación se describen:

1. **Crabs.** Es un conjunto de datos de tamaño 200 con 5 atributos cuantitativos, 0 atributos cualitativos y 2 clases.
2. **Ecoli.** Es un conjunto de datos de tamaño 255 con 6 atributos cuantitativos, 0 atributos cualitativos y 3 clases.
3. **Glass Identification.** Es un conjunto de datos de tamaño 146 con 4 atributos cuantitativos, 0 atributos cualitativos y 2 clases.
4. **Bupa.** Es un conjunto de datos de tamaño 345 con 6 atributos cuantitativos, 0 atributos cualitativos y 2 clases.
5. **Electrode.** Es un conjunto de datos de tamaño 100 con 5 atributos cuantitativos, 0 atributos cualitativos y 2 clases.
6. **Iris.** Es un conjunto de datos de tamaño 150 con 4 atributos cuantitativos, 0 atributos cualitativos y 3 clases.

7. **Titanic.** Es un conjunto de datos de tamaño 300 con 3 atributos cuantitativos, 0 atributos cualitativos y 2 clases.
8. **Banana**. Es un conjunto de datos de tamaño 400 con 2 atributos cuantitativos, 0 atributos cualitativos y 2 clases.

Cuadro 1: Resumen de los datos utilizados

Data	Número observaciones	Variables cuantitativas	Variables cualitativas	Clases
Crabs	200	5	0	2
Ecoli	255	6	0	3
Glass identificacion	146	4	0	2
Bupa	345	6	0	2
Electrode	100	5	0	3
Iris	150	4	0	3
Titanic	300	3	0	2
Banana	400	2	0	2

FUENTE: Elaboración propia

3.3 DISEÑO DE INVESTIGACIÓN

Se desarrolló un diseño no experimental transversal debido a que no se manipulan las variables sólo se observan y es del tipo descriptivo porque busca indagar cuál de las tres técnicas clasifica mejor los individuos en clases tomando en cuenta la Validación Cruzada.

3.4 PROCEDIMIENTO DE ANÁLISIS DE DATOS

Se comparó los tres métodos de clasificación el Análisis Discriminante no Métrico (ADNM) con la Regresión Logística Multivariada (RLM) y el Árbol de Clasificación CHAID.

En el cuadro N°1 se describió cada conjunto de datos utilizados, incluyendo el número de datos (tamaño), variables cuantitativas, variables cualitativas y las clases.

IV. RESULTADOS Y DISCUSIÓN

A continuación se presenta los resultados obtenidos

Cuadro 2: Comparación de clasificadores tomando la tasa promedio de clasificación errónea

CONJUTO DE DATOS	Tasa de mala clasificación (%)		
	ADNM	RL	CHAID
Crabs	0.06639	0.04333	3.53528
Ecoli	0.91133	0.65657	0.75371
Glass identificacion	2.50376	2.50361	3.47529
Bupa	1.12398	1.06937	1.31762
Electrode	0.52667	0.45111	8.08222
Iris	0.33284	0.57728	4.53284
Titanic	1.05679	1.03988	1.15914
Banana	1.18813	1.24361	1.17159

FUENTE: Elaboración propia

Los cálculos de las tasas de mala clasificación de los ocho conjuntos de datos se realizo tomando la validación cruzada con k=10 (10 particiones y 10 iteraciones) debido a que la dispersión es menor a comparación de otros valores que puede tomar k. Finalmente se calculó la media aritmética de las 10 iteraciones para obtener un único resultado.

En esta comparación de clasificadores se puede observar que la técnica de RLM con la data Bupa la cual tiene 6 variables cuantitativas generó la menor tasa de clasificación errónea con un 1.1% seguida del ADNM y CHAID con diferencias del 0.2% entre ellas. En los demás casos cuando se disminuye el número de variables se puede apreciar una disminución

en la tasa de clasificación errónea a excepción de Glass identificación, Titanic y Banana en la cual el mejor clasificador sigue siendo la RLM seguido del ADNM.

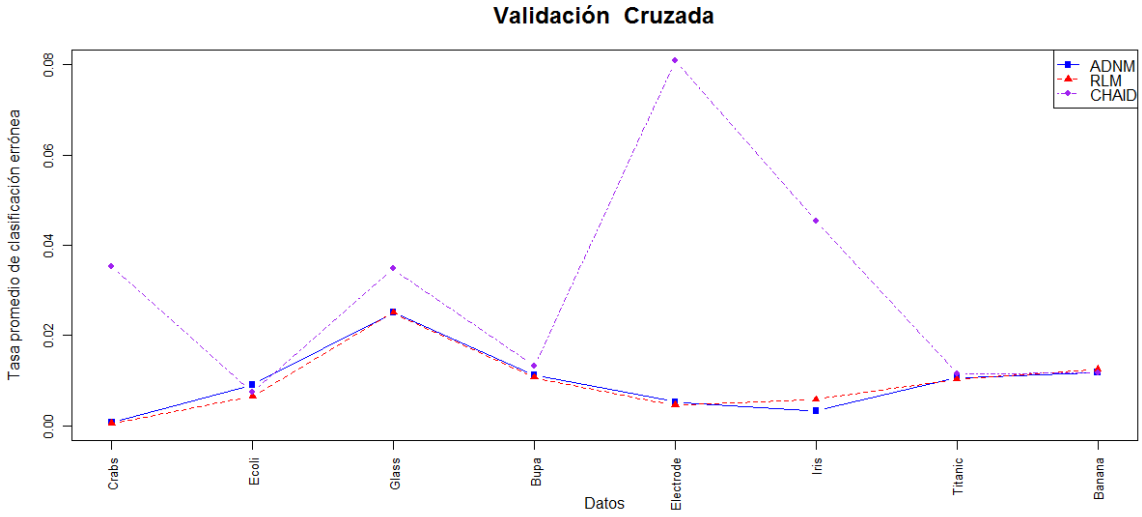


Figura 2: Gráfica de los métodos de clasificación tomando la tasa promedio de clasificación errónea

En la gráfica se puede apreciar que en la mayoría de los casos la RLM presenta la menor tasa de clasificación que el ADNM a excepción de la data iris además el CHAID es la técnica que peor clasifica comparando numéricamente las tasas de clasificación errónea.

En las siguientes figuras se muestra las matrices de correlación de los datos trabajos en esta tesis.

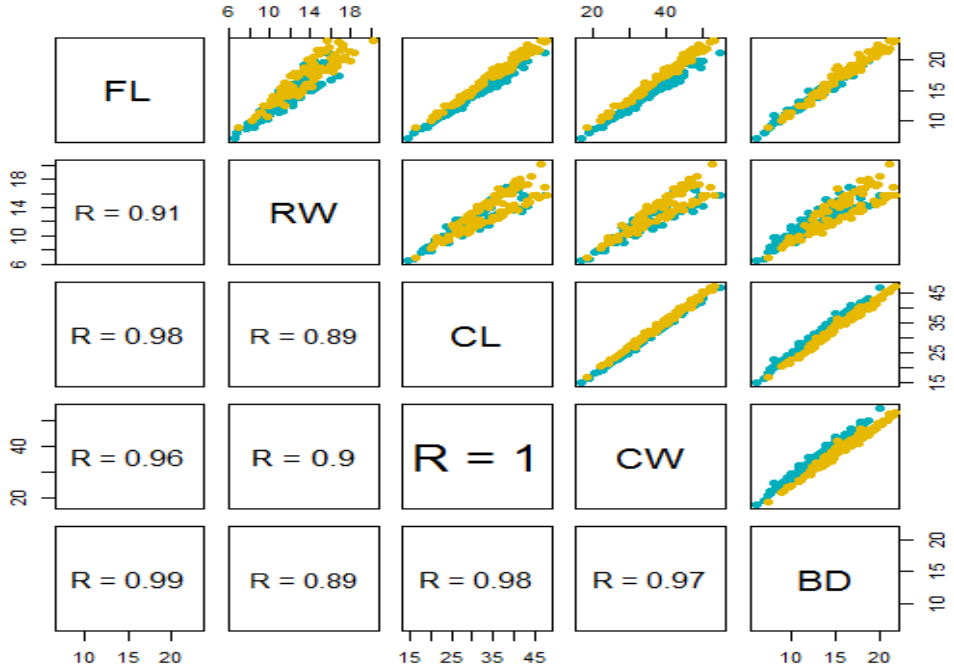


Figura 3: Gráficos de dispersión de las variables de la data Crabs

Las variables tamaño del lóbulo frontal del cangrejo (FL), ancho del trasero (RW), longitud del caparazón (CL), ancho del caparazón (CW) y profundidad del cuerpo (BD) están muy correlacionadas positivamente. Esta relación nos indica que si crece la variable tamaño del lóbulo frontal del cangrejo también crecerá la variable ancho del trasero, la misma relación sucede con la variable longitud del caparazón, ancho del caparazón y profundidad del cuerpo. De igual manera si crece la variable ancho del trasero (RW) también crecerá la variable longitud del caparazón (CL), ancho del caparazón (CW) profundidad del cuerpo (BD) y lóbulo frontal del cangrejo (FL).

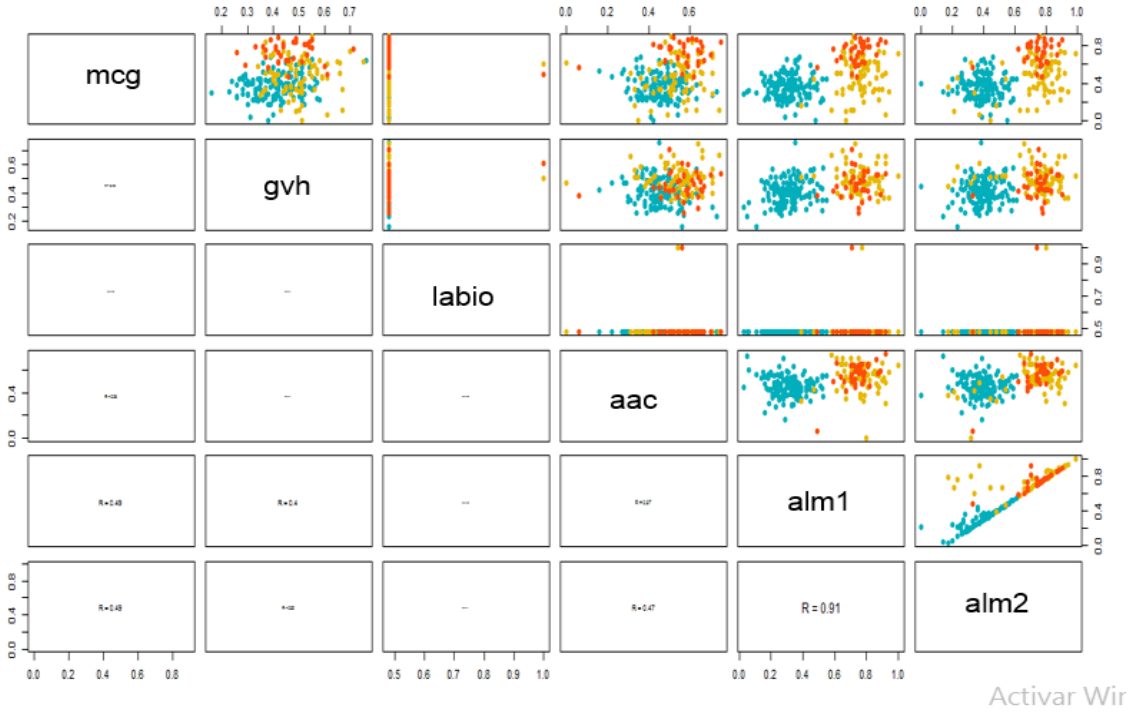


Figura 4: Gráficos de dispersión de las variables de la data Ecoli

La variable puntaje del programa de predicción de la región de extensión de membrana de ALOM (alm1) está correlacionada positivamente con la variable puntuación del programa ALOM después de excluir las regiones putativas de la señal escindible de la secuencia (alm2). Esta relación nos indica que si aumenta el puntaje del programa de predicción de la región de extensión de membrana de ALOM la puntuación del programa ALOM después de excluir las regiones putativas de la señal escindible de la secuencia también aumentará.

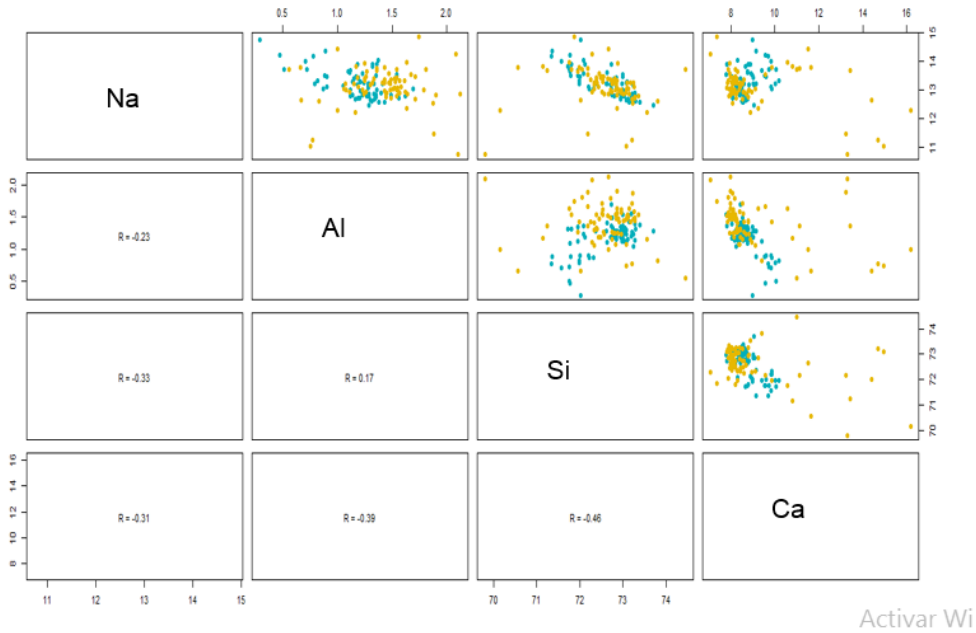
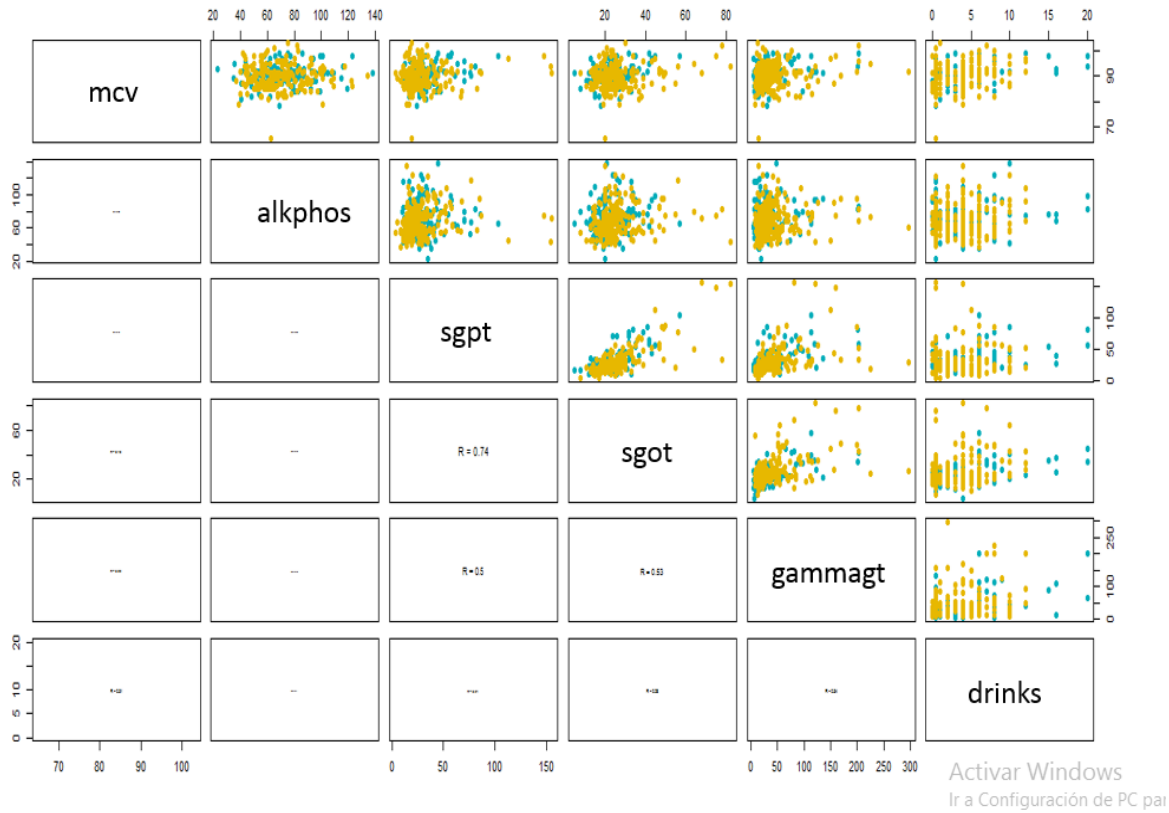


Figura 5: Gráficos de dispersión de las variables de la data Glass identificación

Las variables Sodio (NA), Aluminio (Al), Silicio (Si), Calcio (Ca) tienen una correlación baja, lo que implica que el aumento en una de las variables no necesariamente implica el aumento en la otra variable.



22

Figura 6: Gráficos de dispersión de las variables de la data Bupa

La variable alamina aminotransferasa (sgtp) la cual es una enzina concentrada en el higado está correlacionada positivamente con la variable aspartato aminotransferasa (sgot) enzima concentrada especialmente en el corazón. Esta relación nos indica que si aumenta las enzinas concentrada en el higado también aumentará enzimas concentradas en el corazón.

También se puede observar que la variable aspartato aminotransferasa (sgot) está muy correlacionada positivamente con la variable gamma-glutamil transpeptidasa (gammagt)la cual es una enzima hepática cuya función es el metabolismo del glutatión.

Esta relación nos indica que si aumenta el aspartato aminotransferasa también aumentará las enzimas hepáticas es decir aumentará el metabolismo del glutatión.

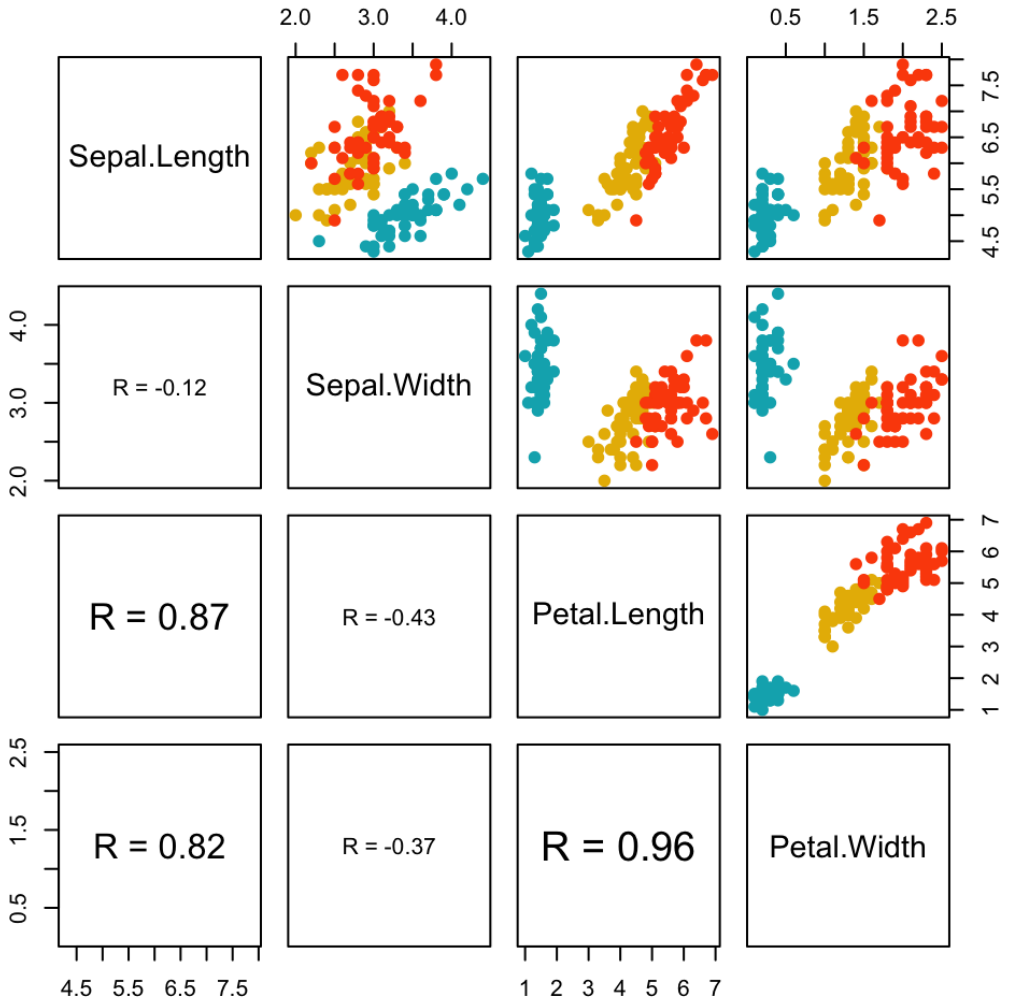


Figura 7: Gráficos de dispersión de las variables de la data Iris

La variable longitud del sépalo está muy correlacionada positivamente con las variables longitud del pétalo y ancho de pétalo. Esta relación nos indica que si aumenta la longitud del sépalo también aumentará longitud y ancho del pétalo.

La variable longitud del pétalo está muy correlacionada positivamente con la variable ancho de pétalo. Esta relación nos indica que si aumenta la longitud del pétalo también aumentará el ancho del pétalo.

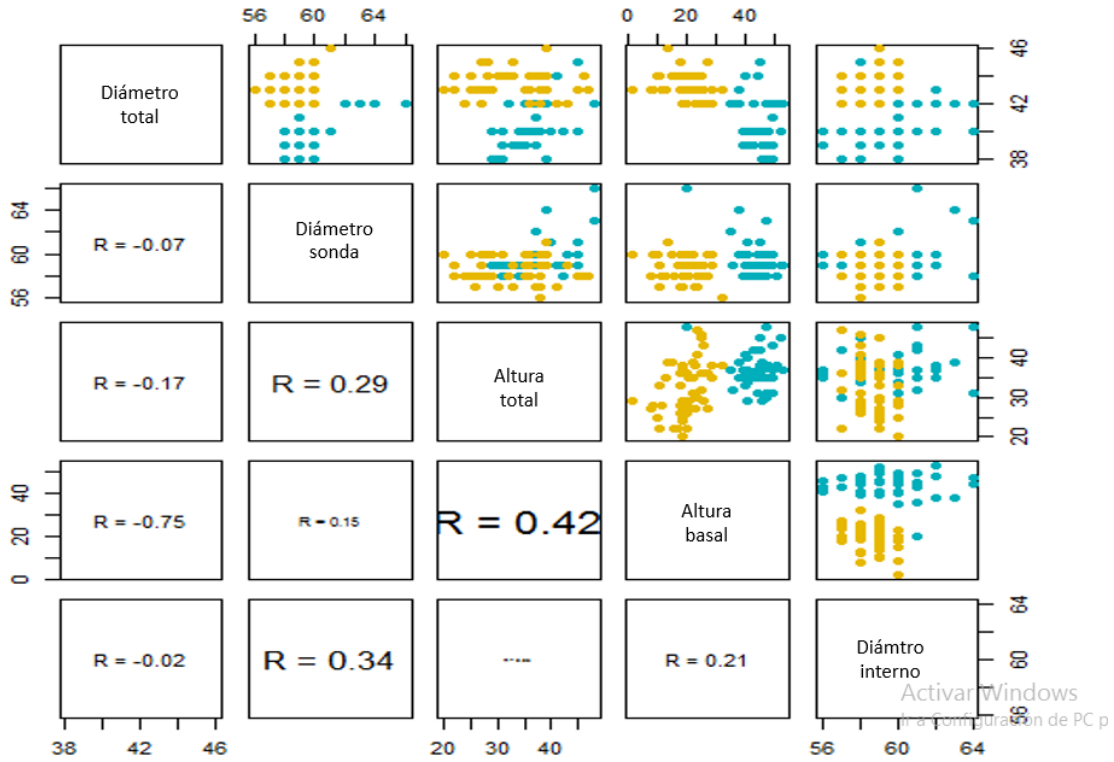


Figura 8: Gráficos de dispersión de las variables de la data Electrode

La variable diámetro total del electrodo está muy correlacionada negativamente con la variable altura basal del electrodo. Esta relación nos indica que si aumenta el diámetro total del electrodo disminuirá la altura basal del electrodo.

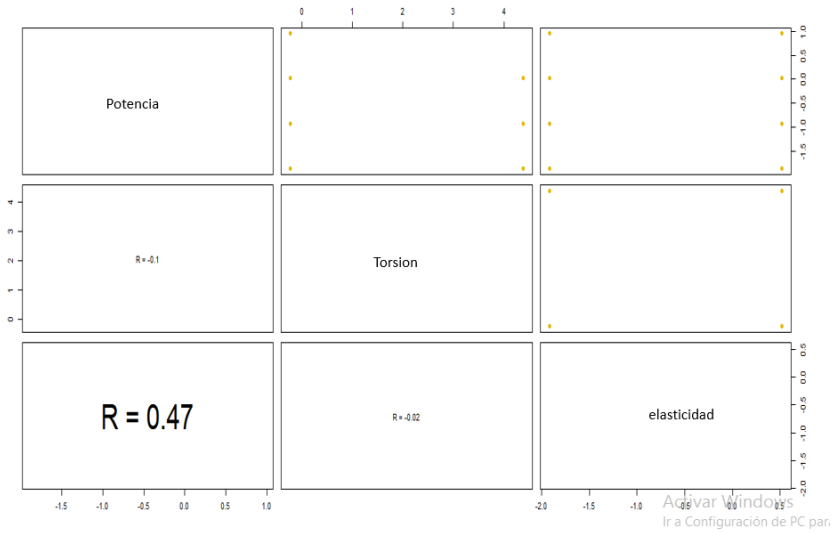


Figura 9: Gráficos de dispersión de las variables de la data Titanic

Las variables Potencia y Elasticidad tienen una correlación moderada, lo que implica que el aumento de la Potencia implica un aumento mesurado de la Elasticidad.

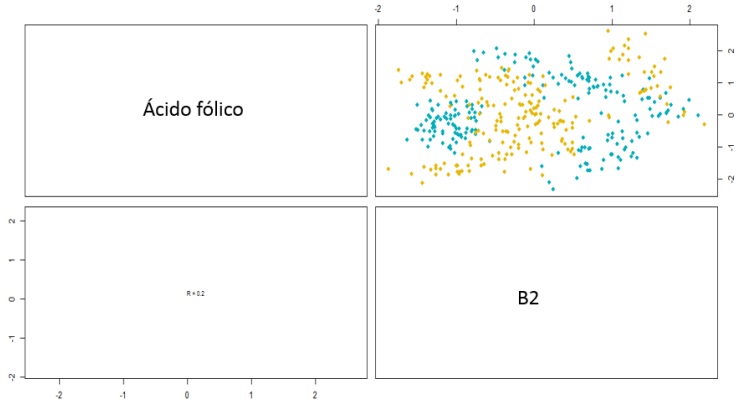


Figura 10: Gráficos de dispersión de las variables de la data Banana

La variable Ácido fólico y la vitamina B2 tienen una correlación positiva baja. Esta relación nos indica que si aumenta la concentración Ácido fólico no necesariamente aumentará la concentración de la vitamina B2.

En las figuras 3, 4, 6 y 7 se puede observar que a medida que aumenta el coeficiente de correlación las tasas de clasificación errónea disminuyen considerablemente.

Cuadro 3: Comparación de Clasificadores tomando el tiempo de procesamiento

CONJUTO DE DATOS	Tiempo de procesamiento(seg)		
	ADNM	**RL**	**CHAID**
Crabs	95.96	1.23	48.06
Ecoli	179.66	1.42	28.08
Glass identificacion	43.41	1.10	2.19
Bupa	464.41	1.31	29.55
Electrode	29.19	1.05	0.68
Iris	49.55	1.12	0.58
Titanic	310.44	2.17	12.49
Banana	235.73	1.20	15.94

FUENTE: Elaboración propia

En esta comparación de clasificadores el 75% de los casos se puede observar que la técnica de RLM tiene el menor tiempo de procesamiento seguida del CHAID y ADNM y en el 25% de los casos el CHAID tiene el menor tiempo de procesamiento seguida de RLM y ADNM. En los demás casos cuando se disminuye el número de variables se puede apreciar una disminución en la tasa de clasificación errónea a excepción de Glass identificación, Titanic y Banana en la cual el mejor clasificador sigue siendo la RLM seguido del ADNM.

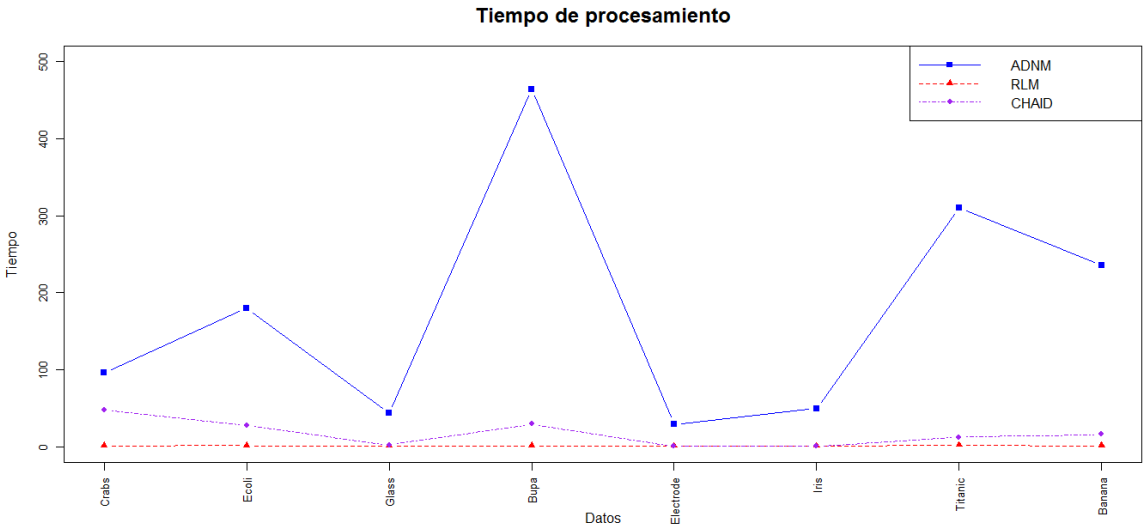


Figura 11: Gráfica de los métodos de clasificación tomando el tiempo de procesamiento

La técnica de RLM tuvo mejor desempeño teniendo en cuenta el tiempo de procesamiento el cual es menor seguida del Árbol de Clasificación CHAID. En todos los demás casos la técnica del ADNM obtuvo un tiempo de procesamiento mucho mayor a la RLM y al Árbol de clasificación CHAID.

V. CONCLUSIONES

Las conclusiones que se obtienen en la presente tesis son:

1. Las técnicas de clasificación que mejor desempeño tuvieron teniendo en cuenta la tasa de mala clasificación promedio fueron RLM y el ADNM. Sus desempeños fueron similares ya que en la gráfica se puede observar que las líneas asociadas se superponen, mostrando de esta manera que las diferencias son mínimas.

   En general, la regresión logística multinomial presenta un mejor desempeño que el análisis discriminante no métrico comparando numéricamente las tasas de clasificación errónea.
2. En la mayoría de los casos el Árbol de Clasificación CHAID obtuvo una tasa de mala clasificación promedio superiores a la RLM y al ADNM.

3. La técnica de clasificación que mejor desempeño tuvo en el estudio teniendo en cuenta el tiempo de procesamiento fue la técnica de RLM seguida del Árbol de Clasificación CHAID. En todos los casos la técnica del ADNM obtuvo un tiempo de procesamiento muy superior a la RLM y al Árbol de clasificación CHAID.

4. En situaciones prácticas donde se presente un problema de clasificación de nuevas observaciones a grupos ya definidos teniendo en cuenta varias variables independientes, se recomienda utilizar principalmente la técnica de RLM seguida del ADNM ya que los resultados obtenidos con ambas técnicas son similares.

5. En aquellos casos donde la tasa de clasificación errónea para los dos procedimientos es similar el análisis discriminante no métrico tiene ventaja con respecto a la regresión logística en el sentido de la interpretación de los resultados, ya que el procedimiento no métrico se interpreta en términos lineales y la regresión logística no.

VI. RECOMENDACIONES

1. En esta tesis se utilizaron siete conjuntos de datos, para un próximo trabajo se recomienda utilizar más conjuntos de datos que tengan sólo variables cuantitativas.

2. Se sugiere comparar el desempeño de las técnicas considerando otro tipo de escenarios en los cuales se puede estudiar aspectos como: mayor número de grupos a clasificar, tamaños muéstrales mayores, estructuras de matrices de varianza y covarianza iguales y diferentes, otros tipos de distribuciones para los grupos.

3. Se recomienda utilizar medidas de desempeño diferente a la tasa de mala clasificación promedio.

4. Se recomienda comparar el Análisis Discriminante no Métrico con otros métodos de clasificación como Redes Neuronales, Algoritmos Genéticos, etc.

VII. REFERENCIA BIBLIOGRÁFICA

[1] Anderson T. W. and Bahadur R. R. (1962). "Classification into two Multivariate Normal Distributions with Different Covariance Matrices". The Annals of Mathematical Statistics, Vol. 33, No. 2. pp. 420-431.

[2] Anderson, J.A. (1972). "Separate sample logistic discrimination". Biometrica. 59, 19-35.

[3] Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone. (1984) Classification and Regression Trees. New York: Chapman & Hall/CRC.

[4] Choulakian, V. and Almhana, J (2001). "An Algorithm for Nonmetric Discriminant Analysis". Computational Statistics & Data Analysis, 35, 253-264.

[5] Carroll, R. and Pederson, S. (1993). "On Robustness in the Logistic Regression Model". Journal of the Royal Statistical Society. Vol. 55, pp. 693-706.

[6] Castrillon, F. (1998). "Comparación de la discriminación normal lineal y Cuadrática con la regresión logística para clasificar vectores en dos poblaciones". Medellín. Tesis Magíster en Estadística. Facultad de Ciencias. Universidad Nacional de Colombia, Sede Medellín.

[7] Cheng, T., Pia. M. and Feser, V. (2002). "High-breakdown estimation of multivariate mean and covariance with missing observations". British Journal of Mathematical and Statistical Psychology. Vol. 55, 317-335.

[8] Cherkaoui, O. and Cleroux, R. (1991). "Comparative study of six classification methods for mixtures of variables" . Computing Science and Statisctics. 23, 233-236.

[9] Chernoff, H. (1972). “The selection of effective attributes for deciding between hypothesesusing linear discriminant functions”. Frontiers of pattern recognition, New York: Academic Press, pp. 55-60.
[10] Clunies, C. W. and Riffenburgh, R. H. (1960). “Geometry and linear discrimination”. Biometrics Vol. 47, No. 1/2. pp. 185-189.

[11] Cornfiel, J. (1962). “Joint dependence of risk of coronary heart disease on serum cholesterol and systolic blood pressure: a discriminant function analysis.”. Proceedings of the Federal American Society of Experimental Biology. Vol 21, 58-61.

[12] Cox, D.R. (1996). “Some procedures associated with the logistic qualitative response curve”. Research papers in statistics: Festschrift for J. Newman., New York Wiley.

[13] Crawley, D. R. (1979). “Logistic discrimination as an alternative to Fisher’s linear function”. New Zealand Statistician. 14, 21-25.

[14] Croux, C. and Dehon, C. (2001). “Robust Linear Discriminant Analysis Using S-Estimators”. The Canadian Journal of Statistics / La Revue Canadienne de Statistique. Vol. 29, No. 3. pp. 473-493.

[15] Díaz, L. (2002). “Estadística multivariada: inferencia y métodos”. Universidad Nacional de Colombia.

[16] Day, N.E. and Kerridge, D.F. (1967). “A general maximum likelihood discriminant”. Biometrics. 23, 313-323.

[17] Efron, B. (1975). “The Efficiency of Logistic Regression compared to Normal Discriminant Analysis”. Journalof the American Statistical Association, 70, 892-898.

[18] Estrada, J., Camacho, J., Restrepo, M. y Parra, C. (1998). “Parámetros antropométricos de la población laboral colombiana 1995 (acopla95)”. Revista de la Facultad Nacional de Salud Pública, 15(2): 112-139.

[19] Fisher R. A. (1936). “The Use of Multiple Measurements in Taxonomic Problems”. Annual Eugenics 7: 179-188.

[20] Guttman, L. (1998). “Eta, disco, odisco and F.”. Psychometrika. 53, 393-405.

[21] Hand, D. (1989), Discriminant Analysis for Psychiatric Screening, 2 edn, John Wiley & Sons, New York, United States.

[22] Hawkins, D.M. and McLachan J. (1997). “High-Breakdown linear discriminant analysis”. Journal of American Statistical Asociation. 92, 136-146.

[23] Kass, G.V. (1980) An Exploratory Technique for Investigating Large Quantities of Categorical Data. Journal of Applied Statistics 29(2): 119-127.

[24] Johnson, R. y Wichern, D. (1998). “Applied Multivariate Statistical Analysis”. London: Prentice-Hall.

[25] Little, R. J. and Smith, P. J. (1987). “Editing and imputing for quantitative survey data”. Journal of the American Statistical Association. 82, 58-68.

[26] Mayorga. J., (2004). “Inferencia estadística”. 1edn,Universidad Nacional de Colombia.114_115

[27] Morrison, D. (1990), Multivariate Statistical Methods, 3 edn, McGraw-Hill, New York, United States.

[28] Pohar, M., Blas, M. & Turk, S. (2004), ‘Comparison of Logistic Regression and Linear Discriminant Analysis: A Simulation Study’, Metodolski Zvezki 1, 143–161.

[29] Pregibon, D. (1981). “Logistic Regression Diagnostics”. The Annals of Statistics. Vol. 9, pp. 705-724.

[30] Rao, C.R. (1948). “The utilization of multiple meausurements in problems of biological classification”. J.Royal Statistic., Vol. 10, 159-193.

[31] Raveh, A. (1983). “Preference structure analysis: A nonmetric approach” .Patter Recognition 16, 253-259.

[32] Raveh, A. (1989). “A Nonmetric Approach to Linear Discriminant Analysis” . Journal of the American Statistical Association. 84, 176-183.

[33] Shelley, B. and Donner, A. (1987). “The Efficiency of Multinomial Logistic Regression compared with Multiple Group Discriminant Analysis”. Journal of American Statistical Association. 82, 1118-1122.

VIII. ANEXOS

ANEXO 1: Programa del Análisis Discrimante No Métrico basado en el algoritmo propuesto por Choulakian y Almhana.

```
#Función para evitar CEROS en caso de ##########################igualdad de observaciones

ajuste<-function(x){   # x es la matriz de diseño con las categorias en# la primer columna

a<-x[,-1]              # a es la matriz de observaciones solamente

n<-nrow(a)             # compara la observacion i y la j, si existe igualdad

for (i in 1:(n-1)){    # en alguna var aplica jitter a la var de la obs de i

for (j in (i+1):n) {

a[i,]=ifelse(a[i,]-a[j,]==0,apply(as.matrix(a[i,]),2,jitter),a[i,])

}

}

cbind(x[,1],a)         # une las categorias con las observaciones modificad

}

########Calculando las matrices B y V ############################

BV<-function(x){  # x es la matriz de datos, en la 1º colum aparecen los grup

G<-nlevels(x[,1]) # Calculando el numero de grupos

p<-ncol(x)-1      # Calculando el numero de variables

Medias<-matrix(0,ncol=p,nrow=G)   # Se crea una matriz para las medias por grupo

Grupos<-as.matrix(levels(x[,1]))  # Esta matriz tiene los nombres de los grupos

for (i in 1:G){

Medias[i,]<- apply(((as.data.frame((split(x=x,f=x[,1]))[i]))[,2:(p+1)]),

2,mean)}       # Esta matriz tiene las medias de los grupos por filas

B<-matrix(0,ncol=p*G,nrow=p*G)

for (i in 1:G){

for (j in 1:G){    # Aqui se calculan todas la matrices Bgh
```

```
f1<-p*i-(p-1)
f2<-p*i
c1<-p*j-(p-1)
c2<-p*j
B[f1:f2,c1:c2]<-(Medias[i,]-Medias[j,])%*%t(Medias[i,]-Medias[j,])
}
}
ndatos<-nrow(x)
x<-ajuste(x)
obs<-as.matrix(x[,2:(p+1)])
V<-matrix(0,ncol=p*ndatos,nrow=p*ndatos)
for (i in 1:ndatos){
for (j in 1:ndatos){    # Aqui se calculan todas la matrices Vgh
f1<-p*i-(p-1)
f2<-p*i
c1<-p*j-(p-1)
c2<-p*j
V[f1:f2,c1:c2]<-(obs[i,]-obs[j,])%*%t(obs[i,]-obs[j,])
}
}
list(B=B,V=V)
}


###Calculando las matrices B(neta) y V(neta) ###################
Bneta<-function(B,neta,nelem){
p<-nrow(neta)
G<-nrow(B)/p
Bneta<-matrix(0,ncol=p,nrow=p)
for (i in 1:(G-1)){
for (j in (i+1):G){
```

```
f1<-p*i-(p-1)

f2<-p*i

c1<-p*j-(p-1)

c2<-p*j

Bneta<-(nelem[i,]*nelem[j,]*B[f1:f2,c1:c2])/as.real(sqrt(t(neta)%*%B[f1:f2,c1:c2]%*%neta)) + Bneta

}

}

Bneta<-Bneta*2

Bneta

}

Vneta<-function(V,neta){

p<-nrow(neta)

ndatos<-nrow(V)/p

Vneta<-matrix(0,ncol=p,nrow=p)

for (i in 1:(ndatos-1)){

for (j in (i+1):ndatos){

f1<-p*i-(p-1)

f2<-p*i

c1<-p*j-(p-1)

c2<-p*j

cuadrado<- as.real(t(neta)%*% V[f1:f2,c1:c2]%*% neta)

Vneta<- ( V[f1:f2,c1:c2]/ sqrt(cuadrado) ) + Vneta

}

}

Vneta<-Vneta*2

Vneta

}

#####Funcion DISCOneta #########

DISCOneta<-function(neta,Bneta,Vneta){

disconeta<-as.real((t(neta)%*%Bneta%*%neta))/as.real((t(neta)%*%Vneta%*%neta))
```

```
disconeta
}
## Funcion de paso##
netak<-function(neta,V,B){
netak<-neta*(1-2*as.real(DISCOneta(neta=neta,Bneta=B,Vneta=V)))+2*solve(V,tol=10^-200)%*%B%*%neta
netak
}
############### Librerias necesarias#
require(fpc)
####    Funcion      NDA   ######
NDA<-function(w,epsilon,n.iteraciones=200){
dat<-ajuste(w)          # w es la matriz de datos, en la 1º col estan las categ
p<-ncol(dat)-1          # epsilon es la diferencia
                    # n.iteraciones es el maximo de iteraciones se cree se necesita
require(fpc)            # n.elem.grup es un arreglo con los numeros de obs x grup
nele<- as.matrix(table(w[1]))   # Esta debe ser una matriz de 1 col por g filas
neta0<-as.matrix(discrcoord(xd=dat[,2:(p+1)], # Aqui se halla el discriminante
clvecd=dat[,1], pool = "n")$units[,1])       # canónico
neta0<-neta0/max(abs(neta0))                # Aqui se escala
B<-BV(dat)$B                            # Las matrices B y V
V<-BV(dat)$V
nitera<-n.iteraciones
B0<-Bneta(B=B,neta=neta0,nelem=nele)
V0<-Vneta(V=V,neta=neta0)
DISCO0<-DISCOneta(neta=neta0,Bneta=B0,Vneta=V0)
Bn<-Vn<-matrix(0,ncol=p,nrow=p)
Bn<-B0
Vn<-V0
iteraciones<-matrix(0,ncol=(p+1),nrow=nitera)
iteraciones[1,1]<-DISCO0
```

```
iteraciones[1,2:(p+1)]<-t(neta0)

convergencia<-matrix(0,ncol=1,nrow=nitera)

convergencia[1,1]<-1

i<-2

while(convergencia[i-1,1]>0)  {

iteraciones[i,2:(p+1)]<-as.matrix(netak(neta=as.matrix(iteraciones[i-1,2:(p+1)]),V=Vn,B=Bn))

Bn<-Bneta(B=B,neta=as.matrix(iteraciones[i,2:(p+1)]),nelem=nele)

Vn<-Vneta(V=V,neta=as.matrix(iteraciones[i,2:(p+1)]))

iteraciones[i,1]<-DISCOneta(neta=as.matrix(iteraciones[i,2:(p+1)]),Bneta=Bn,Vneta=Vn)

convergencia[i,1]<-ifelse(abs(iteraciones[i,1]-iteraciones[(i-1),1])>epsilon,1,0)

i<-i+1

}

nconver<-sum(convergencia,na.rm =T)

par(bg='lightyellow',cex.axis=0.7)

plot(x=iteraciones[1:nconver,1],ylab='Disco',xlab='Iteración',

main='Comportamiento de Disco',pch=20,ylim=c(0,1.01),col='blue')

abline(a=1,b=0,col='red')

axis(1, 1:nconver)

list(NETA=iteraciones[nconver,-1],NETACERO=t(neta0),DISCO=iteraciones[nconver,1],Niteraciones=nconver)

}

# El conjunto de datos A CLASIFICAR se denota por "datos.predic"

# El conjunto de datos de entreno de denota por "datos"

require(klaR)

NDA.CLA<-function(datos,datos.predic,resultado){

vector<-resultado$NETA   # Vector o Función Discriminante

scores.clas<-as.matrix(datos[,-1])%*%as.matrix(vector)   # Scores para clasificar

n.grupos<-nlevels(datos[,1])

G<-datos[,1]

S<-scores.clas

x<-data.frame(G=factor(datos[,1]),S=scores.clas)
```

```
y.clas<-x[order(S,G),]     # Y es la matriz con los grupos y los scores de clasificac

plot(x,ylab='Scores',main='Distribución de los Scores por grupo',xlab='Grupos')

                                # Distribución de los scores por grupo

prediccion<-datos.predic[,1]

scores.predic<-as.matrix(datos.predic[,-1])%*%as.matrix(vector)     # Scores para las nuevas observaciones
cortes<-lapply(split(y.clas[,2],y.clas[,1]),mean,3)                                                    distancias<-
matrix(0,ncol=n.grupos,nrow=nrow(datos.predic))

for (i in 1:nrow(datos.predic)){ distancias[i,]= abs (unlist(t(as.matrix(cortes))) - scores.predic[i,1]) }

distancias

predicciones<-names(cortes[max.col(-distancias)])

tabla<-errormatrix(true=datos.predic[,1], predicted=predicciones, relative = FALSE) # Tabla de clasificacion

n.clas.erradas<-errormatrix(true=datos.predic[,1], predicted=predicciones,

relative = FALSE)[n.grupos+1,n.grupos+1]  # Nº de clasificaciones que fueron clasificadas incorrectamente

tasa.clas.incorr<-n.clas.erradas/nrow(datos.predic)

list(TABLA=tabla,N.clas.erradas=n.clas.erradas,tasa.clas.incorr=tasa.clas.incorr)

}
```

#Trabajando con la data Iris

```
crossval(iris,10,10,5)

$EVCP

[1] 0.0033284

system.time(crossval(iris,10,10,5))[3]

elapsed

  49.55
```

#Trabajando con la data Crabs

```
#Package(MASS)

data(crabs)

crabs2=crabs[,-2]

crabs3=crabs2[,-2]

crossval(crabs3,10,10,1)

$EVCP
```

[1] 0.0006638889

system.time(crossval(crabs3,10,10,1))[3]

elapsed

95.96

#Trabajando con la data Ecoli

ecoli3=read.table("ecoli3.txt",header=F)

crossval(ecoli3,10,10,7)

$EVCP

[1] 0.00911327

system.time(crossval(ecoli3,10,10,7))[3]

elapsed

179.66

#Trabajando con la data Electrode

#Package(flury)

data(electrode)

crossval(electrode,10,10,1)

$EVCP

[1] 0.00526667

system.time(crossval(electrode,10,10,1))[3]

elapsed

29.19

#Trabajando con la data Glass identificación

glass=read.table("glaas2.txt",header=F)

crossval(glass,10,10,5)

$EVCP

[1] 0.02503756

system.time(crossval(glass,10,10,5))[3]

elapsed

43.41

#Trabajando con la data Bupa

```
bupa=read.table("bupa.txt",header=F)

crossval(bupa,10,10,7)

$EVCP

[1] 0.01123975

system.time(crossval(bupa,10,10,7))[3]

elapsed

464.41
```

#Trabajando con la data Titanic

```
tita=read.table("tita.txt",header=F)

crossval(tita,10,10,4)

$EVCP

[1] 0.0105679

system.time(crossval(tita,10,10,4))[3]

elapsed

310.44
```

#Trabajando con la data Banana

```
bana=read.table("bana.txt",header=F)

crossval(bana,10,10,3)

$EVCP

[1] 0.01188125

system.time(crossval(bana,10,10,3))[3]

elapsed

235.73
```

#Regresión Logística Multinomial

```
crossval=function(data,repet,K,y)
{
library(nnet)
  n=dim(as.matrix(data))[1]
  p=dim(as.matrix(data))[2]
  EVC=rep(0, repet)
  for (i in 1:repet)
```

```
  {
   resid=matrix(0,1,K)
   indices=sample(1:n,n,replace=F)
   azar=data[indices,]
   subm=floor(n/K)
   for (j in 1:K)
   {
    unid=((j-1)*subm+1):(j*subm)
    if (j == K)
    {
     unid=((j-1)*subm+1):n
    }
    datap=azar[unid,]
    datae=azar[-unid,]
    yp=datap[,y]
    xp=datap[,-y]
    dataps=cbind(yp,xp)
    ye=datae[,y]
    xe=datae[,-y]
    dataes=cbind(ye,xe)
result=multinom(yp~.,dataps)
Ypred=predict(result,newdata=dataes)
t = table(ye, Ypred)
a=1-sum(diag(t))/sum(t)
resid[j]=a
}
EVC[i]=sum(resid)/n
  }
EVCP=mean(EVC)
return (list(EVC=EVC, EVCP=EVCP))
}
```

#Trabajando con la data Iris

```
crossval(iris,10,10,5)

$EVCP

[1] 0.00577284

system.time(crossval(iris,10,10,5))[3]

elapsed

1.12
```

#Trabajando con la data Crabs

```
#Package(MASS)

data(crabs)

crabs2=crabs[,-2]
```

```
crabs3=crabs2[,-2]

crossval(crabs3,10,10,1)

system.time(crossval(crabs3,10,10,1))[3]

$EVCP

[1] 0.0004333333

system.time(crossval(can,10,10,10))[3]

elapsed

1.23
```

#Trabajando con la data Ecoli

```
ecoli3=read.table("ecoli3.txt",header=F)

crossval(ecoli3,10,10,7)

$EVCP

[1] 0.00656569

system.time(crossval(ecoli3,10,10,7))[3]

elapsed

1.42
```

#Trabajando con la data Electrode

```
#Package(flury)

data(electrode)

crossval(electrode,10,10,1)

$EVCP

[1] 0.00451111

system.time(crossval(electrode,10,10,1))[3]

elapsed

1.05
```

#Trabajando con la data Glass identificación

```
glass=read.table("glaas2.txt",header=F)

crossval(glass,10,10,5)

$EVCP

[1] 0.02503608
```

system.time(crossval(glass,10,10,5))[3]

elapsed

1.1

#Trabajando con la data Bupa

bupa=read.table("bupa.txt",header=F)

crossval(bupa,10,10,7)

$EVCP

[1] 0.01069364

system.time(crossval(bupa,10,10,7))[3]

elapsed

1.31

#Trabajando con la data Titanic

tita=read.table("tita.txt",header=F)

crossval(tita,10,10,4)

$EVCP

[1] 0.01039877

system.time(crossval(tita,10,10,4))[3]

elapsed

2.17

#Trabajando con la data Banana

bana=read.table("bana.txt",header=F)

crossval(bana,10,10,3)

$EVCP

[1] 0.01243611

system.time(crossval(bana,10,10,3))[3]

elapsed

1.2

#Trabajando con la data Iris

#Arbol de clasificación CHAID

```
crossval=function(data,repet,K,y)
 {
library(CHAID)
v1=ordered(cut(data[,1],breaks = 10))
v2=ordered(cut(data[,2],, breaks = 10))
v3=ordered(cut(data[,3],,breaks = 10))
v4=ordered(cut(data[,4], breaks = 10))
data=data.frame(v1,v2,v3,v4,v5=data[,5])
n=dim(as.matrix(data))[1]
p=dim(as.matrix(data))[2]
EVC=rep(0, repet)
for (i in 1:repet)
  {
   resid=matrix(0,1,K)
   indices=sample(1:n,n,replace=F)
   azar=data[indices,]
   subm=floor(n/K)
   for (j in 1:K)
   {
    unid=((j-1)*subm+1):(j*subm)
    if (j == K)
    {
     unid=((j-1)*subm+1):n
   }
    datap=azar[unid,]
    datae=azar[-unid,]
    yp=datap[,y]
    xp=datap[,-y]
    dataps=cbind(yp,xp)
    ye=datae[,y]
```

```
    xe=datae[,-y]
    dataes=cbind(ye,xe)
result=chaid(yp~.,dataps)
Ypred=predict(result,newdata=dataes)
t = table(ye, Ypred)
a=1-sum(diag(t))/sum(t)
    resid[j]=a
   }
   EVC[i]=sum(resid)/n
  }
  EVCP=mean(EVC)
 return (list(EVC=EVC, EVCP=EVCP))
 }
 crossval(iris,10,10,5)
$EVCP
[1] 0.0453284
system.time(crossval(iris,10,10,5))[3]
elapsed
0.58.
```

#Trabajando con la data Crabs
#Arbol de clasificación CHAID

```
crossval=function(data,repet,K,y)
 {
library(CHAID)
v2=ordered(cut(data[,2],, breaks = 10))
v3=ordered(cut(data[,3],,breaks = 10))
v4=ordered(cut(data[,4], breaks = 10))
v5=ordered(cut(data[,5], breaks = 10))
v6=ordered(cut(data[,6],breaks = 10))
data=data.frame(v2,v3,v4,v5,v6,v1=data[,1])
```

```
n=dim(as.matrix(data))[1]
p=dim(as.matrix(data))[2]
EVC=rep(0, repet)
for (i in 1:repet)
 {
  resid=matrix(0,1,K)
  indices=sample(1:n,n,replace=F)
  azar=data[indices,]
  subm=floor(n/K)
  for (j in 1:K)
  {
   unid=((j-1)*subm+1):(j*subm)
   if (j == K)
   {
    unid=((j-1)*subm+1):n
  }
   datap=azar[unid,]
   datae=azar[-unid,]
   yp=datap[,y]
   xp=datap[,-y]
   dataps=cbind(yp,xp)
   ye=datae[,y]
   xe=datae[,-y]
   dataes=cbind(ye,xe)
result=chaid(yp~.,dataps)
Ypred=predict(result,newdata=dataes)
t = table(ye, Ypred)
a=1-sum(diag(t))/sum(t)
   resid[j]=a
  }
```

```
  EVC[i]=sum(resid)/n
 }
 EVCP=mean(EVC)
 return (list(EVC=EVC, EVCP=EVCP))
 }
#Package(MASS)
data(crabs)
crabs2=crabs[,-2]
crabs3=crabs2[,-2]
crossval(crabs3,10,10,1)
$EVCP
[1] 0.03535278
system.time(crossval(crabs3,10,10,1))[3]
elapsed
48.06
```

#Trabajando con la data Ecoli
#Arbol de clasificación CHAID

```
crossval=function(data,repet,K,y)
 {
library(CHAID)
v1=ordered(cut(data[,1],breaks = 10))
v2=ordered(cut(data[,2],, breaks = 10))
v3=ordered(cut(data[,3],,breaks = 10))
v4=ordered(cut(data[,4], breaks = 10))
v5=ordered(cut(data[,5],breaks = 10))
v6=ordered(cut(data[,6],, breaks = 10))
data=data.frame(v1,v2,v3,v4,v5,v6,v7=data[,7])
n=dim(as.matrix(data))[1]
p=dim(as.matrix(data))[2]
EVC=rep(0, repet)
```

```
for (i in 1:repet)
 {
  resid=matrix(0,1,K)
  indices=sample(1:n,n,replace=F)
  azar=data[indices,]
  subm=floor(n/K)
  for (j in 1:K)
  {
   unid=((j-1)*subm+1):(j*subm)
   if (j == K)
   {
    unid=((j-1)*subm+1):n
  }
   datap=azar[unid,]
   datae=azar[-unid,]
   yp=datap[,y]
   xp=datap[,-y]
   dataps=cbind(yp,xp)
   ye=datae[,y]
   xe=datae[,-y]
   dataes=cbind(ye,xe)
result=chaid(yp~.,dataps)
Ypred=predict(result,newdata=dataes)
t = table(ye, Ypred)
a=1-sum(diag(t))/sum(t)
    resid[j]=a
  }
  EVC[i]=sum(resid)/n
 }
 EVCP=mean(EVC)
```

```
 return (list(EVC=EVC, EVCP=EVCP))
 }
ecoli3=read.table("ecoli3.txt",header=F)
crossval(ecoli3,10,10,7)
$EVCP
[1] 0.0075371
system.time(crossval(ecoli3,10,10,7))[3]
elapsed
28.08
```

#Trabajando con la data Electrode
#Arbol de clasificación CHAID

```
crossval=function(data,repet,K,y)
 {
library(CHAID)
v2=ordered(cut(data[,2],, breaks = 10))
v3=ordered(cut(data[,3],,breaks = 10))
v4=ordered(cut(data[,4], breaks = 10))
v5=ordered(cut(data[,5],breaks = 10))
data=data.frame(v2,v3,v4,v5,v1=data[,1])
n=dim(as.matrix(data))[1]
p=dim(as.matrix(data))[2]
EVC=rep(0, repet)
for (i in 1:repet)
 {
  resid=matrix(0,1,K)
  indices=sample(1:n,n,replace=F)
  azar=data[indices,]
  subm=floor(n/K)
  for (j in 1:K)
  {
```

```
   unid=((j-1)*subm+1):(j*subm)
   if (j == K)
   {
    unid=((j-1)*subm+1):n
  }
   datap=azar[unid,]
   datae=azar[-unid,]
   yp=datap[,y]
   xp=datap[,-y]
   dataps=cbind(yp,xp)
   ye=datae[,y]
   xe=datae[,-y]
   dataes=cbind(ye,xe)
result=chaid(yp~.,dataps)
Ypred=predict(result,newdata=dataes)
t = table(ye, Ypred)
a=1-sum(diag(t))/sum(t)
    resid[j]=a
   }
  EVC[i]=sum(resid)/n
 }
 EVCP=mean(EVC)
 return (list(EVC=EVC, EVCP=EVCP))
 }
#Package(flury)
data(electrode)
crossval(electrode,10,10,1)
$EVCP
[1] 0.08082222
system.time(crossval(electrode,10,10,1))[3]
```

```
elapsed

0.68
```

#Trabajando con la data Glass identificación
#Arbol de clasificación CHAID

```
crossval=function(data,repet,K,y)

 {

library(CHAID)

v1=ordered(cut(data[,1],breaks = 10))

v2=ordered(cut(data[,2],, breaks = 10))

v3=ordered(cut(data[,3],,breaks = 10))

v4=ordered(cut(data[,4], breaks = 10))

data=data.frame(v1,v2,v3,v4,v5=data[,5])

n=dim(as.matrix(data))[1]

p=dim(as.matrix(data))[2]

EVC=rep(0, repet)

for (i in 1:repet)

 {

  resid=matrix(0,1,K)

  indices=sample(1:n,n,replace=F)

  azar=data[indices,]

  subm=floor(n/K)

  for (j in 1:K)

  {

   unid=((j-1)*subm+1):(j*subm)

   if (j == K)

   {

    unid=((j-1)*subm+1):n

  }

   datap=azar[unid,]

   datae=azar[-unid,]
```

```
    yp=datap[,y]
    xp=datap[,-y]
    dataps=cbind(yp,xp)
    ye=datae[,y]
    xe=datae[,-y]
    dataes=cbind(ye,xe)
result=chaid(yp~.,dataps)
Ypred=predict(result,newdata=dataes)
t = table(ye, Ypred)
a=1-sum(diag(t))/sum(t)
    resid[j]=a
   }
   EVC[i]=sum(resid)/n
  }
  EVCP=mean(EVC)
 return (list(EVC=EVC, EVCP=EVCP))
 }
glass=read.table("glaas2.txt",header=F)
crossval(glass,10,10,5)
$EVCP
[1] 0.03475286
system.time(crossval(glass,10,10,5))[3]
elapsed
2.19
```

#Trabajando con la data Bupa
#Arbol de clasificación CHAID

```
crossval=function(data,repet,K,y)
 {
library(CHAID)
v1=ordered(cut(data[,1],breaks = 10))
```

```
v2=ordered(cut(data[,2],, breaks = 10))

v3=ordered(cut(data[,3],,breaks = 10))

v4=ordered(cut(data[,4], breaks = 10))

v5=ordered(cut(data[,5],, breaks = 10))

v6=ordered(cut(data[,6],,breaks = 10))

data=data.frame(v1,v2,v3,v4,v5,v6,v7=data[,7])

n=dim(as.matrix(data))[1]

p=dim(as.matrix(data))[2]

EVC=rep(0, repet)

for (i in 1:repet)

 {

  resid=matrix(0,1,K)

  indices=sample(1:n,n,replace=F)

  azar=data[indices,]

  subm=floor(n/K)

  for (j in 1:K)

  {

   unid=((j-1)*subm+1):(j*subm)

   if (j == K)

   {

    unid=((j-1)*subm+1):n

  }

   datap=azar[unid,]

   datae=azar[-unid,]

   yp=datap[,y]

   xp=datap[,-y]

   dataps=cbind(yp,xp)

   ye=datae[,y]

   xe=datae[,-y]

   dataes=cbind(ye,xe)
```

```
result=chaid(yp~.,dataps)

Ypred=predict(result,newdata=dataes)

t = table(ye, Ypred)

a=1-sum(diag(t))/sum(t)

    resid[j]=a

   }

   EVC[i]=sum(resid)/n

  }

  EVCP=mean(EVC)

 return (list(EVC=EVC, EVCP=EVCP))

 }

bupa=read.table("bupa.txt",header=F)

crossval(bupa,10,10,7)

$EVCP

[1] 0.01317616

system.time(crossval(bupa,10,10,7))[3]

elapsed

29.55
```

#Trabajando con la data Titanic
#Arbol de clasificación CHAID

```
crossval=function(data,repet,K,y)

 {

library(CHAID)

v1=ordered(cut(data[,1],, breaks = 10))

v2=ordered(cut(data[,2],, breaks = 10))

v3=ordered(cut(data[,3],,breaks = 10))

data=data.frame(v1,v2,v3,v4=data[,4])

n=dim(as.matrix(data))[1]

p=dim(as.matrix(data))[2]

EVC=rep(0, repet)
```

```
for (i in 1:repet)
 {
  resid=matrix(0,1,K)
  indices=sample(1:n,n,replace=F)
  azar=data[indices,]
  subm=floor(n/K)
  for (j in 1:K)
  {
   unid=((j-1)*subm+1):(j*subm)
   if (j == K)
   {
    unid=((j-1)*subm+1):n
  }
   datap=azar[unid,]
   datae=azar[-unid,]
   yp=datap[,y]
   xp=datap[,-y]
   dataps=cbind(yp,xp)
   ye=datae[,y]
   xe=datae[,-y]
   dataes=cbind(ye,xe)
result=chaid(yp~.,dataps)
Ypred=predict(result,newdata=dataes)
t = table(ye, Ypred)
a=1-sum(diag(t))/sum(t)
    resid[j]=a
  }
  EVC[i]=sum(resid)/n
 }
 EVCP=mean(EVC)
```

```
 return (list(EVC=EVC, EVCP=EVCP))
 }
tita=read.table("tita.txt",header=F)
crossval(tita,10,10,4)
$EVCP
[1] 0.01159136
system.time(crossval(tita,10,10,4))[3]
elapsed
  12.49

#Trabajando con la data Banana
#Arbol de clasificación CHAID
crossval=function(data,repet,K,y)
 {
library(CHAID)
v1=ordered(cut(data[,1],breaks = 10))
v2=ordered(cut(data[,2],, breaks = 10))
data=data.frame(v1,v2,v3=data[,3])
n=dim(as.matrix(data))[1]
p=dim(as.matrix(data))[2]
EVC=rep(0, repet)
for (i in 1:repet)
  {
   resid=matrix(0,1,K)
   indices=sample(1:n,n,replace=F)
   azar=data[indices,]
   subm=floor(n/K)
   for (j in 1:K)
   {
    unid=((j-1)*subm+1):(j*subm)
```

```
    if (j == K)
    {
     unid=((j-1)*subm+1):n
   }
    datap=azar[unid,]
    datae=azar[-unid,]
    yp=datap[,y]
    xp=datap[,-y]
    dataps=cbind(yp,xp)
    ye=datae[,y]
    xe=datae[,-y]
    dataes=cbind(ye,xe)
result=chaid(yp~.,dataps)
Ypred=predict(result,newdata=dataes)
t = table(ye, Ypred)
a=1-sum(diag(t))/sum(t)
     resid[j]=a
    }
   EVC[i]=sum(resid)/n
  }
  EVCP=mean(EVC)
 return (list(EVC=EVC, EVCP=EVCP))
 }
bana=read.table("bana.txt",header=F)
crossval(bana,10,10,3)
$EVCP
[1] 0.01171597
system.time(crossval(bana,10,10,3))[3]
elapsed
  15.94
```

ANEXO 2: Funciones NDA y NDA.CLA en R para Análisis Discriminante no Métrico.

1. Función NDA

- Descripción: esta función permite encontrar el vector de clasificación para el método de Análisis Discriminante No Métrico usando el algoritmo sugerido por Choulakian y Almhama (2001).
- Sintaxis de la función:

  NDA(w= , epsilon=10^-7 , n.iteraciones=)
- Argumentos

  w: marco de datos donde la primera columna contiene los nombres de los grupos a los cuales pertenece cada observación. Es importante que los nombres estén ordenados.

  epsilon: valor sugerido por Choulakian y Almhama (2001) para detener el proceso iterativo.

  n.iteraciones: número máximo de iteraciones se cree va a necesitar el algoritmo para la convergencia. Si el número total de iteraciones que tomó el algoritmo para converger es igual al que el usuario dió (n.iteraciones) se recomienda volver a correr la función y aumentar este parámetro; sin embargo, la ventaja de este algoritmo es que converge en no más de treinta iteraciones por lo general.
- Valores

  NETA: corresponde al vector que sirve como función de clasificación.

  NETACERO: corresponde al vector de clasificación de Análisis Discriminante Canónico que sirve como semilla para el algoritmo.

  DISCO: valor máximo del discriminante cuando converge el algoritmo.

  Niteraciones: número de iteraciones que tomó el algoritmo para alcanzar la convergencia.

  Adicionalmente se obtiene un gráfico donde se puede ver el comportamiento del discriminante, en el eje horizontal van las iteraciones y en el vertical el valor de DISCO.
- **Ejemplo**

  En este ejemplo se tomará la base de datos Iris, se obtendrá la función de clasificación con NDA.

  ```
  # Colocando en la primera columna los nombres de los grupos de cada observación
  datos<-cbind(iris[,5],iris[,-5])
  ```

NDA(w=datos,epsilon=10^-7,n.iteraciones=200)

2. NDA.CLA

- **Descripción**: esta función tiene como objetivo clasificar un conjunto de observaciones

  en uno de varios grupos utilizando el resultado obtenido por la función NDA.

- **Sintaxis de la función**:

  NDA.CLA(datos= , datos.predic= , resultado=)

- **Argumentos**

  datos: marco de datos que se sirvieron para encontrar el vector de clasificación con la función NDA.

  datos.predict: marco de datos que se quiere clasificar.

  resultado: objeto que correponde a la salida de la función NDA.

- **Valores**

  TABLA: aparece una tabla donde se muestran las clasificaciones correctas y erradas.

  N.clas.erradas: número de clasificaciones erradas.

  tasa.clas.incorr: tasa de clasificaciones incorrectas.

- Ejemplo

  En este ejemplo se tomará la base de datos Iris, se obtendrá la función de clasificación con NDA y luego se validará la clasificación con el mismo conjunto.

  ```
  datos<-cbind(iris[,5], iris[,-5])
  resul.NDA<-NDA(w=datos,epsilon=10^-7,n.iteraciones=200)
  datos.predic<-datos
  z.NDA<-NDA.CLA(datos=datos,datos.predic=datos.predic,resultado=resul.NDA)
  z.NDA
  ```

3. Referencias

- Choulakian, V. and Almhana, J (2001). “An Algorithm for Nonmetric Discriminant Analysis”. Computational Statistics & Data Analysis, 35, 253-264.

ANEXO 3: Código en R para el gráfico de líneas para la Validación Cruzada

```
y1<-c(0.0006639,0.00911,0.02503756,0.01123975,0.00526667,0.0033284,0.0105679, 0.01188125)
y2<-c(0.000433333,0.00656569,0.02503608,0.01069364,0.00451111,0.00577284,0.01039877,0.0124361)
y3<-c(0.03535278,0.0075371,0.03475286,0.01317616,0.08082222,0.0453284,0.01159136,0.0117159)
x<-1:8
nombre<-c("Crabs","Ecoli","Glass","Bupa","Electrode","Iris","Titanic","Banana")
plot(x,y1,xaxt="n",cex.axis=0.8,pch=1,bg="gray",col="blue",cex=0.1,main="Validación
Cruzada",xlab="Datos",ylab="Tasa promedio de clasificación errónea",cex.main=1.5, ylim=c(0,0.08))
axis(1,at=1:8,lab=nombre,las=2,cex.axis=0.8)
lines(x,y1,type="b",pch=15, lty=1,col="blue")
lines(x,y2,type="b",pch=17, lty=2,col="red")
lines(x,y3,type="b",pch=19, lty=4,col="purple")
legend("topright",legend=c("ADNM","RLM","CHAID")
col=c("blue","red","purple"),pch=c(15,17,19),lty=c(1,2,4))
```

ANEXO 4: Código en R para el gráfico de líneas para el tiempo de procesamiento

```
y1<-c(95.96,179.66,43.41,464.41,29.19,49.55,310.44,235.73)
y2<-c(1.23,1.42,1.1,1.31,1.05,1.12,2.17,1.20)
y3<-c(48.06,28.08,2.19,29.55,0.68,0.58,12.49,15.94)
x<-1:8
nombre<-c("Crabs","Ecoli","Glass","Bupa","Electrode","Iris","Titanic","Banana")
plot(x,y1,xaxt="n",cex.axis=0.8,pch=1,bg="gray",col="blue",cex=0.1,main="Tiempo de
procesamiento",xlab="Datos",ylab="Tiempo",cex.main=1.5, ylim=c(0,500))
axis(1,at=1:8,lab=nombre,las=2,cex.axis=0.8)
lines(x,y1,type="b",pch=15, lty=1,col="blue")
lines(x,y2,type="b",pch=17, lty=2,col="red")
lines(x,y3,type="b",pch=19, lty=4,col="purple")
legend("topright",legend=c("ADNM","RLM","CHAID"),
col=c("blue","red","purple"),pch=c(15,17,19),lty=c(1,2,4))
```

ANEXO 5: Prueba de la proposición 1

Usando la identidad simple que se cumple para cualquier escalar a, $|a| = \sqrt{a}$ el disco en (1.7) puede reescribirse como

$$\text{disco}(\eta) = \frac{u(\eta)}{l(\eta)} = \frac{\sum_{g=1}^{G}\sum_{h=1}^{G} n_h n_g (\eta' B_{gh}\eta)^{1/2}}{\sum_{g=1}^{G}\sum_{h=1}^{G}\sum_{i=1}^{n_g}\sum_{j=1}^{n_h}\left[\eta' V_{gh}(i,j)\eta\right]^{1/2}}, \qquad (A.1)$$

donde B_{gh} y $V_{gh}(i,j)$ son dado en (1.8) y (1.9) respectivamente. Asumimos que

$\eta' V_{gh}(i,j)\eta \neq 0$ para i=1,… n_g , l=1,… n_h , g y h=1,… G (A. 2)

Observamos que (A. 2) implica $\eta' B_{gh}\eta \neq 0$.

Sea A una matriz simétrica. Se ve fácilmente que el vector de derivadas parciales de $(\eta' A\eta)^{1/2}$ con respecto a η es

$$\partial(\eta' A\eta)^{1/2}/\partial\eta = \frac{A\eta}{(\eta' A\eta)^{1/2}}, \quad (A.3)$$

Usando (A.3), obtenemos el gradiente de disco

$$\nabla disco(\eta) = \sum_{g=1}^{G}\sum_{h=1}^{G}\left[\frac{n_g n_h B_{gh}\eta}{(\eta' B_{gh}\eta)^{\frac{1}{2}}}\right]/l(\eta)$$

$$-u(\eta)\sum_{g=1}^{G}\sum_{h=1}^{G}\sum_{i=1}^{g}\sum_{j=1}^{h}\left[\frac{V_{gh}(i,j)\eta}{\{\eta' V_{gh}(i,j)\eta\}^{\frac{1}{2}}}\right]/[l(\eta)]^2,$$

y por (A.1),

$$\nabla disco(\eta) = \left(\sum_{g=1}^{G}\sum_{h=1}^{G}\left[\frac{n_g n_h B_{gh}\eta}{(\eta' B_{gh}\eta)^{\frac{1}{2}}}\right]\right.$$

$$\left.-disco(\eta)\sum_{g=1}^{G}\sum_{h=1}^{G}\sum_{i=1}^{g}\sum_{j=1}^{h}\left[\frac{V_{gh}(i,j)\eta}{\{\eta' V_{gh}(i,j)\eta\}^{\frac{1}{2}}}\right]\right)/l(\eta),$$

y por (1.11) y (1.12) obtenemos el resultado requerido (1.13).

Prueba del teorema 1. (a) Primero simplificamos el término B $(\eta)\,\eta$ en (1.16). Utilizando (1.8) y $(\eta' B_{gh}\eta)^{1/2} = |\eta'(\bar{X}(g) - \bar{X}(h))|$, tenemos

$$\frac{B_{gh}\eta}{(\eta' B_{gh}\eta)^{1/2}} = [\bar{X}(g) - \bar{X}(h)]sgn(\eta'[\bar{X}(g) - \bar{X}(h)]), \quad (A.4)$$

que es una función vectorial constante, donde sgn (x) es la función signo tomando valores -1,1,0. Por (1.11) y (A.4), vemos que B(η)η es una suma de funciones constantes, por lo que la matriz de las derivadas parciales de B(η)η con respecto a η es cero, y por la regla de la cadena la matriz de las derivadas parciales de $V(\eta)^{-1}B(\eta)\eta$ en (1.16) con respecto a η es cero. La matriz de las derivadas parciales de η en (1.16) con respecto a η es la matriz identidad, I; el vector de filas de las derivadas parciales de disco (η) es $\nabla disco(\eta)'$. Así que el Jacobiano del $g_2(\eta)$ en (1.16) será:

$$\nabla g_2(\eta) = I[1 - \alpha disco(\eta)] - \alpha\eta\nabla\text{disco}(\eta)'. \quad (A.5)$$

(b) Primero observamos el siguiente resultado: Sea a y b vectores de longitud n, entonces la matriz ab′ es de rango uno que admite (n-1) 0 valores propios y un valor propio de b′a asociado con el vector propio a. Usando (1.13) y (A.1), tenemos

$$\eta'\nabla disco(\eta) = \frac{[u(\eta) - disco(\eta)l(\eta)]}{l(\eta)} = 0, \quad (A.6)$$

y usando el resultado anterior, vemos que la matriz $\eta\nabla disco(\eta)'$, que es de rango uno, tiene todos sus valores propios 0. De aquí se deduce que los valores propios de $\nabla g_2(\eta)$ y de su transposición son todos iguales a 1-αdisco(η); y $\|\nabla g_\alpha(\eta)\|_2 = |1 - \alpha disco(\eta)|$.

(c) se deduce fácilmente de (b).