UNIVERSIDAD NACIONAL DE COLOMBIA

Comparación entre métodos para clasificación usando algunas distribuciones multivariadas

Catalina Inés Cortés Vélez

Universidad Nacional de Colombia
Escuela de Estadística
Medellín, Colombia
2014

Comparación entre métodos para clasificación usando algunas distribuciones multivariadas

Catalina Inés Cortés Vélez

Tesis de grado presentada como requisito parcial para optar al título de:
Magister en Ciencias-Estadística

Director(a):
Juan Carlos Salazar Uribe, Ph.D.

Línea de Investigación:
Bioestadística

Universidad Nacional de Colombia
Escuela de Estadística
Medellín, Colombia
2014

El éxito en la vida se mide con la vara de los objetivos que te has fijado

Apostolos Doxiadis

Agradecimientos

Agradezco al profesor Juan Carlos Salazar Uribe, profesor asociado, Escuela de Estadística, Universidad Nacional de Colombia, Medellín, por su paciencia y compromiso para orientar mi tesis de grado y a todos los profesores de la Escuela que se comprometieron con mi formación en la maestría.

Resumen

El problema de establecer similitudes o diferencias en áreas como la genética, biología, ciencias médicas, ingeniería, entre otras, es llamado problema de clasificación, consiste en asignar una pertenencia a determinado individuo ya sea por sus características, orden o estructura. En un trabajo previo Salazar, Vélez y Salazar [1] comparan vía simulación la eficiencia de las máquinas de soporte vectorial y la Regresión Logística, para datos que necesiten la clasificación en dos grupos y que posean una distribución univariada.
En este trabajo se compara la eficiencia de Regresión Logística, Máquinas de Soporte Vectorial, Análisis Discriminante y Clasificador Fuzzy, para clasificar un grupo de datos en dos categorías mutuamente excluyentes, en el escenario de datos multivariados provenientes de poblaciones con distribución normal multivariada, normal asimétrica y t multivariada. Dicha eficiencia o desempeño se medirá con la tasa de clasificación errónea.

Palabras clave: Clasificación, Máquinas de Soporte Vectorial, Regresión Logística, Análisis Discriminante Lineal, Tasa de clasificación errónea.

Abstract

The problem of establishing similarities or differences in fields such as genetics, medical sciences, engineering, just to mention some of them is known as classification. This process consists on assigning a subject to a specific group according to his/her features, order or structure.
In a previous work, Salazar and Salazar compared the efficiency of both Support Vector Machines -SVM- and Logistic Regression -LR-, using two groups and univariate distributions by means of a simulation study.
In this work, we compare the efficiency of the following classifiers to classify a dataset in two category mutually exclusive: Support Vector Machines -SVM-, Logistic Regression -LR-, Discriminant Analysis -DA- and Fuzzy Classifier. The comparison is carried out using multivariate data coming from several multivariate populations. Such efficiency is measured through the False Discovery Rate -FDR-.

Keywords: Classification, Support Vector Machines, Logistic Regression, Linear Discriminant Analysis , False Discovery Rate.

[1] Salazar D.A., Vélez J.I., Salazar J.C. (2012). Comparison between SVM and Logistic Regression: which one is better to discriminate? *Revista Colombiana de Estadística.***35**(2),223-237

Contenido

1. Introducción

Día a día el mundo se encuentra en una necesidad constante separar o dividir la información de acuerdo al área de desempeño como búsquedas en internet [Bull, 2012], clasificación de clientes en un banco [Ravi Kumar and Ravi, 2007], clasificación de un paciente en un centro médico [Anderer et al., 1994], identificación de enfermendades [Jen et al., 2012], detección de imágenes [Press, 2006], entre otras , lo que hace que la información adquirida deba ser más eficiente y concisa.

El hecho de determinar subgrupos de una población tal que sean homogéneos o compartan cierta información intragrupal basándose en información proporcionada por las observaciones o mediciones es el principal objetivo del Análisis Discriminante Lineal (*Linear Dirscriminant Analysis* - LDA) [Fisher, 1936, Fisher, 1938].El Análisis Discriminante Lineal fue introducido a partir de un estudio taxonómico para determinar la diferencia entre los individuos de distintos grupos y las similitudes entre los individuos de un mismo grupo por medio de una función discriminante que corresponde a una combinación lineal de las variables de los datos conocidos.

La Regresión Logística (*Logistic Regression* - LR) [Hosmer and Lemeshow, 2000] es propuesta como alternativa de clasificación [Johnson and Wichern, 2007] permite predecir la pertenencia de un nuevo individuo dentro de una categoría a partir de un modelo que cuantifica la probabilidad de pertenencia de éste nuevo individuo, convirtiéndose en un método muy usado dentro de la literatura estadística.

Las Maquinas de Soporte Vectorial (*Support Vector Machine* - SVM), introducidas en la decada de los noventa por Vapnik [Cortes and Vapnik, 1995, Vapnik, 1998], son una técnica muy competitiva en el estudio de clasificación, ya que cuenta de un alto nivel teórico y su gran aproximación al campo aplicado, por medio de una transformación lineal a espacios de dimensión superior a la dimensión del conjunto de datos, permitiendo de una manera más sencilla la predicción de nuevos individuos de acuerdo a una variable categórica.

Por otro lado se encuentra el Clasificador Fuzzy (*Fuzzy Classifiers* - FC) [Bezdek et al., 1984] basado en la lógica difusa o lógica borrosa que pretende agrupar objetos de un universo en grupos cuyos niveles de pertenencia son vistos como grados difusos, es decir, como ponderaciones que determinan la pertenencia a una categoría, el algoritmo más conocido en este

clasificador es *fuzzy isodata*.

En estudios realizados para evaluar algunos clasificadores, se han considerado datos univariados [Salazar et al., 2012] y bivariados[Hernández and Correa, 2009] pero poco se conoce acerca de los clasificadores para datos multivariados, por ello surge la pregunta ¿Cuáles pueden ser los clasificadores más eficientes en presencia de datos multivariados?

Es por ello que el presente trabajo tiene como objetivo comparar la eficiencia o desempeño de los clasificadores SVM, FC, LR y LDA usando datos multivariados provenientes de distribuciones multivariadas. El desempeño es medido por la tasa de clasificación errónea, que determinará en cuál distribución y bajo qué circunstancias es más adecuado el uso de uno de los clasificadores, estableciendo parámetros para futuras investigaciones relacionadas con el tema de clasificación.

En los estudios anteriores [Salazar et al., 2012, Hernández and Correa, 2009], de análisis de clasificadores, (SVM, LR y LDA) se requiere conocer un conjunto de datos previo relacionado al estudio, que contenga las clases a las cuales pertenence cada dato, para así lograr separar los datos nuevos de acuerdo a su clase. En este estudio se adiciona el Clasificador Fuzzy, el cual posee la ventaja de no necesitar un conjunto de datos previo para determinar la pertenencia a la clase de la cual proviene.
El desarrollo del presente trabajo se divide en cuatro capítulos. En el capítulo 1, llamado algunos clasificadores multivariados, se describen los principios de cada clasificador estudiado; en el capítulo 2, llamado algunas distribuciones multivariadas, se describe los tópicos correspondientes a las distribuciones normal multivariada, normal asimétrica multivariada y t multivariada; el capítulo 3, llamado metodología del estudio, comprende la estructura del estudio de simulación bajo los diferentes escenarios considerados y el respectivo algoritmo usado para la simulación. En el capítulo 4 mostramos los resultados obtenidos en el proceso de simulación con los gráficos correspondientes de acuerdo a cada escenario. En el capítulo 5 se encuentra la aplicación en poblaciones con características similares a las contempladas en los tres escenarios. Al final del trabajo se encuentra un capítulo de anexos donde se consignan las tablas numéricas con las tasas de clasificación errónea del capítulo de resultados y los programas usados en el paquete estadístico R [R Core Team, 2013].

2. Algunos clasificadores multivariados

A continuación se presenta una descripción de los cuatro clasificadores estadísticos usados en el estudio, ellos son SVM, FC, LR y LDA.

2.1. Máquinas de Soporte Vectorial

Las SVM son una técnica implementada por [Cortes and Vapnik, 1995] donde trata teóricamente los supuestos para encontrar un espacio que permita conocer la pertenencia a un grupo específico de un dato desconocido. En 1998 [Vapnik, 1998] generaliza la teoría de aprendizaje, estima los vectores soporte y determina la fundamentación estadística de la teoría de aprendizaje.
Las SVM han tenido diversos usos dentro de los más estudiados esta el reconocimiento de manuscritos, el reconocimiento de objetos, la identificación de voz, la detección y reconocimiento de imágenes, la minería de datos, entre otros.
En este trabajo se expondrán las SVM como un clasificador de datos, el cual permitirá separar en dos clases un conjunto de datos; dicha separación se hace por medio del aprendizaje automático, permitiendo separar un nuevo grupo de datos que no pertenezca al conjunto de aprendizaje pero que si provengan de la misma población. A continuación se explican los casos de las SVM: linealmente separable, no separable linealmente y la función kernel.

2.1.1. Caso linealmente separable

Las SVM para el caso separable consiste en un grupo de datos en un espacio determinado los cuales pueden ser clasificados en dos clases por medio de una margen o linea.
Como se muestra, por ejemplo, en la **Figura 2-1** para el caso bidimensional, pueden existir diferentes márgenes que separen los datos, pero el objetivo es encontrar aquella única margen que separe los datos (llamada *hiperplano de separación*) tal que se maximice la distancia entre los puntos más cercanos de las distintas clases.
Considérese un conjunto de m datos de entrenamiento $\{\boldsymbol{x}_k, \boldsymbol{y}_k\}_{k=1}^{m}$ tal que $\boldsymbol{x}_k \in R^n$ es un dato de entrada que es clasificado en el proceso de entrenamiento a la clase perteneciente asignándole el valor de $\boldsymbol{y}_k \in \{-1, 1\}$, según sea el caso ($-1 \in \mathcal{C}_1$, $1 \in \mathcal{C}_2$).

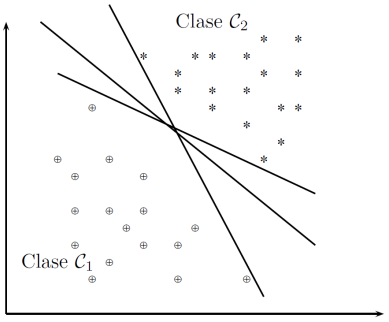

Figura 2-1.: Diferentes margenes a trazar en un problema con dos clases linealmente separables.

Para este caso se tiene el clasificador lineal

$$f(\boldsymbol{x}) = sign(\boldsymbol{w}^T\boldsymbol{x} + b)$$

donde $sign(\cdot)$ es la función signo, $\boldsymbol{w}$ es un vector normal al hiperplano de separación y b es un escalar Cuando los datos pertenecientes a las dos clases son separables linealmente, se puede afirmar lo siguiente:

$$\begin{aligned} \boldsymbol{w}^T\boldsymbol{x}_k + b \geq +1 \quad &\text{si} \quad y_k = +1 \\ \boldsymbol{w}^T\boldsymbol{x}_k + b \leq -1 \quad &\text{si} \quad y_k = -1 \end{aligned}$$

El vector $\boldsymbol{w}$ y el escalar b, se determinan de tal forma que maximicen la separación o distancia entre el hiperplano de separación y los puntos más cercanos de las dos clases.
Lo anterior, se puede formular como un problema de primalidad[1] (PP) en $\boldsymbol{w}$ así:

$$\begin{aligned} PP: \quad &\min_{\boldsymbol{w},b} J(\boldsymbol{w}) = \tfrac{1}{2}\boldsymbol{w}^T\boldsymbol{w} \\ &\text{tal que} \quad y_k[\boldsymbol{w}^T\boldsymbol{x}_k + b] \geq 1, \quad k = 1, 2, \cdots, m \end{aligned}$$

Aquellos puntos que cumplen la igualdad son llamados *vector soporte* y se encuentran sobre la margen del hiperplano de separación tal como lo muestra la **Figura 2-2**.

[1]El problema de primalidad consiste en encontrar el mínimo valor de la función $J(\boldsymbol{w})$ tal que $\boldsymbol{w}$ y b cumplan la desigualdad especificada.
La función $J(\boldsymbol{w})$ se define como un medio del producto punto entre el vector transpuesto, por si mismo.

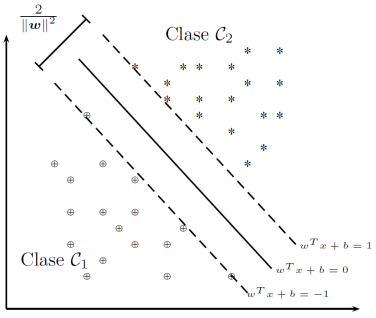

Figura 2-2.: Máxima margen $\frac{2}{\|\boldsymbol{w}\|^2}$ del hiperplano de separación y los vectores soporte de las dos clases.

2.1.2. Caso no linealmente separable

El caso no linealmente separable se presenta cuando los datos de los dos grupos a ser clasifiados se encuentran traslapados, esto es, no existe un hiperplano de separación capaz de separarlos sin que algunos datos de una clase permanezcan en la otra, como lo muestra la **Figura 2-3**.

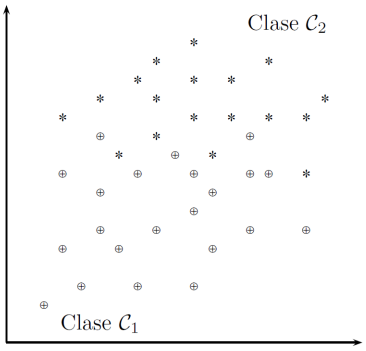

Figura 2-3.: Dos clases no linealmente separables.

Para este caso se considera el conjunto de datos de entrenamiento clasificados de la misma forma que el caso linelmente separable, se les asigna -1 o 1 de acuerdo al grupo de pertenencia, además, se debe introducir una variable de holgura ξ_k, que corresponde a la medida del error, esto es, si la variable de holgura $\xi_k > 0$, indica que el dato se encuentra en el lado incorrecto y si es $\xi_k = 0$ indica que la variable está al lado correcto.
Al ingresar esta variable de holgura se modifican las desigualdades de clasificación:

$$(\boldsymbol{w}^T X_k + b) + \xi_k \geq +1$$
$$(\boldsymbol{w}^T X_k + b) - \xi_k \leq -1$$

La suma de todas las variables de holgura $\sum_{k=1}^{m} \xi_k$, para aquellos casos donde $\xi_k \neq 0$, se puede tomar como tipo de medida para el error de clasificación.
Para el caso no linealmente separable, el problema de primalidad (PP) se convierte en:

$$\begin{aligned} PP: \quad & \min_{\boldsymbol{w},b,\xi} J(\boldsymbol{w},\xi) = \tfrac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + c\sum_{k=1}^{m}\xi_k \\ & \text{tal que} \;\; y_k[\boldsymbol{w}^T X_k + b] \geq 1 - \xi_k, \quad \xi_k \geq 0, \quad k = 1, 2, \cdots, m \end{aligned}$$

donde c es una constante real positiva.
En este caso de variables no linealmente separables es necesario realizar un mapeo de los datos a un espacio de dimensión superior, llamado *espacio característico*, en el cual se puedan separar los datos linealmente.
Dicho mapeo se logra con una transformación biyectiva no lineal, $\varphi(\cdot)$, la cual convierte el problema no separable en un problema separable de un espacio de dimensión mayor. De dicha transformación no es necesario conocer su estructura, lo que importa es conocer su producto interno $K(\boldsymbol{x}, \boldsymbol{z})$, llamado función kernel, el cual permitirá convertir el problema no separable en uno separable, tal que:

$$\begin{aligned} \boldsymbol{w}^T\varphi(X_k) + b \geq +1 \qquad & \text{si} \quad y_k = +1 \\ \boldsymbol{w}^T\varphi(X_k) + b \leq -1 \qquad & \text{si} \quad y_k = -1 \end{aligned}$$

Así se logra obtener un hiperplano de separación no lineal para los datos, como lo muestra la **Figura 2-4**:

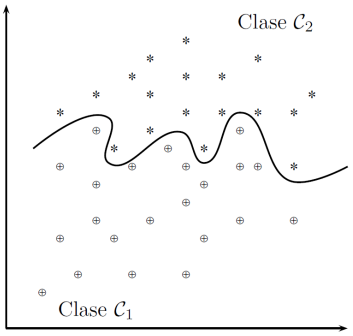

Figura 2-4.: Hiperplano de separación para las dos clases no linealmente separables.

2.1.3. Función Kernel

La función Kernel es aquel mapeo que permite convertir un problema de separación no lineal a un problema de separación lineal en un *espacio característico*. Para que una función sea

considerada como Kernel debe cumplir las condiciones de continuidad, simetría y semidefinida positiva.
Las funciones Kernel más usadas se muestran en la **Tabla 2-1**:

Kernel	Función
Lineal	$K_L(\boldsymbol{x}, \boldsymbol{z}) = \langle \boldsymbol{x}, \boldsymbol{z} \rangle = \sum_{i=1}^{m} x_i z_i$
Polinómico	$K_P(\boldsymbol{x}, \boldsymbol{z}) = (\langle \boldsymbol{x}, \boldsymbol{z} \rangle)^p$
	donde p es el grado del polinómio
Gaussiano	$K_G(\boldsymbol{x}, \boldsymbol{z}) = \exp\{-\frac{\|\boldsymbol{x}-\boldsymbol{z}\|^2}{2\sigma^2}\}$
	donde $\sigma^2 > 0$

Tabla 2-1.: Kernel más usados en la literatura de SVM.

2.2. Clasificador Fuzzy

La teoría de los conjuntos difusos y la lógica difusa (*Fuzzy*) fue introducida [Zadeh, 1965] como una forma de trabajar la no linealidad de una manera eficiente, permitiendo la decisión a partir de reglas lingüísticas.
Este clasificador no necesita usar un conjunto de datos de entrenamiento para formar las reglas de decisión, él mismo se corrige hasta encontrar la clasificación adecuada, sólo conociendo el número de clases (cluster) en las cuales se desean separar los datos. Posee características que permiten establecer diferentes grados de valor a la relación (pertenencia) que hay entre el dato y la clase a la que pertenece.
Este clasificador ha sido usado para reconocimiento de patrones, análisis geoestadístico, análisis de imágenes médicas, control automatizado, minería de datos, entre otros.
A continuación se estudiarán los clasificadores nitido y difuso como métodos para clasificación de un conjunto de datos en dos clases.

2.2.1. Cluster nítido

Es utilizado cuando el conjunto de datos está separado linealmente. Los clusters corresponden a la partición o subconjuntos de un conjunto mayor de datos.
Sea X un conjunto de n datos o individuos a clasificar $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \cdots, \boldsymbol{x}_n\}$, cada vector $\boldsymbol{x}_k$ es definido por m características (o variables) $\boldsymbol{x}_k = (x_{k1}, x_{k2}, \cdots, x_{km})^T$.
El número de clases es definido por el investigador de acuerdo a la cantidad de clases c en las que desee separar el conjunto de datos y que no debe ser superior al número de datos,

estos se definen como $\{A_i\}_{i=1}^{c}$ y cumplen las siguientes propiedades:

$$\begin{aligned}
&\bigcup_{i=1}^{c} A_i = X \\
&A_i \cap A_j = \emptyset \quad i \neq j \\
&A_i \subset X
\end{aligned}$$

En los cluster nítidos cada dato del conjunto a clasificar pertenece a uno y solo un cluster, para lo cual se le asigna a $\boldsymbol{x}_k$ un valor $\chi_{A_i}(\boldsymbol{x}_k)$ de 1 o 0 indicando que pertenece o no pertenece, respectivamente, al cluster A_i, esto es:

$$\chi_{A_i} = \chi_{ij} = \begin{cases} 1 & x_k \in A_i \\ 0 & x_k \notin A_i \end{cases}$$

Donde i corresponde al número de clusters y j corresponde al número de observaciones. Estos valores de pertenencia deben cumplir lo siguiente:

$$\begin{aligned}
&\sum_{i=1}^{c} \chi_{ij} = 1 \qquad\qquad \text{para} \quad j = 1, 2, 3, \ldots, n \\
&0 < \sum_{j=1}^{n} \chi_{ij} < n \qquad \text{para} \quad i = 1, 2, 3, \ldots, c
\end{aligned}$$

Cada valor de pertenencia corresponde a las componentes de la matriz $\boldsymbol{U} = [\chi_{ij}]$, llamada *matriz de representación de la partición*, y pertenece al espacio de partición M_P:

$$M_P = \left\{ U | \chi_{ij} \in \{0, 1\}; \sum_{i=1}^{c} \chi_{ij} = 1; 0 < \sum_{i=1}^{n} \chi_{ij} < n \right\}$$

De todo el espacio de partición, la matriz $\boldsymbol{U}$ corresponde a la matriz de clasificacón óptima que minimiza la función de error cuadrático $J(U, \nu)$, dada por:

$$J(U, \nu) = \sum_{k=1}^{n} \sum_{i=1}^{c} \chi_{ik} (d_{ik})^2$$

El término d_{ik} corresponde a la distacia euclidiana medida en un espacio de dimensión m entre el vector de datos x_k y el centroide ν_i del cluster i , como se muestra a continuación

$$\begin{aligned}
d_{ik} &= \|\nu_i - x_i\| \\
&= \left[\sum_{j=1}^{m} (x_{kj} - \nu_{ij})^2 \right]^{1/2}
\end{aligned}$$

para $k = 1, 2, \ldots, n$ y $i = 1, 2, \ldots, c$.
Las m componentes del centroide ν_i del cluster i se determinan a partir del cociente

$$\nu_{ij} = \frac{\sum_{k=1}^{n} \chi_{ik} x_{kj}}{\sum_{k=1}^{n} \chi_{ik}}$$

La partición óptima, que esta representada por la matriz $\boldsymbol{U}^*$, se logra fijando el número de cluster y la matriz $U^{(0)} \in M_c$, que se escoge de manera aleatoria y donde sus valores se actualizan en cada iteración por:

$$\chi_{A_i}^{(r+1)} = \begin{cases} 1 & d_{ik}^{(r)} = \text{mín}\left(d_{ik}^{(r)}\right), \quad \forall j \in c \\ 0 & \text{en otro caso.} \end{cases}$$

de tal forma que $\left\|U^{(r+1)} - U^r\right\| \leq \epsilon$, donde ϵ es un nivel de tolerancia escogido por el investigador, usualmente es de 10^{-6}.

2.2.2. Cluster difuso

El cluster difuso es utilizado cuando los datos de diferentes clases se traslapan entre si, esto es, se cuenta con un caso de datos no separables linealmente.
Sea X un conjunto de n datos o individuos a clasificar $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \cdots, \boldsymbol{x}_n\}$, cada vector $\boldsymbol{x}_k$ es definido por m características (o variables) $\boldsymbol{x}_k = (x_{k1}, x_{k2}, \cdots, x_{km})^T$.
Los clusters difusos,$\{B_i\}_{i=1}^c$, satisfacen las siguientes condiciones:

$$\begin{aligned} B_i &\neq \emptyset & i &= 1, 2, \ldots, c \\ B_i \cap B_j &= \emptyset & i &\neq j \\ \bigcup_{i=1}^{c} B_i &= X \end{aligned}$$

En el cluster difuso se tiene que cada dato del conjunto a clasificar pertenece ponderadamente a cada cluster, para lo cual se le asigna un vector de pertenencia, $\mu_{B_i}(\boldsymbol{x}_k)$, con sus componentes, μ_{ik}, en el intervalo $[0, 1]$. Estas componentes deben cumplir las siguientes restricciones.

$$\begin{aligned} \sum_{i=1}^{c} \mu_{ik} &= 1 & k &= 1, 2, \cdots, n \\ 0 < \sum_{k=1}^{n} \mu_{ik} &< n & i &= 1, 2, \cdots, c \end{aligned}$$

donde i es el indice para el closter y k el indice para la obesrvación.
En este caso el espacio de partición difusa, M_P , estará dado por el conjunto:

$$M_P = \{\boldsymbol{U} | \mu_{ij} \in [0, 1]; \textstyle\sum_{i=1}^{c} \mu_{ik} = 1; 0 < \sum_{i=1}^{n} \mu_{ik} < n\}$$

Para agrupar n datos en c clusters difusos se debe contar con el parámetro m' que controla la cantidad de difusión en el proceso de clasificación el cual puede tener valores en el intervalo $[1, +\infty)$, usualmente se emplea 1 o 2.
Se define la distiancia euclidiana, igual que en el cluster nitido como:

$$\begin{aligned} d_{ik} &= d(x_k - v_i) \\ &= \left[\textstyle\sum_{j=1}^{m} (x_{kj} - v_{ij})^2\right]^{1/2} \end{aligned}$$

Los centroides, ν_i, de cada cluster se calculan a partir de la expresión:

$$\nu_{ij} = \frac{\sum_{k=1}^{n} \mu_{ik}^{m'} x_{kj}}{\sum_{k=1}^{n} \mu_{ik}^{m'}}$$

La función cuadrática media a optimizar, para este caso, está dada por:

$$J_m(U, v) = \sum_{k=1}^{n} \sum_{i=1}^{c} \mu_{ik}^{m'} (d_{ik})^2$$

La partición óptima se encuentra siguiendo un proceso de iteración fijando c, m' e inicializando con la matriz $\boldsymbol{U}^{(0)} \in M_P$ de tal foma que los valores de la matriz de partición se actualizará por:

$$\mu_{ik}^{(r+1)} = \left[\sum_{j=1}^{c} \left(\frac{d_{ik}^{(r)}}{d_{jk}^{(r)}} \right)^{\frac{2}{m'-1}} \right]^{-1}$$

hasta que la diferencia de matrices sea menor a un nivel de tolerancia ϵ, $\left\| U^{(r+1)} - U^r \right\| \leq \epsilon$.

2.3. Análisis Discriminante

El LDA [Fisher, 1936, Fisher, 1938] parte de n datos o individuos donde se han medido m variables cuantitativas y una variable cualitativa (clasificativa), que posee las categorías o clases de pertenencia de los datos, para construir una función tal que permita clasificar un dato desconocido a la clase que pertenence. La función discriminante de Fisher D, necesaria para determinar la pertenencia de un dato a una clase, es una función lineal de m variables explicativas

$$D = u_1 X_1 + u_2 X_2 + \cdots + u_m X_m,$$

donde u_j, $j = 1, 2, \cdots, m$ corresponden a los coeficientes de ponderación, variando su valor de acuerdo a la categoría a la que pertenezca.
Considerando que el conjunto de entrenamiento cuenta con n individuos, se puede expresar la función discriminante, para cada uno, así:

$$D_i = u_1 X_{1i} + u_2 X_{2i} + \cdots + u_m X_{mi}, \quad i = 1, 2, \cdots, n$$

donde D_i corresponde a la puntuación discriminante de la i-ésima observación.
Si las m variables explicativas se expresan como desviaciones respecto a la media, cada función discriminante estará expresada como desviaciones respecto a la media, permitiendo escribir las funciones discriminantes en la forma matricial

$$\boldsymbol{D} = \boldsymbol{X}\boldsymbol{u}$$

así, la variabilidad de la función discriminante puede ser expresada como

$$\boldsymbol{D}^T \boldsymbol{D} = \boldsymbol{u}^T \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{u},$$

donde $\boldsymbol{X}^T\boldsymbol{X}$ es una matriz simétrica que puede descomponerse como la suma entre la matriz intragrupal $\boldsymbol{F}$ y la matriz residual $\boldsymbol{V}$

$$\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{F} + \boldsymbol{V}$$

así se tiene que:

$$\begin{aligned}\boldsymbol{D}^T\boldsymbol{D} &= \boldsymbol{u}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{u} \\ &= \boldsymbol{u}^T\boldsymbol{F}\boldsymbol{u} + \boldsymbol{u}^T\boldsymbol{V}\boldsymbol{u}\end{aligned}$$

Las matrices $\boldsymbol{F}$ y $\boldsymbol{V}$ se pueden obtener de los datos muestrales mientras que los coeficientes u_i son determinados según el criterio discriminante de Fisher, que maximiza la razón, $\boldsymbol{\lambda}$, entre la variabilidad intragrupal y la variabilidad de los residuales, esto es:

$$\boldsymbol{\lambda} = \frac{\boldsymbol{u}^T\boldsymbol{F}\boldsymbol{u}}{\boldsymbol{u}^T\boldsymbol{V}\boldsymbol{u}}$$

El máximo valor de $\boldsymbol{\lambda}$ es obtenido derivando parcialmente, de manera numérica, con respecto a $\boldsymbol{u}$ e igualando a cero, de aquí se obtiene el primer eje discriminante

$$\boldsymbol{V}^{-1}\boldsymbol{F}\boldsymbol{u} = \boldsymbol{\lambda}\boldsymbol{u}$$

que corresponde al vector propio $\boldsymbol{u}$ asociado a la matriz no simétrica cuyo valor propio $\boldsymbol{\lambda}$ es el mayor encontrado.
El número de ejes discriminantes corresponde al mín$(G-1, m)$ donde G es el número de clases en las que se desean clasificar los datos y m es el número de variables que se miden en los datos, que serán seleccionados de acuerdo a los valores propios asociados de mayor valor (ordenados de manera descendente), para nuestro caso en estudio, donde $G = 2$, sólo se necesitará de un eje discriminante que corresponde al mayor valor propio del sistema.

2.3.1. Puntuaciones discriminantes

Las puntuaciones discriminantes son los valores que resultan de evaluar la información de cada individuo en la función discriminante

$$D = u_1X_1 + u_2X_2 + \cdots + u_mX_m$$

que corresponde a la proyección de cada individuo sobre el eje discriminante.

2.3.2. Centroides

Los centroides son vectores de medias que resumen la información sobre los grupos. Los centroides de cada grupo están dados por:

$$\bar{X}_I = \begin{bmatrix} \bar{X}_{1,I} \\ \bar{X}_{2,I} \\ \vdots \\ \bar{X}_{m,I} \end{bmatrix} \bar{X}_{II} = \begin{bmatrix} \bar{X}_{1,II} \\ \bar{X}_{2,II} \\ \vdots \\ \bar{X}_{m,II} \end{bmatrix}$$

Para los grupos I y II se tiene que la puntuación discriminante para el vector de medias, está dado por:

$$\bar{D}_I = u_1\bar{X}_{1,I} + u_2\bar{X}_{2,I} + \cdots + u_m\bar{X}_{m,I}$$
$$\bar{D}_{II} = u_1\bar{X}_{1,II} + u_2\bar{X}_{2,II} + \cdots + u_m\bar{X}_{m,II}$$

2.3.3. Punto de corte discriminante

El punto de corte discriminante se calcula promediando la puntuación discriminante de los vectores de medias para cada grupo

$$C = \frac{\bar{D}_I + \bar{D}_{II}}{2}$$

Así, el criterio de clasificación para el i-ésimo individuo será

$$i \in I, \quad \text{si} \quad D_i - C < 0$$
$$i \in II, \quad \text{si} \quad D_i - C > 0$$

2.4. Regresión Logística

Para aplicar el modelo de LR no es necesario que las variables predictoras presenten una distribución normal y en las que puede presentarse variables numéricas y/o categóricas. Dicho modelo estudia la dependencia entre una variable binomial y un conjunto de explicativas.
El modelo LR se objeta en la ecuación

$$Y = log\{\frac{p}{1-p}\}$$

donde p es la probabilidad de que la variable dependiente tome el valor de 1 y Y es el valor de la variable dependiente.
Al expresar el modelo LR en forma de regresión, se obtiene:

$$p = \frac{e^{(\alpha+\beta x)}}{1+e^{(\alpha+\beta x)}}$$

donde α y β son los coeficientes de la ecuación.
Para el caso en que existan varias variables predicatoras, se obtiene la regresión:

$$p = \frac{e^{\alpha+\beta_1 x_1+\cdots+\beta_m x_m}}{1-e^{\alpha+\beta_1 x_1+\cdots+\beta_m x_m}}$$

donde p es la probabilidad de que la variable dependiente tome el valor de 1 con variable predictora $\boldsymbol{Y} = (x_1, x_2, \cdots, x_m)$ y los coeficientes α y β_i para $i = 1, 2, \cdots, m$ de la ecuación regresora que son estimados por el método de máxima verosimilitud.
En el caso en que p tome un valor mayor a 0,5 se asume que la variable dependiente toma el valor de 1 y si p toma un valor menor a 0,5 se asume que la variable dependiente toma el valor de 0.

De acuerdo a la importancia de clasificación que se presenta en los estudios estadísticos con datos reales, en este estudio se plantean los clasificadores SVM, LR y LDA de acuerdo al importante aporte que han hecho en la literatura estadística y por otro lado se elige FC como método alternativo de clasificación ya que su importancia radica en no necesitar un conjunto de entrenamiento para realizar el proceso de clasificación.

3. Algunas distribuciones multivariadas

A continuación se muestran las funciones de distribución de probabilidad multivariada que serán usadas en el presente estudio.

3.1. Distribución Normal Multivariada

La distribución normal multivariada (Multivariate Normal Distribution - MND) también conocida como la distribución Gaussiana multivariada es una generalización de la distribución normal univariada, como lo indica [Anderson, 1958].
El vector aleatorio p-dimensional $\mathbf{X} = (x_1, x_2, \cdots, p)^T$ tiene distribución normal p-variada con vector de medias $\boldsymbol{\mu}$ y matriz de covarianza $\boldsymbol{\Sigma}$ si su función de densidad de probabilidad está dada por:

$$f(\mathbf{X}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{X}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu})\right\},$$

donde $\boldsymbol{\Sigma} > 0$ y se denota como $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
Algunas propiedades de la distribucion normal multivariada son:

- Si un vector p-dimensional posee una distribución normal multivariada, las distribuciones marginales son normales univariadas.
- La distribución normal multivariada condicionada también es normal multivariada, de igual forma se cumple para la normal univariada condicionada.
- Si el conjunto de variables aleatorias marginales son normales univariadas e independientes, las covarianzas entre las variables son nulas.
- Si dos variables aleatorias normales multivariadas son independientes la matriz de correlación entre ellas es nula.

3.2. Distribución Normal Asimétrica

Según lo afirma Genton [Genton, 2004], las distribuciones multivariadas, en el campo aplicado, parecen ser no muy útiles para tratar datos no normales. Para el caso de la familia de la distribución normal multivariada asimétrica (Multivariate Skew Normal Distribution - MSND) [Azzalini and Dalla Valle, 1996, Azzalini and Capitanio, 1999] ocurre lo contrario,

son útiles para tratar datos no normales, sobre todo cuando la asimetría de las marginales es moderada.
Un vector aleatorio p-dimensional $\mathbf{z} = (Z_1, Z_2, \ldots, Z_p)^T$ sigue una distribución normal asimétrica, denotado por $\mathbf{z} \sim SN_p(\boldsymbol{\Omega}, \boldsymbol{\alpha})$, si la función de distribución de probabilidad de $\mathbf{z}$ es:

$$g(\mathbf{z}) = 2\varphi_p(\mathbf{z}; \boldsymbol{\Omega})\Phi(\boldsymbol{\alpha}^T\mathbf{z}), \qquad \mathbf{z} \in R^p,$$

donde $\varphi_p(\mathbf{z}; \boldsymbol{\Omega})$ denota la función de distribución de probabilidad de una distribución normal multivariada p-dimensional, con marginales estandarizadas y matriz de correlación $\boldsymbol{\Omega}$, $\Phi(\boldsymbol{\alpha}^T\mathbf{z})$ denota la función de distribución acumulada de la distribución normal multivariada p-dimensional y $\boldsymbol{\alpha}$ es el vector de p números reales.
La función de distribución acumulada de la distribución normal asimétrica multivariada muestra una estrecha relación entre las distribuciones multivariadas normal sesgada y normal.
Si $\mathbf{z} \sim SN_p(\boldsymbol{\Omega}, \boldsymbol{\alpha})$ entonces la función de distribución acumulada corresponde a:

$$\begin{aligned} G(\mathbf{z}) &= P(Z_1 \leq z_1, \ldots, Z_p \leq z_p) \\ &= 2\int_{-\infty}^{z_1} \cdots \int_{-\infty}^{z_p} \varphi_p(\mathbf{z}; \boldsymbol{\Omega})\Phi(\boldsymbol{\alpha}^T\mathbf{z}) dz_1 \cdots dz_2 \end{aligned}$$

para $\mathbf{z} \in R^p$
El cálculo de la función generadora de momentos de $\mathbf{z}$ con distribución normal asimétrica multivariada está dada por:

$$M(t) = 2\exp\left\{\tfrac{1}{2}\mathbf{t}^T\Omega\mathbf{t}\right\}\Phi\left\{\tfrac{\boldsymbol{\alpha}^T\boldsymbol{\Omega}\mathbf{t}}{1+\boldsymbol{\alpha}^T\boldsymbol{\Omega}\boldsymbol{\alpha}}\right\}^{1/2}$$

Por medio de la función generadora de momentos es fácil hallar el vector de media y la matriz de varianza, dadas respectivamente, por:

$$\begin{aligned} E(\mathbf{z}) &= \left(\tfrac{2}{\pi}\right)^{1/2}\boldsymbol{\delta} \\ Var(\mathbf{z}) &= \boldsymbol{\Omega} - \tfrac{2}{\pi}\boldsymbol{\delta}\boldsymbol{\delta}^T \end{aligned}$$

Para los cuales se tiene que:

$$\boldsymbol{\delta} = \tfrac{1}{(1+\boldsymbol{\alpha}^T\boldsymbol{\Omega}\boldsymbol{\alpha})^{1/2}}\boldsymbol{\Omega}\boldsymbol{\alpha}$$

3.3. Distribución t Multivariada

Sea U es una variable aleatoria independiente que se distribuye normal estándar y χ^2_ν es una variable aleatoria independiente que se distribuye chi-cuadrado con ν grados de libertad.
La función de densidad de la distribución t es:

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi}\sigma}\left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

donde $-\infty < t < \infty$ y parámetro $\nu > 0$ y $\sigma > 0$.
La gráfica de la función de densidad de probabilidad de una distribución t, es más platicúrtica con respecto a la distribución normal, se acerca a la de una distribución normal cuando $n \to \infty$. Esto indica que para valores de ν grandes la gráfica se hace menos platicúrtica.
Teniendo en cuenta la anterior definición de la distribución t univariada se define la distribución t multivariada (Multivariate T Distribution - MTD) [Kotz and Nadarajah, 2004] (centrada $\delta = 0$, no centrada $\delta \neq 0$). Por tanto, la función de densidad de un número p de variables independientes, que se distribuyen t multivariada, es simplemente el producto de las funciones de densidad individuales con distribución t univariada.

$$f(\mathbf{y}) = \prod_{j=1}^{p} f(y_j)$$

Así:

$$f(\mathbf{y}) = \frac{\Gamma\left(\frac{\nu+p}{2}\right)}{(\pi\nu)^{p/2}\Gamma\left(\frac{\nu}{2}\right)|\mathbf{R}|^{1/2}}(1 + \nu^{-1}\mathbf{y}^T\mathbf{R}^{-1}\mathbf{y})^{\frac{\nu+p}{2}},$$

donde $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_p$ tienen una distribución normal multivariada con $\mathbf{R}$ la matriz de varianza-covarianza y $\mathbf{y}$ el vector con distribución t multivariada.

La mayoría de los estudios estadísticos usan datos provenientes de la distribución normal (univariada o multivariada), pero en la mayoría de los casos reales los datos no se comportan como dicha distribución, es por ello que en este estudio se tomaron las distribuciones MND como el caso más general en la literatura, MSND que es una distribución considerada más cercana a los datos reales y MTD cuya importancia radica en tener colas más pesadas o menos pesadas que una distribución MND.

4. Metodología del estudio

Este estudio de simulación tiene como objetivo comparar por medio de la tasa de clasificación errónea [1] (TCE), los clasificadores estadísticos: SVM, FC, LR y LDA para el caso con dos grupos de datos provenientes de poblaciones multivariadas con distribución MND, MSND y MTD.
Cada escenario corresponde a dos conjuntos de datos con 2, 3 y 4 números de variables, con tamaños de muestra 20, 50 y 100 y diferentes distancias entre los vectores de medias (2, 3, 4, 5 y 6) que determinan una fuerte superposición entre el par de grupos de datos hasta una leve superposición entre estos mismos. Además, en cada escenario se hace la simulación con dos tipos de matrices de varianza-covarianza entre los dos grupos de datos.

4.1. Escenario 1: Dos variables

En este escenario se consideran conjuntos de datos que poseen dos variables y son generados con las distribuciones multivariadas MND, MSND y MTD con distintos tamaños de muestra (20, 50 y 100) y distintas distancias entre los vectores de medias de cada grupo (2, 3, 4, 5 y 6), con las matrices de varianza-covarianza:

$$\begin{bmatrix} 1 & 0.3 \\ 0.3 & 1 \end{bmatrix} \begin{bmatrix} 10 & 3 \\ 3 & 2 \end{bmatrix}$$

La matriz del lado izquierdo ha sido utilizada en diferentes estudios [Salazar et al., 2012, Hernández and Correa, 2009] hechos previamente y la matriz del lado derecho es tomada arbitrariamente en este estudio.
Para los diferentes casos considerados en este escenario a sido necesario fijar algunos parámetros para las distribuciones MSND y MTD. En la distribución MSND se fija el parámetro de forma consistiendo en el vector $[-1, 1]$ y en la distribución MTD se fija los grados de libertad con $\nu = 10$.
Así, son analizados cinco casos en los que se se tienen diferentes vectores de medias con sus diferentes distancias y los diferentes números de datos.

Caso 1: los vectores de medias para los dos grupos tienen una distancia de 2, los vectores de medias correspondientes son $[-1, 0]$ y $[1, 0]$.

[1] La tasa de clasificación errónea corresponde al porcentaje de datos mal clasificados, es decir, el cociente entre los datos mal clasificados y el total de datos.

Caso 2: los vectores de medias para los dos grupos tienen una distancia de 3, los vectores de medias correspondientes son $[-1, 0]$ y $[2, 0]$.

Caso 3: los vectores de medias para los dos grupos tienen una distancia de 4, los vectores de medias correspondientes son $[-2, 0]$ y $[2, 0]$.

Caso 4: los vectores de medias para los dos grupos tienen una distancia de 5, los vectores de medias correspondientes son $[-2, 0]$ y $[3, 0]$.

Caso 5: los vectores de medias para los dos grupos tienen una distancia de 6, los vectores de medias correspondientes son $[-3, 0]$ y $[3, 0]$.

4.2. Escenario 2: Tres variables

En este escenario se consideran conjuntos de datos que poseen tres variables y son generados con las distribuciones multivariadas MND, MSND y MTD con distintos tamaños de muestra (20, 50 y 100) y distintas distancias entre los vectores de medias de cada grupo (2, 3, 4, 5 y 6), con las matrices de varianza-covarianza:

$$\begin{bmatrix} 1 & 0.3 & 0.3 \\ 0.3 & 1 & 0.3 \\ 0.3 & 0.3 & 1 \end{bmatrix} \begin{bmatrix} 4 & 3 & 2 \\ 3 & 3 & 2 \\ 2 & 2 & 2 \end{bmatrix}$$

La matriz del lado izquierdo se toma de tal forma que guarde una coherencia con los valores de la matriz tomada en el esccenario 1 y la matriz del lado derecho es tomada arbitrariamente en este estudio.
Para los diferentes casos considerados en este escenario a sido necesario fijar algunos parámetros para las distribuciones MSND y MTD. En la distribución MSND se fija el parámetro de forma consistiendo en el verctor $[-2, 0, 2]$ y en la distribución MTD se fina los grados de libertad con $\nu = 10$.
Así, son analizados cinco casos en los que se se tienen diferentes vectores de medias con sus diferentes distancias y los diferentes números de datos.

Caso 1: los vectores de medias para los dos grupos tienen una distancia de 2, los vectores de medias correspondientes son $[-1, 0, 0]$ y $[1, 0, 0]$.

Caso 2: los vectores de medias para los dos grupos tienen una distancia de 3, los vectores de medias correspondientes son $[-1, 0, 0]$ y $[2, 0, 0]$.

Caso 3: los vectores de medias para los dos grupos tienen una distancia de 4, los vectores de medias correspondientes son $[-2, 0, 0]$ y $[2, 0, 0]$.

Caso 4: los vectores de medias para los dos grupos tienen una distancia de 5, los vectores de medias correspondientes son $[-2, 0, 0]$ y $[3, 0, 0]$.

Caso 5: los vectores de medias para los dos grupos tienen una distancia de 6, los vectores de medias correspondientes son $[-3, 0, 0]$ y $[3, 0, 0]$.

4.3. Escenario 3: Cuatro variables

En este escenario se consideran conjuntos de datos que poseen cuatro variables y son generados con las distribuciones multivariadas MND, MSND y MTD con distintos tamaños de muestra (20, 50 y 100) y distintas distancias entre los vectores de medias de cada grupo (2, 3, 4, 5 y 6), con las matrices de varianza-covarianza:

$$\begin{bmatrix} 1 & 0.3 & 0.3 & 0.3 \\ 0.3 & 1 & 0.3 & 0.3 \\ 0.3 & 0.3 & 1 & 0.3 \\ 0.3 & 0.3 & 0.3 & 1 \end{bmatrix} \begin{bmatrix} 4 & 3 & 2 & 1 \\ 3 & 3 & 2 & 1 \\ 2 & 2 & 2 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

La matriz del lado izquierdo se toma de tal forma que guarde una coherencia con los valores de la matriz tomada en los esccenarios 1 y 2 y la matriz del lado derecho es tomada arbitrariamente en este estudio.
Para los diferentes casos considerados en este escenario a sido necesario fijar algunos paámetros para las distribuciones MSND y MTD. En la distribución MSND se fina el parámetro de forma consistiendo en el vector $[-2, -1, 1, 2]$ y en la distribución MTD se fija los grados de libertad con $\nu = 10$.
Así, son analizados cinco casos en los que se tienen diferentes vectores de medias con sus diferentes distancias y los diferentes números de datos.

Caso 1: los vectores de medias para los dos grupos tienen una distancia de 2, los vectores de medias correspondientes son $[-1, 0, 0, 0]$ y $[1, 0, 0, 0]$.

Caso 2: los vectores de medias para los dos grupos tienen una distancia de 3, los vectores de medias correspondientes son $[-1, 0, 0, 0]$ y $[2, 0, 0, 0]$.

Caso 3: los vectores de medias para los dos grupos tienen una distancia de 4, los vectores de medias correspondientes son $[-2, 0, 0, 0]$ y $[2, 0, 0, 0]$.

Caso 4: los vectores de medias para los dos grupos tienen una distancia de 5, los vectores de medias correspondientes son $[-2, 0, 0, 0]$ y $[3, 0, 0, 0]$.

Caso 5: los vectores de medias para los dos grupos tienen una distancia de 6, los vectores de medias correspondientes son $[-3, 0, 0, 0]$ y $[3, 0, 0, 0]$.

4.4. Procedimiento de comparación

Las cuatro técnicas de clasificación (SVM, FC, LR y LD) en sus diferentes escenarios, fueron comparadas a partir de una secuencia de pasos descritos a continuación:

Paso 1: Se simularon dos conjuntos de datos de cada distribución, agrupados en un arreglo llamado *conjunto a clasificar*.

Paso 2: Se simularon dos conjuntos de datos de cada distribución, con un tamaño del 50 % de los datos a clasificar, con idénticas características del conjunto de datos a clasificar.

Paso 3: Con los datos simulados en el Paso 2 se realizó el entrenamiento de los clasificadores SVM, LR y LD [2].

Paso 4: Se clasificaron los datos obtenidos en el Paso 1.

Paso 5: Se calcularon las TCE para las clasificaciones obtenidas en el Paso 4.

Paso 6: Se repitio 5000 veces del Paso 1 al Paso 5, luego se hizo un promedio de todas las TCE obtenidas en estas 5000 repeticiones.

Todo el proceso de comparación que se realizó en este trabajo fue llevado a cabo por medio del paquete de uso público R [R Core Team, 2013]. Para SVM se usó la función `svm()` del paquete `e1071` que depende del paquete `class`, para FC se usó la función `cmeans()` del paquete `e1071`, para LR se usó la función `glm()` del paquete `stats` y para LDA se usó la función `lda()` del paquete `MASS`.

[2] FC no se incluye en este paso, ya que no necesita de conjunto de entrenamiento para clasificar.

5. Resultados

En este capítulo se presentan los resultados obtenidos en los tres escenarios de simulación. Las tablas con las TCE se encuentran en el Anexo A.
En los escenarios considerados se calcula la TCE para dos tipos de matrices de varianza-covarianza, la matriz considerada a partir de estudios previos se denominó matriz previa y la matriz arbitraria se llamó matriz arbitraria.

5.1. Escenario 1: Dos variables

En la **Figura 5-1** se presentan los resultados obtenidos con la matriz previa. Se puede observar como en los diferentes casos considerados el LDA es el clasificador que presentó mayor eficiencia obteniendo la menor TCE seguido de LR. Para las distribuciones MND y MTD se observa que el clasificador de menor eficiencia es SVM mientras que para la distribución MSND el de menor eficiencia es FC. A medida que disminuye el traslape entre los conjuntos de datos disminuye la TCE siendo muy cercana a cero.

En la **Figura 5-2** se presentan los resultados obtenidos con la matriz arbitraria. Se observa que cuando los datos poseen una distancia de 2 unidades entre los vectores de medias para los diferentes casos de este escenario, el clasificador FC presenta una menor TCE en comparación al SVM, con una diferencia entre 1 % y 4 %, aunque a medida que aumenta la distancia entre los vectores de medias mejora la eficiencia de SVM comparada con FC. Se presenta la constanter de ser LDA la de mayor eficiencia (menor TCE) seguida de LR.

5.2. Escenario 2: Tres variables

De los resultados de la simulación presentados en la **Figura 5-3** con la matriz previa se observa que para las distribuciones MND y MTD el clasificador FC tiene un mejor desempeño que el clasificador SVM sin importar la distancia entre las medias de los dos grupos de datos considerados.
Para la distribución MSND se observa que SVM posee un mejor desempeño que FC y que en el caso en que las distancias entre medias aumentan, las TCE en dicha distribución, tiende a comportarse igual que en la distribución MND obteniendo valores muy semejantes, caso

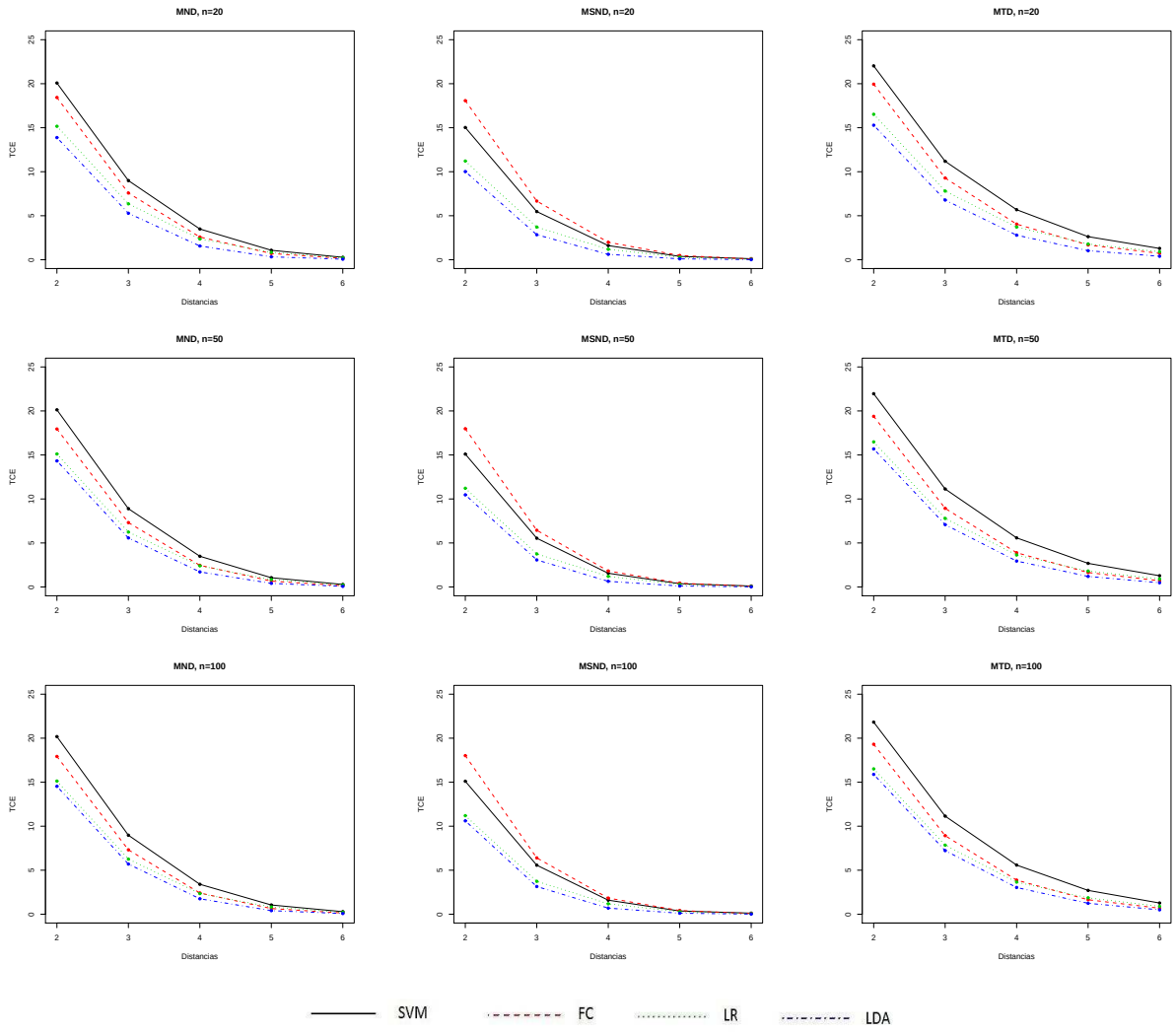

Figura 5-1.: TCE para el escenario 1, con dos variables, diferentes distancias entre medias y matriz previa. De arriba hacia abajo se cuentan los tamaños de muestra 20, 50 y 100, respectivamente, y de izquierda a derecha se observan las distribuciones MND, MSND y MTD, respectivamente.

contrario con la distribución MTD que se alcanza a observar una leve diferencia entre la eficiencia de los clasificadores.
En la distribución MTD, cuando aumenta la distancia entre la media de los dos grupos de datos mejora la eficiencia del clasificador FC siendo superior a LR. Para los casos considerados es constante la mayor eficiencia de LDA.

En la simulación usando la matriz arbitraria se observa, de acuerdo a la **Figura 5-4**, que en los cinco casos considerados en los diferentes tamaños de muestra es constante el com-

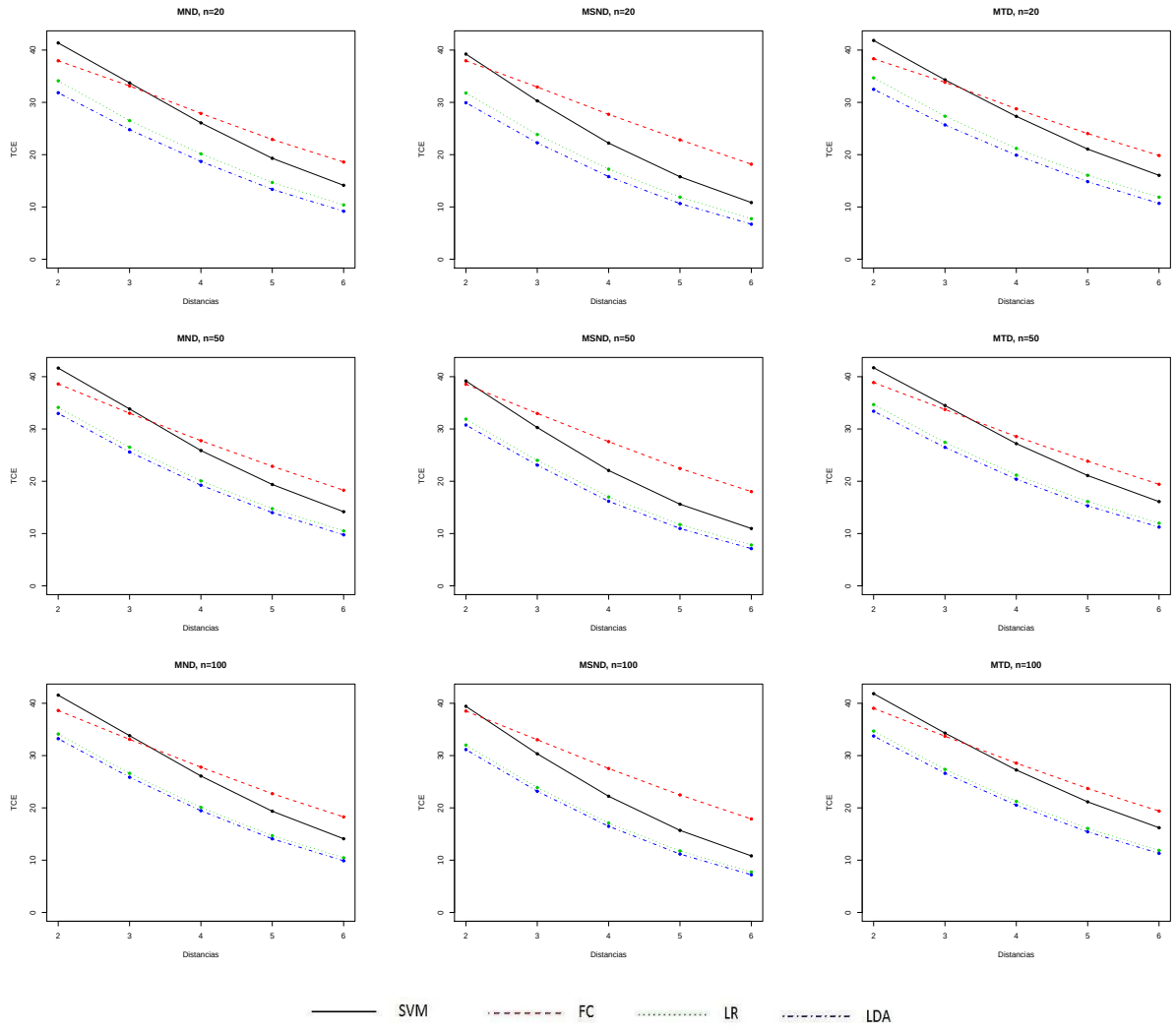

Figura 5-2.: Se muestran los gráficos para el escenario 1, con dos variables, con diferentes distancias entre medias y matriz arbitraria. De arriba hacia abajo se cuentan los tamaños de muestra 20, 50 y 100, respectivamente, y de izquierda a derecha se observan las distribuciones MND, MSND y MTD, respectivamente.

portamiento de los diferentes clasificadores, el clasificador que mejor se desempeña es LDA, seguido de LR, SVM y por último está FC.

5.3. Escenario 3: Cuatro variables

De la **Figura 5-5** para el caso donde se hace la simulación con la matriz previa se observa que en las distribuciones MND y MTD se presenta el mismo comportamiento de TCE para los clasificadores en los diferentes casos, es decir, para una distancia muy pequeña entre

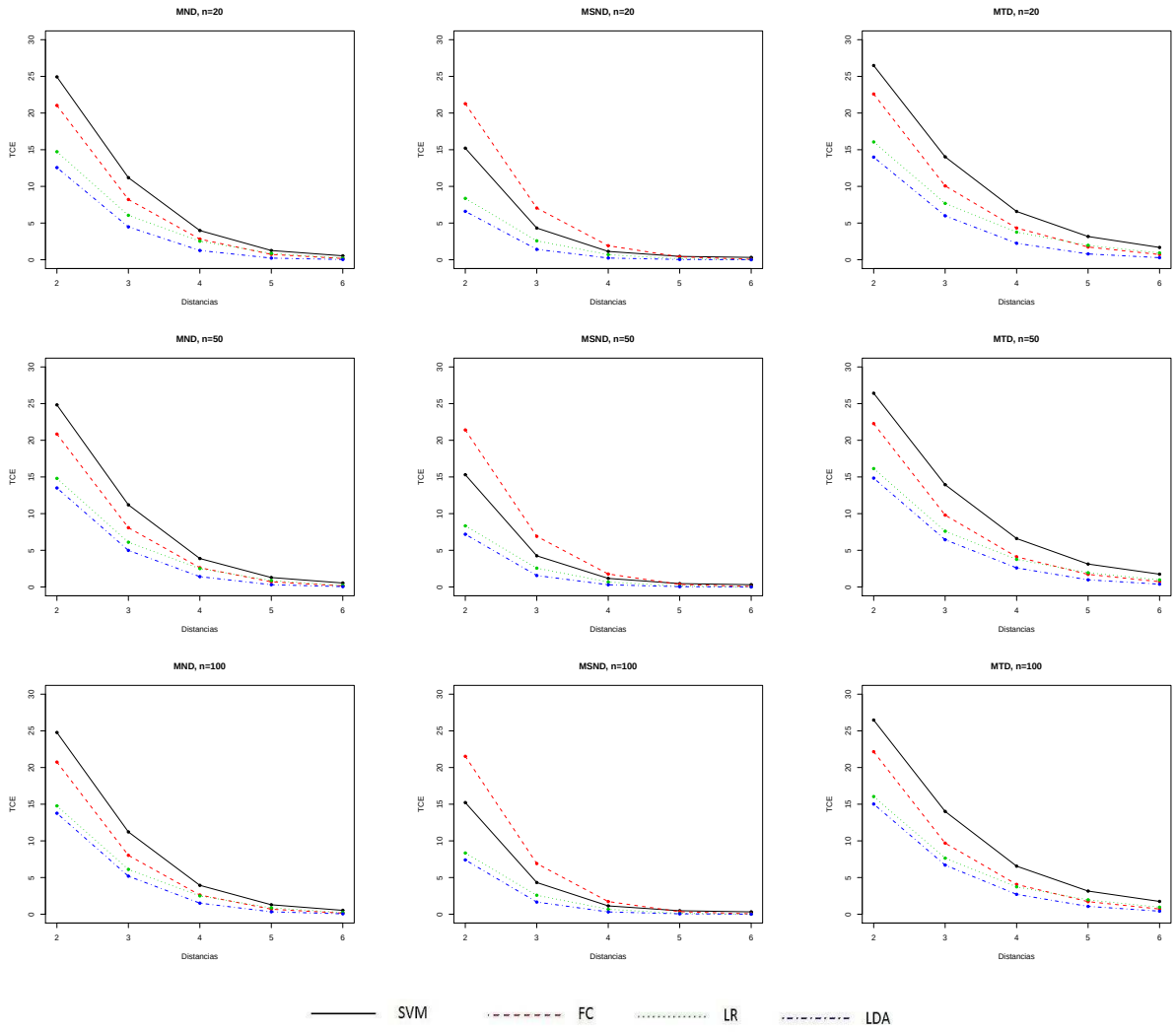

Figura 5-3.: TCE para el escenario 2, con tres variables, con diferentes distancias entre medias y matriz previa. De arriba hacia abajo se cuentan los tamaños de muestra 20, 50 y 100, respectivamente, y de izquierda a derecha se observan las distribuciones MND, MSND y MTD, respectivamente.

medias de los grupos de datos, SVM y FC presentan igual comportamiento y a medida que aumenta dicha distancia FC presenta mejor eficiencia que SVM y LR.
Para el caso de la distribución MSND se observa que a menor distancia entre medias es menor la eficiencia de FC respecto a SVM pero cuando la distancia entre las medias supera las 4 unidades, FC supera en eficiencia a SVM. Se presenta constante la eficiencia de LDA como la mejor.

De la **Figura 5-6** para la matriz arbitraria se observa que en los cinco casos considerados y para los diferentes tamaños de muestra se observa que es constante el comportamiento de

los diferentes clasificadores, el clasificador que mejor se comporta es LDA, seguido de LR, SVM y por último está FC.

Se puede observar que en los casos considerados para cada escenario se presentó una TCE representativa para cada clasificador (SVM, FC, LR y LDA) en cada una de las distribuciones consideradas (MND, MSND y MTD) haciendo competitivo cada clasificador de acuerdo a la situación y logrando el objetivo primordial de este trabajo que es comparar que tan eficientes son los métodos de discriminación para cada una de las distribuciones multivariadas elegidas de acuerdo a la TCE.

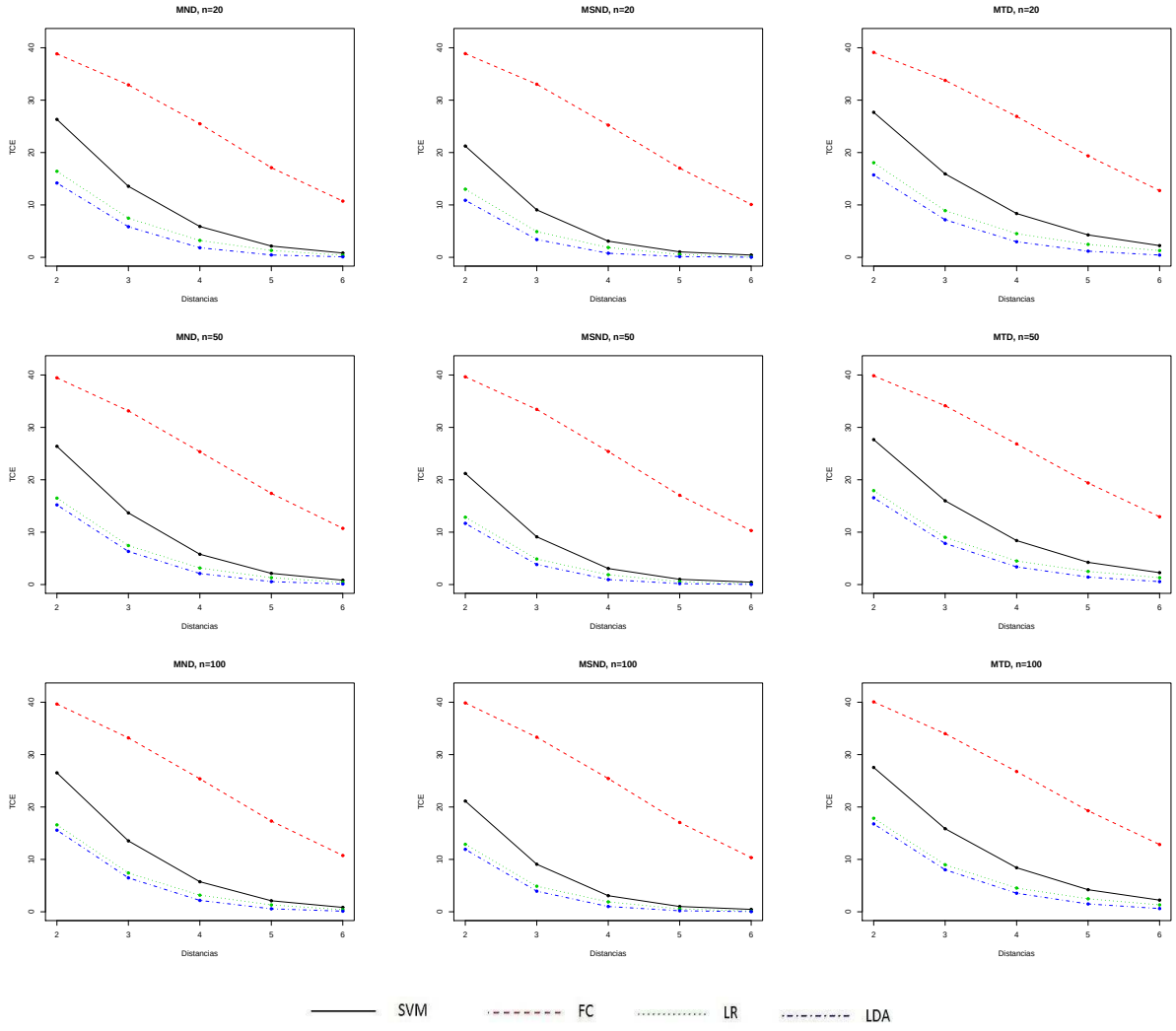

Figura 5-4.: TCE para el escenario 2, con tres variables, con diferentes distancias entre medias y matriz arbitraria. De arriba hacia abajo se cuentan los tamaños de muestra 20, 50 y 100, respectivamente, y de izquierda a derecha se observan las distribuciones MND, MSND y MTD, respectivamente.

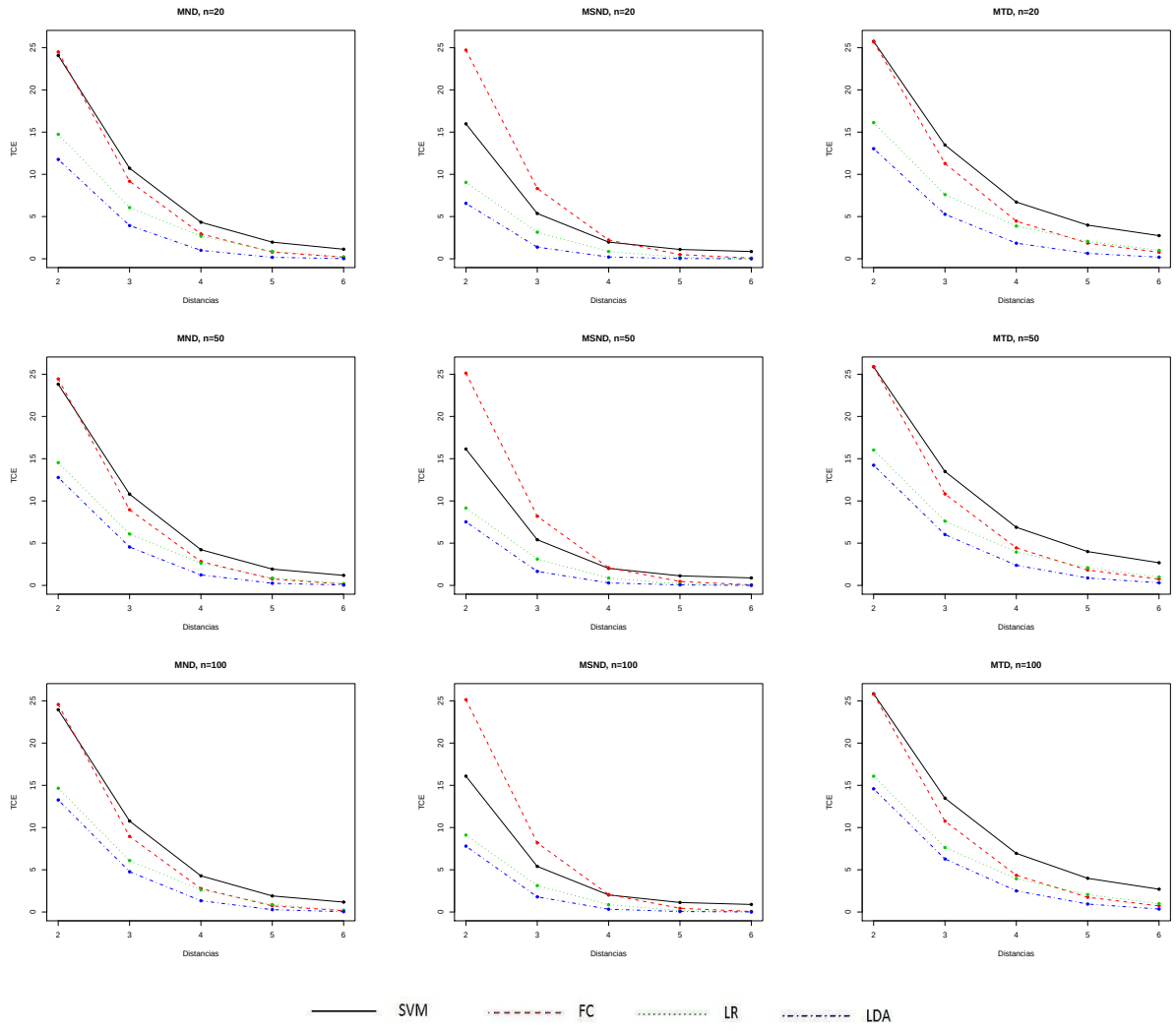

Figura 5-5.: TCE para el escenario 3, con cuatro variables, con diferentes distancias entre medias y matriz previa. De arriba hacia abajo se cuentan los tamaños de muestra 20, 50 y 100, respectivamente, y de izquierda a derecha se observan las distribuciones MND, MSND y MTD, respectivamente.

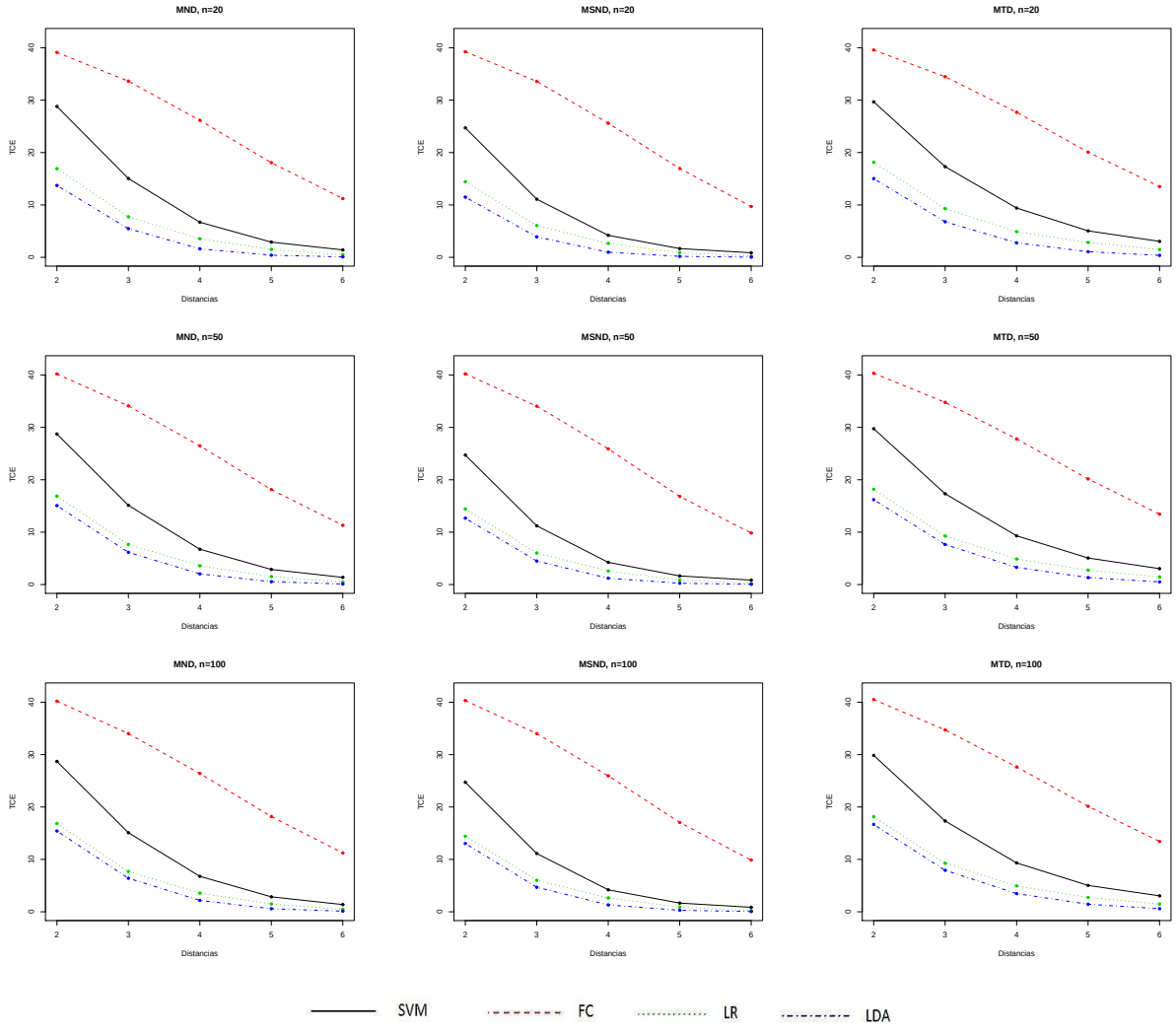

Figura 5-6.: TCE para el escenario 3, con cuatro variables, con diferentes distancias entre medias y varianza arbitraria. De arriba hacia abajo se cuentan los tamaños de muestra 20, 50 y 100, respectivamente, y de izquierda a derecha se observan las distribuciones MND, MSND y MTD, respectivamente.

6. Aplicaciones

En el presente capítulo se aplicaron los clasificadores estudiados en datos que semejaron los escenarios considerados. Para ello se tuverion en cuenta datos con dos variables, tres variables y cuatro variables, cuyos datos poseian una variable categórica con dos clases como indicador de la clasificación.

6.1. Aplicación para escenario 1

Para la aplicación del escenario 1 se tomaron los datos provenientes de un estudio realizado en Alaska y Canadá por el Departamento de Pesca y caza del estado de Alaska, Estados Unidos [Johnson and Wichern, 2007]. Dicho estudio ayuda a identificar cuando un salmón pertenece a aguas de Alaska o a aguas de Canadá. En este estudio se ha identificado que el salmón nacido en aguas dulces de Alaska trae consigo anillos de escamas más pequeños que el salmón nacido en aguas dulces Canadienses.
El estudio toma la muestra del salmón en agua dulce y agua de mar, de Alaska y Canadá en el primer año de vida. La tabla de los datos **Tabla C-1** [Anexo C], muestra los diámetros de los anillos de acuerdo a la región y el agua.

Las variables consideradas son
X_1: la medida del diámetro de los anillos para el primer año de crecimiento en agua dulce, en *centésimas de una pulgada*
X_2: la medida del diámetro de los anillos para el primer año de crecimiento en agua de mar, en *centésimas de una pulgada*

Por medio de la prueba de Shapiro-Wilk multivariada del programa R [1], se determina que los datos no tienen una distribución normal multivariada ya que se tiene un p-valor muy inferior al valor de alfa, esto es p-valor$=2{,}099e-05 < 0{,}05$
Para determinar las TCE de los clasificadores en el grupo de datos, se tomaron muestras de los datos para entrenar los clasificadores SVM, LR y LDA, luego se realizaron 5000 iteraciones de clasificación y se promediaron, los resultados obtenidos son
Además, se puede observar en la **Figura 6-1** que cada marginal no posee una distribución

[1] Para la prueba de Shapiro-Will multivariada se debe cargar la librería `mvnormtest`, cargar el paquete con el mismo nombre y se usa la función `mshapiro.test()`

TCE: Datos de Salmón			
SVM	FC	LR	LDA
11,4840	43,000	51,8484	48,000

Tabla 6-1.: TCE para los datos proveneintes del estudio al Salmón en agua dulce y salada de Alaska y Canadá

normal univariada.

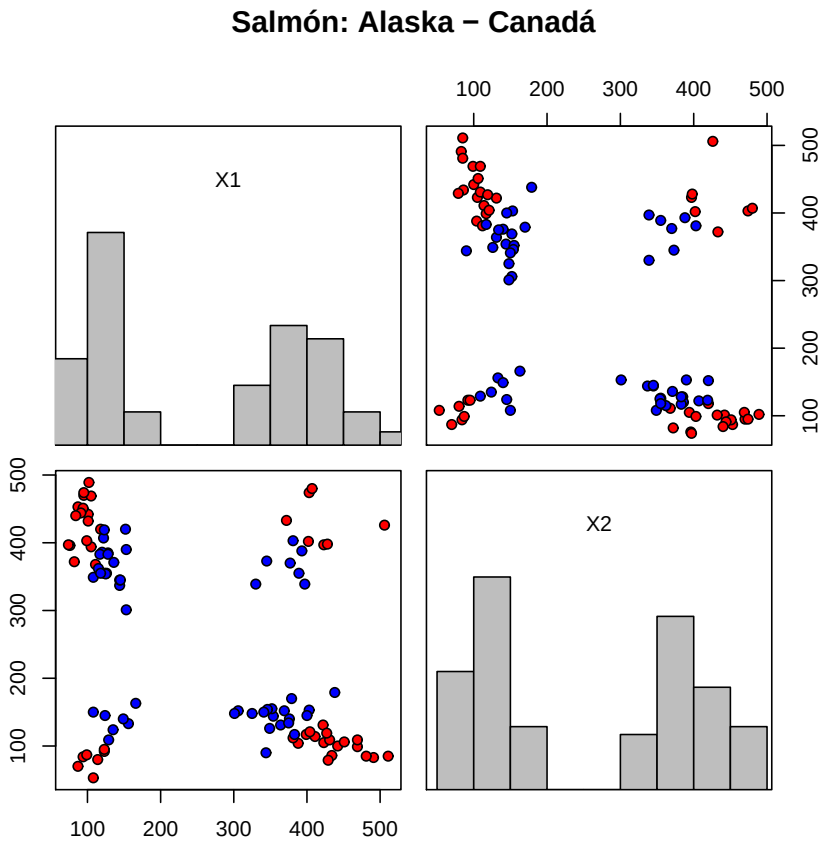

Figura 6-1.: Diagramas de dispersión para las variables X_1 y X_2 para los peces en agua dulce y agua salada de Alaska (azul) y Canadá (rojo). En el panel diagonal se presentan los histogramas marginales.

De acuerdo a lo anterior se puede considerar SVM como el mejor clasificador para este conjunto de datos seguido de FC, de este sigue LDA y por último está el clasificador LR que, de acuerdo al resultado en este conjunto de datos, hay mayor probabilidad de clasificar mal los datos.

6.2. Aplicación para escenario 2

Para la aplicación del escenario 2 se tomaron los datos provenientes de un estudio realizado a la tortuga pintada de Midlan, por el Departamento de Biología de la Universidad de Monterreal, Canadá [Jolicoeur and Mosimann, 1960]. Dicho estudio mide el dimorfismo sexual [2] en la tortuga pintada de Midlan (*Chrysemys picta marginata*) de una pequeña charca estancada del Valle de San Lorenzo, en Coteau Landing, a 35 millas al suroeste de Montreal, Canadá. En el estudio se consideran las variables:
X_1: longitud del caparazón de la tortuga en *milímetro más cercano*
X_2: máximo ancho del caparazón de la tortuga en *milímetro más cercano*
X_3: altura del caparazón de la tortuga en *milímetro más cercano*

Los datos tomados en el estudio se muestran en la **Tabla C-2** [Anexo C]
De acuerdo a la prueba de Shapiro-Will multivariada del programa R, se verifica que los datos obtenidos en el estudio de las tortugas no poseen una distribución normal multivariada. Esto se verifica con el p-valor=0,000993 $< 0{,}05$

De acuerdo a la **Figura 6-2**, se puede verificar que cada marginal no corresponde a una normal univariada.
Para determinar las TCE de los clasificadores en el grupo de datos, se tomaron muestras de los datos para entrenar los clasificadores SVM, LR y LDA, luego se realizaron 5000 iteraciones de clasificación y se promediaron, los resultados obtenidos son presentados en la **Tabla 6-2**

TCE: Datos de Tortugas pintadas de Midlan			
SVM	FC	LR	LDA
15,2233	18,7500	87,8904	10,4167

Tabla 6-2.: TCE para los datos provenientes del estudio de las tortugas pintadas de Midlan, en Montreal, Canadá.

Lo que determina que, en orden ascendente por TCE obtenida en cada clasificador, el que mejor clasifico los datos, en este estudio en particular, es LDA seguido de SVM y FC, presentándose que LR en este caso obtiene una TCE muy superior lo que determina que no es apropiado en este caso.

6.3. Aplicación para escenario 3

Para la aplicación del escenario 3 se tomaron los datos provenientes de un estudio donde se consideran dos especies de pequeños escarabajos (*Haltica oleracea*, *Haltica carduorum*)

[2] Conjunto de diferencias morfológicas y fisiológicas que caracterizan y diferencian a los dos sexos de una misma especie.

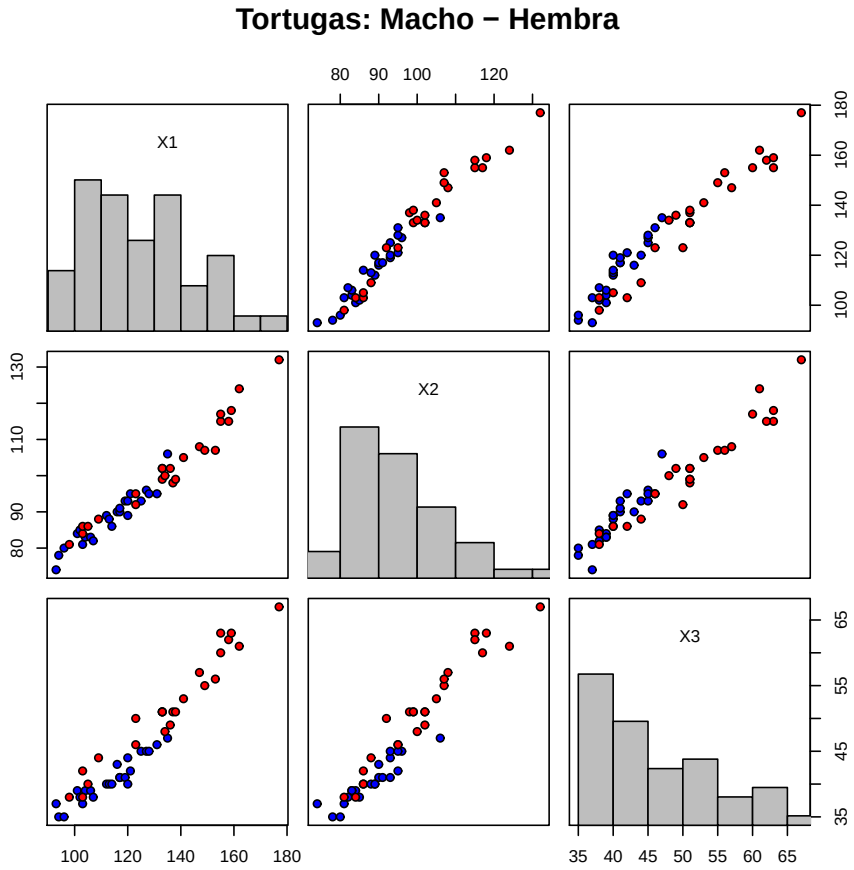

Figura 6-2.: Diagramas de dispersión para las variables X_1, X_2 y X_3 para el estudio de dimorfismo sexual de la tortuga pintada de Midlan. Macho (rojo) y Hembra (azul). En el panel diagonal se presentan los histogramas marginales.

comunes en Rusia, Alemania y Francia [Lubischew, 1962]. Dicho estudio mide 21 variables de las cuales se escogen las cuatro variables más relevantes, de acuerdo al coeficiente discriminante, para estudiar la pertenencia a cada especie.
Las variables más relevantes en este estudio son
X_5: distancia de la ranura transversal desde el borde posterior del protórax en *micras*
X_{14}: longitud de la elytra en 0,01 *mm*
X_{17}: longitud de segundo conjunto antenal en *micras*
X_{18}: longitud del tercer conjunto antenal en *micras*
Los datos usados en esta aplicación son mostrados en la Tabla **C-3** [Anexo C]

De acuerdo a la prueba de Shapiro-Will multivariada se evidencia que los datos no poseen una distribución normal multivariada, esto se ve en los resultados de la prueba con un

p-valor$= 0{,}03226 < 0{,}05$.

Puede observarse en la **Figura 6-3** que la distribuciones de las marginales de los datos en la matriz de dispersión difieren un poco de la distribución normal univariada.

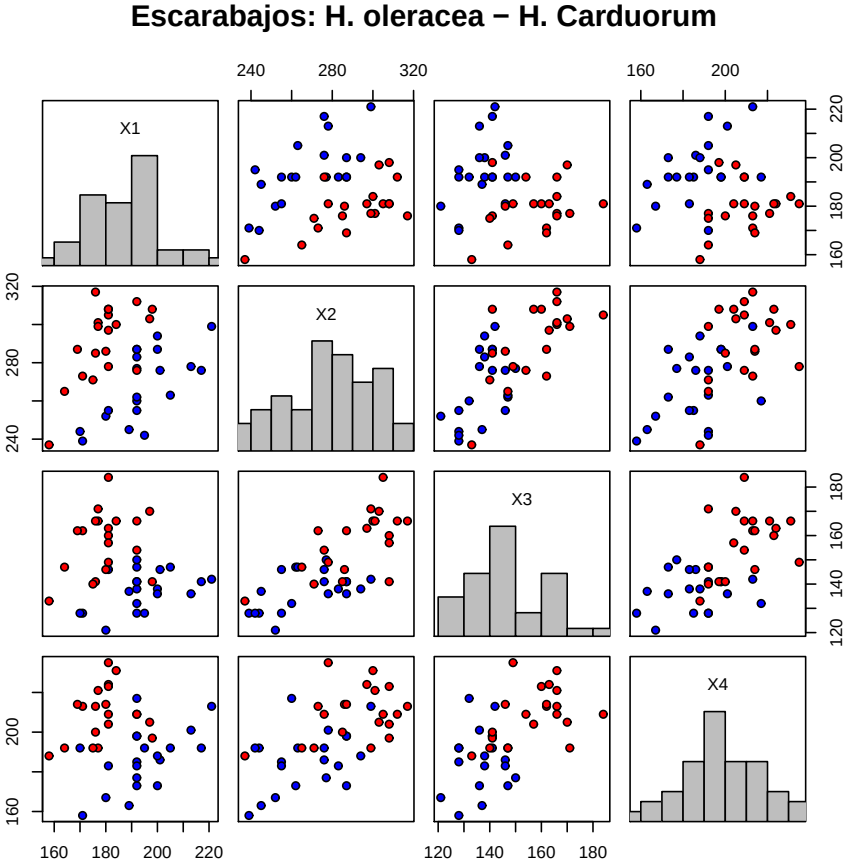

Figura 6-3.: Diagramas de dispersión para las variables X_1, X_2 , X_3 y X_4 para estudiar la pertenencia a las especies *Haltica oleracea* (rojo) y *Haltica carduorum* (azul). En el panel diagonal se presentan los histogramas marginales.

Para determinar las TCE de los clasificadores en el grupo de datos, se tomaron muestras de los datos para entrenar los clasificadores SVM, LR y LDA, luego se realizaron 5000 iteraciones de clasificación y se promediaron, los resultados obtenidos son
Para los datos de este estudio, en particular, se observa que SVM se comporta mejor clasificando que FC y LDA, LR presenta una eficiencia muy baja lo cual lo hace menos apropiado para clasificar en este estudio.

De acuerdo a los resultados obtenidos en las aplicaciones para los tres escenarios, se evidencia

TCE: Datos de Escarabajos Haltica			
SVM	FC	LR	LDA
12,4390	17,5000	89,1600	25,0000

Tabla 6-3.: TCE para los datos proveneintes del estudio de los escarabajos *Haltica oleracea* y *Haltica carduorum*

que los datos reales no se comportan normalmente sin importar el número de variables, lo que hace apropiado el acercamiento a distribuciones distintas de la MND.

7. Conclusiones y recomendaciones

7.1. Conclusiones

1. Se observa que los clasificadores estudiados tienen desempeños diferentes de acuerdo a la varianza de los datos ya que se ha percibido que:

   - En el escenario 1, en la matriz previa los valores de TCE son más cercanos a cero cuando aumenta la distancia entre los vectores de medias, mientras que para la matriz arbitraria al aumentar las distancias entre los vectores de medias el TCE disminuye hasta quedar entre un 10 % y 25 %.

   - El escenario 2 y escenario 3, presentan igual conducta, notando que los clasificadores tienen un comportamiento descendente en las TCE a medida que aumenta la distancia entre las medias, el clasificador FC tiene inferior desempeño con respecto a SVM para el caso de la matriz arbitraria, mientras que para la matriz previa el comportamiento es superior a SVM.

2. Para los diferentes casos considerados en los tres escenarios se observa que hay una constante en el comportamiento de los clasificadores LR y LDA, siendo este segundo más eficiente que el primero conservando una diferencia entre 1 % y 2 %. Estos dos clasificadores presentan superioridad con respecto a FC y SVM.

3. Aunque los clasificadores LDA y LR muestran mejor desempeño en todos los casos considerados para los diferentes escenarios y SVM muestra mejor desempeño en la distribución MSND, con respecto a FC, se encuentra que este desempeño es comparable cuando se cuenta con un conjunto de datos de entrenamiento, cuando esto no sucede (conjunto de entrenamiento) solo se puede contar con el clasificador FC lo que lo hace más ventajoso que los otros clasificadores considerados en este estudio.

4. De acuerdo a los escenarios considerados en las aplicaciones se observa que LR no fue muy conveniente para la clasificación de los datos ya que obtuvo una TCE superior al 50 %, mientras que en los otros clasificadores es diferenciable su desempeño.

7.2. Recomendaciones

1. Cuando un investigador se encuentre con un conjunto de datos en el cual deba clasificar en dos conjuntos mutuamente excluyentes y no cuente con un conjunto de entrenamiento previo a cada conjunto excluyente, el clasificador que en este caso debe usar es FC ya que no necesita conjunto de entrenamiento.

2. En futuras investigaciones se puede considerar un estudio exhaustivo acerca del comportamiento de los clasificadores de acuerdo a la matriz de varianza-covarianza de los datos en estudio para considerar si hay relación con el número de clases en las cuales se deba hacer la clasificación.

A. Anexo: Tablas de resultados con la matriz previa

Las tablas a continuación presentan las FDR calculados a los clasificadores estadísticos SVM, FC, LR y LDA en el proceso de simulación en escenarios donde se tienen en cuenta datos con 2, 3 y 4 variables provenientes de las distribuciones multivariadas MND, MSND y MTD y con parámetros establecidos prebiamente y donde permanece como consante la variabilidad entre los datos (se usa la matriz previa).
Lo terminos D1, D2, D3, D4 y D5 corresponde a las diferentes distancias entre los vectores de medias de los dos grupos considerados en cada caso de los escenarios y los respectivos valores son 2, 3, 4, 5 y 6.

		SVM	FC	LR	LDA
MND	D1	20,079	18,435	15,164	13,879
	D2	8,9950	7,5825	6,3530	5,2800
	D3	3,4930	2,5810	2,3585	1,5605
	D4	1,0745	0,7090	0,8420	0,3305
	D5	0,2770	0,1575	0,2235	0,0645
MSND	D1	15,0230	18,0740	11,2115	10,0160
	D2	5,4725	6,6660	3,7125	2,8415
	D3	1,6095	1,9875	1,1870	0,6170
	D4	0,381	0,482	0,293	0,103
	D5	0,1165	0,0910	0,0430	0,0125
MTD	D1	22,0245	19,9455	16,5220	15,2820
	D2	11,1840	9,2870	7,8005	6,7820
	D3	5,6855	4,0460	3,7140	2,7915
	D4	2,6235	1,6760	1,7875	1,0280
	D5	1,2840	0,7265	0,9105	0,3965

Tabla A-1.: TCE para datos de dos variables y tamaño de muestra 20

		SVM	FC	LR	LDA
MND	D1	20,1274	17,9484	15,1128	14,3226
	D2	8,8966	7,3166	6,2388	5,5744
	D3	3,4902	2,4624	2,4032	1,7140
	D4	1,0534	0,6696	0,8440	0,4108
	D5	0,2884	0,1486	0,2244	0,0696
MSND	D1	15,0968	17,9792	11,2146	10,4674
	D2	5,5404	6,4450	3,7550	3,0746
	D3	1,5608	1,8194	1,2022	0,6428
	D4	0,3738	0,4338	0,2920	0,1108
	D5	0,1122	0,0730	0,0510	0,0148
MTD	D1	21,9566	19,3858	16,4854	15,6854
	D2	11,1340	8,9340	7,7882	7,0844
	D3	5,5796	3,8912	3,6334	2,9464
	D4	2,6836	1,6424	1,8136	1,1990
	D5	1,2780	0,7060	0,9102	0,4792

Tabla A-2.: TCE para datos de dos variables y tamaño de muestra 50

		SVM	FC	LR	LDA
MND	D1	20,1783	17,9231	15,1273	14,5297
	D2	8,9752	7,3165	6,2629	5,6985
	D3	3,4013	2,4134	2,3466	1,7550
	D4	1,0486	0,6456	0,8203	0,4055
	D5	0,2915	0,1370	0,2181	0,0766
MSND	D1	15,1057	18,0197	11,1976	10,6089
	D2	5,5710	6,3992	3,7365	3,1474
	D3	1,5956	1,8349	1,1844	0,6937
	D4	0,3684	0,4189	0,2877	0,1117
	D5	0,1219	0,0740	0,0479	0,0145
MTD	D1	21,8227	19,3089	16,5112	15,8854
	D2	11,1444	8,9236	7,8442	7,2279
	D3	5,5844	3,8767	3,6493	3,0367
	D4	2,7062	1,6451	1,8627	1,2413
	D5	1,2738	0,6807	0,8941	0,4993

Tabla A-3.: TCE para datos de dos variables y tamaño de muestra 100

		SVM	FC	LR	LDA
MND	D1	24,9315	21,0350	14,7265	12,5620
	D2	11,1935	8,2145	6,0550	4,4805
	D3	3,9815	2,8150	2,5305	1,2625
	D4	1,2570	0,7270	0,8645	0,2220
	D5	0,5355	0,1715	0,2080	0,0320
MSND	D1	15,2130	21,2715	8,3610	6,5955
	D2	4,3220	7,0495	2,5810	1,4140
	D3	1,1330	1,9115	0,6920	0,2325
	D4	0,4555	0,4005	0,1295	0,0345
	D5	0,3180	0,0640	0,0220	0,0055
MTD	D1	26,4795	22,5905	16,0610	13,9875
	D2	14,0295	10,0645	7,6820	5,9885
	D3	6,5835	4,3240	3,7625	2,2460
	D4	3,168	1,735	1,966	0,792
	D5	1,6755	0,7095	0,9325	0,2825

Tabla A-4.: TCE para datos de tres variables y tamaño de muestra 20

		SVM	FC	LR	LDA
MND	D1	24,8266	20,8452	14,8042	13,4994
	D2	11,1986	8,0846	6,1038	4,9872
	D3	3,8800	2,6238	2,5088	1,4088
	D4	1,2890	0,7044	0,8768	0,3062
	D5	0,5430	0,1494	0,2194	0,0490
MSND	D1	15,3070	21,4084	8,3322	7,1862
	D2	4,2482	6,9060	2,5686	1,5608
	D3	1,1700	1,7690	0,6756	0,3004
	D4	0,4608	0,3536	0,1206	0,0442
	D5	0,3134	0,0608	0,0172	0,0048
MTD	D1	26,4140	22,2752	16,1380	14,8408
	D2	13,9442	9,7958	7,6004	6,4420
	D3	6,6060	4,1144	3,7628	2,6024
	D4	3,1270	1,7178	1,9418	0,9742
	D5	1,7444	0,7200	0,9620	0,3836

Tabla A-5.: TCE para datos de tres variables y tamaño de muestra 50

		SVM	FC	LR	LDA
MND	D1	24,7795	20,7427	14,7708	13,7664
	D2	11,2212	8,0299	6,1119	5,1890
	D3	3,9368	2,6172	2,5007	1,5105
	D4	1,2874	0,6836	0,8691	0,3307
	D5	0,5275	0,1459	0,2285	0,0548
MSND	D1	15,2221	21,5278	8,3441	7,4012
	D2	4,3216	6,9083	2,5867	1,6577
	D3	1,1351	1,7264	0,6730	0,3111
	D4	0,4756	0,3457	0,1174	0,0535
	D5	0,3216	0,0531	0,0187	0,0071
MTD	D1	26,4693	22,1564	16,0423	15,0206
	D2	14,0146	9,6894	7,6410	6,6965
	D3	6,5530	4,0622	3,7390	2,7092
	D4	3,1558	1,7143	1,9602	1,0681
	D5	1,7429	0,7095	0,9553	0,4163

Tabla A-6.: TCE para datos de tres variables y tamaño de muestra 100

		SVM	FC	LR	LDA
MND	D1	24,0760	24,4910	14,7365	11,7615
	D2	10,7210	9,1750	6,0630	3,345
	D3	4,3270	2,9480	2,6830	1,0015
	D4	1,9730	0,7960	0,8855	0,1730
	D5	1,1350	0,1895	0,2335	0,0215
MSND	D1	15,9805	24,7085	9,0485	6,5650
	D2	5,3640	8,3260	3,1505	1,3850
	D3	1,9785	2,2290	0,8620	0,2130
	D4	1,1065	0,4980	0,1705	0,0350
	D5	0,8605	0,0835	0,0295	0,0020
MTD	D1	25,7465	25,6810	16,1195	13,0370
	D2	13,4720	11,2890	7,6005	5,2695
	D3	6,7185	4,4585	3,8725	1,8505
	D4	3,9875	1,8560	2,0675	0,6445
	D5	2,7430	0,7795	1,0010	0,1920

Tabla A-7.: TCE para datos de cuatro variables y tamaño de muestra 20

		SVM	FC	LR	LDA
MND	D1	23,8172	24,4402	14,5366	12,7764
	D2	10,8030	8,9586	6,0838	4,5452
	D3	4,2286	2,8234	2,6370	1,2432
	D4	1,9286	0,7600	0,8754	0,2546
	D5	1,1882	0,1548	0,2072	0,0362
MSND	D1	16,1442	25,1298	9,1520	7,5222
	D2	5,4064	8,1862	3,1186	1,6570
	D3	2,0152	2,0920	0,8692	0,2940
	D4	1,1334	0,4652	0,1804	0,0558
	D5	0,8874	0,0664	0,0282	0,0066
MTD	D1	25,8752	25,9002	16,0256	14,2242
	D2	13,4882	10,8236	7,6100	6,0208
	D3	6,8780	4,4424	3,9474	2,3794
	D4	4,0058	1,8178	2,0960	0,8806
	D5	2,6832	0,7442	0,9836	0,3052

Tabla A-8.: TCE para datos de cuatro variables y tamaño de muestra 50

		SVM	FC	LR	LDA
MND	D1	23,9492	24,5697	14,6550	13,2641
	D2	10,7729	8,9495	6,0914	4,7707
	D3	4,2849	2,7771	2,6210	1,3359
	D4	1,9033	0,7246	0,8857	0,2776
	D5	1,1741	0,1503	0,2189	0,0407
MSND	D1	16,0846	25,1375	9,1070	7,7995
	D2	5,3984	8,1875	3,1283	1,7937
	D3	2,0190	2,0705	0,8663	0,3341
	D4	1,1280	0,4422	0,1737	0,0577
	D5	0,8945	0,0690	0,0268	0,0066
MTD	D1	25,8418	25,7891	16,0764	14,5832
	D2	13,4752	10,7565	7,6313	6,2851
	D3	6,9418	4,3458	3,9415	2,5029
	D4	3,9907	1,7484	2,0626	0,9422
	D5	2,7007	0,7249	0,9879	0,3523

Tabla A-9.: TCE para datos de cuatro variables y tamaño de muestra 100

B. Anexo: Tabla de resultados con la matriz arbitraria

Las tablas a continuación presentan las FDR calculados a los clasificadores estadísticos SVM, FC, LR y LDA en el proceso de simulación en escenarios donde se tienen en cuenta datos con 2, 3 y 4 variables provenientes de las distribuciones multivariadas MND, MSND y MTD y con parámetros establecidos prebiamente y donde permanece como consante la variabilidad entre los datos (se usa la matriz arbitraria).
Lo terminos D1, D2, D3, D4 y D5 corresponde a las diferentes distancias entre los vectores de medias de los dos grupos considerados en cada caso de los escenarios y los respectivos valores son 2, 3, 4, 5 y 6.

		SVM	FC	LR	LDA
MND	D1	41,3610	37,9550	34,1205	31,8455
	D2	33,7290	33,1320	26,5110	24,7775
	D3	26,078	27,882	20,154	18,702
	D4	19,3020	22,9035	14,6790	13,3565
	D5	14,1415	18,5870	10,3845	9,1755
MSND	D1	39,2365	37,9560	31,7870	29,9320
	D2	30,2875	32,9335	23,8495	22,2670
	D3	22,2120	27,7160	17,2275	15,7990
	D4	15,7735	22,8330	11,8800	10,6500
	D5	10,8380	18,1725	7,7515	6,6980
MTD	D1	41,8415	38,3335	34,6820	32,4885
	D2	34,303	33,840	27,353	25,689
	D3	27,3270	28,7600	21,1935	19,9005
	D4	21,0605	24,0325	16,0560	14,8410
	D5	16,0470	19,8395	11,8725	10,6845

Tabla B-1.: TCE para datos de dos variables y tamaño de muestra 20

		SVM	FC	LR	LDA
MND	D1	41,6466	38,6090	34,1176	32,9644
	D2	33,8258	33,0050	26,5192	25,5926
	D3	25,8726	27,7568	20,0706	19,2418
	D4	19,3806	22,8600	14,7634	14,0044
	D5	14,1758	18,2842	10,5010	9,7532
MSND	D1	39,1782	38,5552	31,8758	30,7582
	D2	30,2760	32,9698	23,9978	23,0818
	D3	22,0686	27,5854	16,9986	16,1778
	D4	15,6136	22,4484	11,7204	10,9744
	D5	10,9628	18,0224	7,8172	7,1196
MTD	D1	41,7052	38,8736	34,6692	33,3922
	D2	34,4924	33,7318	27,4408	26,4796
	D3	27,1954	28,5472	21,1710	20,3916
	D4	21,0960	23,8294	16,1190	15,2982
	D5	16,1110	19,4238	11,9764	11,2356

Tabla B-2.: TCE para datos de dos variables y tamaño de muestra 50

		SVM	FC	LR	LDA
MND	D1	41,5540	38,6178	34,1229	33,2333
	D2	33,8096	33,1090	26,6035	25,8573
	D3	26,0888	27,7798	20,0922	19,4575
	D4	19,3548	22,7141	14,6753	14,0797
	D5	14,0968	18,2712	10,4531	9,8832
MSND	D1	39,4210	38,5090	32,0009	31,1496
	D2	30,3461	33,0264	23,8794	23,1730
	D3	22,2203	27,5318	17,1018	16,4760
	D4	15,6959	22,4741	11,7604	11,1711
	D5	10,8189	17,8785	7,7375	7,2013
MTD	D1	41,8471	39,0411	34,6854	33,7403
	D2	34,3024	33,7041	27,3592	26,6112
	D3	27,2402	28,5509	21,2455	20,5162
	D4	21,1519	23,7124	16,0580	15,4169
	D5	16,2035	19,3876	11,8724	11,2963

Tabla B-3.: TCE para datos de dos variables y tamaño de muestra 100

		SVM	FC	LR	LDA
MND	D1	26,3235	38,8400	16,4160	14,1860
	D2	13,5620	32,9150	7,4755	5,8380
	D3	5,8835	25,4895	3,2190	1,8295
	D4	2,1505	17,1160	1,3160	0,4500
	D5	0,8260	10,7310	0,3915	0,0955
MSND	D1	21,2415	38,8815	13,0020	10,8855
	D2	9,0595	33,0405	4,9070	3,3975
	D3	3,0870	25,2285	1,8750	0,7760
	D4	1,0435	17,0250	0,5620	0,1400
	D5	0,4345	10,0850	0,1190	0,0200
MTD	D1	27,6785	39,1105	18,0550	15,7205
	D2	15,9265	33,7700	8,9100	7,1650
	D3	8,3610	26,9035	4,5190	2,9735
	D4	4,2775	19,3605	2,4665	1,1750
	D5	2,2365	12,7405	1,2940	0,4260

Tabla B-4.: TCE para datos de tres variables y tamaño de muestra 20

		SVM	FC	LR	LDA
MND	D1	26,3832	39,4604	16,4972	15,2064
	D2	13,6818	33,1518	7,4442	6,3258
	D3	5,7848	25,3572	3,1632	2,1036
	D4	2,1380	17,4054	1,3110	0,5560
	D5	0,8312	10,7278	0,4230	0,1062
MSND	D1	21,2056	39,6542	12,8496	11,6806
	D2	9,1108	33,4354	4,8796	3,8234
	D3	3,0640	25,4018	1,8748	0,9526
	D4	1,0028	17,0372	0,5498	0,1762
	D5	0,4532	10,3068	0,1266	0,0246
MTD	D1	27,6636	39,8602	17,9286	16,5638
	D2	15,9890	34,1330	9,0126	7,8560
	D3	8,3970	26,8454	4,4976	3,3752
	D4	4,2388	19,3926	2,5056	1,4288
	D5	2,2688	12,9422	1,3104	0,5618

Tabla B-5.: TCE para datos de tres variables y tamaño de muestra 50

		SVM	FC	LR	LDA
MND	D1	26,5212	39,6513	16,5865	15,5764
	D2	13,5361	33,1972	7,4105	6,4806
	D3	5,7367	25,3820	3,1752	2,1732
	D4	2,1055	17,3168	1,3091	0,5692
	D5	0,8388	10,7430	0,4006	0,1162
MSND	D1	21,1302	39,8615	12,8740	11,9169
	D2	9,0729	33,3219	4,8795	3,9317
	D3	3,0714	25,4332	1,8927	1,0076
	D4	1,0089	17,0478	0,5582	0,1958
	D5	0,4618	10,3443	0,1265	0,0319
MTD	D1	27,5503	40,0627	17,8492	16,7743
	D2	15,8591	34,0000	8,9831	8,0243
	D3	8,4094	26,7743	4,5339	3,5460
	D4	4,2138	19,2877	2,4772	1,4960
	D5	2,2283	12,8423	1,3247	0,6184

Tabla B-6.: TCE para datos de tres variables y tamaño de muestra 100

		SVM	FC	LR	LDA
MND	D1	28,8025	39,1100	16,9260	13,7180
	D2	15,0475	33,6260	7,7325	5,4725
	D3	6,7000	26,1565	3,5510	1,6210
	D4	2,9040	18,0645	1,5030	0,4120
	D5	1,4185	11,2195	0,4790	0,0835
MSND	D1	24,7055	39,2555	14,4180	11,4875
	D2	11,0955	33,5885	6,0570	3,8980
	D3	4,2010	25,5910	2,6515	0,9770
	D4	1,6690	16,9550	0,9025	0,1870
	D5	0,8695	9,6985	0,2275	0,0280
MTD	D1	29,6805	39,5675	18,1320	15,0220
	D2	17,3065	34,4740	9,2925	6,7890
	D3	9,4090	27,6880	4,8975	2,7645
	D4	5,0540	20,0865	2,8355	1,0745
	D5	3,0385	13,5085	1,4855	0,3725

Tabla B-7.: TCE para datos de cuatro variables y tamaño de muestra 20

		SVM	FC	LR	LDA
MND	D1	28,7308	40,1916	16,8858	15,0638
	D2	15,1292	34,0980	7,6596	6,1594
	D3	6,7448	26,4780	3,5872	2,0418
	D4	2,8832	18,1250	1,5154	0,5462
	D5	1,3718	11,3132	0,4732	0,1084
MSND	D1	24,7106	40,2210	14,4298	12,6656
	D2	11,2098	34,0482	6,0204	4,4914
	D3	4,2390	25,9024	2,5796	1,2122
	D4	1,6438	16,8174	0,8784	0,2510
	D5	0,8510	9,8310	0,2248	0,0462
MTD	D1	29,7500	40,3268	18,1868	16,2016
	D2	17,3318	34,7750	9,2544	7,6394
	D3	9,3066	27,8012	4,8820	3,2852
	D4	5,0680	20,1622	2,7468	1,3278
	D5	3,0374	13,4332	1,4582	0,5138

Tabla B-8.: TCE para datos de cuatro variables y tamaño de muestra 50

		SVM	FC	LR	LDA
MND	D1	28,6976	40,1997	16,8362	15,4205
	D2	15,0957	34,0215	7,6870	6,4014
	D3	6,7713	26,3989	3,5636	2,1729
	D4	2,8555	18,1747	1,4848	0,5802
	D5	1,3791	11,2252	0,4790	0,1201
MSND	D1	24,7162	40,3218	14,4226	13,0359
	D2	11,1250	34,0081	6,0113	4,6843
	D3	4,1945	25,9516	2,6478	1,3088
	D4	1,6631	17,0513	0,9100	0,2897
	D5	0,8545	9,8647	0,2382	0,0488
MTD	D1	29,8536	40,5036	18,1466	16,6762
	D2	17,3286	34,7146	9,2546	7,9199
	D3	9,3233	27,6491	4,9365	3,4775
	D4	5,0344	20,1332	2,7445	1,4472
	D5	3,0465	13,4219	1,4747	0,5860

Tabla B-9.: TCE para datos de cuatro variables y tamaño de muestra 100

C. Anexo: Tablas de datos para aplicación

Las tablas presentadas a continuación muestran los datos utilizados en el capítulo de aplicaciones para los diferentes escenarios considerados en la metodología del estudio.

Alaska		Canadá	
X_1	X_2	X_1	X_2
108	368	129	420
131	355	148	371
105	469	179	407
86	506	152	381
99	402	166	377
87	423	124	389
94	440	156	419
117	489	131	345
79	432	140	362
99	403	144	345
114	428	149	393
123	372	108	330
123	372	135	355
109	420	170	386
112	394	152	301
104	407	153	397
111	422	152	301
126	423	136	438
105	434	122	306
119	474	148	383
114	396	90	385
100	470	145	337

Continua en la página siguiente

Continuación de la tabla			
Alaska		Canadá	
X_1	X_2	X_1	X_2
84	399	123	364
102	429	145	376
101	469	115	354
85	444	134	383
109	397	117	355
106	442	126	345
82	431	118	379
118	381	120	369
105	388	153	403
121	403	150	354
85	451	154	390
83	453	155	349
53	427	109	325
95	411	117	344
76	442	128	400
95	426	144	403
87	402	163	370
70	397	145	355
84	511	133	375
91	469	128	383
74	451	123	349
101	474	144	373
80	398	140	388
95	433	150	339
92	404	124	341
99	481	125	346
94	491	153	352
87	480	108	339

Tabla C-1.: Datos del salmón proveniente del estado de Alaska y Canadá.

Macho			Hembra		
X_1	X_2	X_3	X_1	X_2	X_3
93	74	37	98	81	38
94	78	35	103	84	38
96	80	35	103	86	42
101	84	39	105	86	40
102	85	38	109	88	44
103	81	37	123	92	50
104	83	39	123	95	46
106	83	39	133	99	51
107	82	38	133	102	51
112	89	40	133	102	51
113	88	40	134	100	48
114	86	40	136	102	49
116	90	43	137	98	51
117	90	41	138	99	51
117	91	41	141	105	53
119	93	41	147	108	57
120	89	40	149	107	55
120	93	44	153	107	56
121	95	42	155	115	63
125	93	45	155	117	60
127	96	45	158	115	62
128	95	45	159	118	63
131	95	46	162	124	61
135	106	47	177	132	67

Tabla C-2.: Datos de las caparazones de las tortugas Machos y Hembras

Haltica oleracea				Haltica carduorum			
X_5	X_{14}	X_{17}	X_{18}	X_5	X_{14}	X_{17}	X_{18}
189	245	137	163	181	305	184	209
192	260	132	217	158	237	133	188
217	276	141	192	184	300	166	231
221	299	142	213	171	273	162	213
171	239	128	158	181	297	163	224
192	262	147	173	181	308	160	223
213	278	136	201	177	301	166	221
192	255	128	185	198	308	141	197
170	244	128	192	180	286	146	214
201	276	146	186	177	299	171	192
195	242	128	192	176	317	166	213
205	263	147	192	192	312	166	209
180	252	121	167	176	285	141	200
192	283	138	183	169	287	162	214
200	294	138	188	164	265	147	192
192	277	150	177	181	308	157	204
200	287	136	173	192	276	154	209
181	255	146	183	181	278	149	235
192	287	141	198	175	271	140	192
				197	303	170	205

Tabla C-3.: Variables para el estudio de los escarabajos Haltica oleracea y Haltica carduorum.

D. Anexo: Programa en R

Antes de comenzar las rutinas es necesario cargar las librerias `MASS`, `nnet`, `e1071` y `rpart`, además de los paquetes `class`, `mvtnorm`, `rpart`, `e1071`, `mlbench`, `graphics`, `mnormt`, `sn`, `stats`. Los datos de simulación fueron generados a partir de funciones que involucran las distribuciones correspondientes.
La función que genera los datos MND es

```
NormalM<-function(n2,m1,m2,S2){
                                Datos1<-rmnorm(n2,m1,S2)
                                Datos2<-rmnorm(n2,m2,S2)
                                B<-rbind(Datos1,Datos2)
                                Grupo<-as.factor(rep(2:1,c(n2,n2)))
                                B<-B
                                DtM<-data.frame(B,Grupo)
                                retval<-list(matriz=B,tabla=DtM)
                                }
```

La función que genera los datos MSND es

```
NormalS<-function(n2,m1,m2,S2,A2){
                                   Datos1<-rmsn(n2,m1,S2,A2)
                                   Datos2<-rmsn(n2,m2,S2,A2)
                                   B<-rbind(Datos1,Datos2)
                                   Grupo<-as.factor(rep(2:1,c(n2,n2)))
                                   B<-B
                                   DtM<-data.frame(B,Grupo)
                                   retval<-list(matriz=B,tabla=DtM)
                                   }
```

La función que genera los datos MTD es

```
TM<-function(n2,m1,m2,S2,g2){
                              Datos1<-rmt(n2,m1,S2,g2)
                              Datos2<-rmt(n2,m2,S2,g2)
                              B<-rbind(Datos1,Datos2)
                              Grupo<-as.factor(rep(2:1,c(n2,n2)))
```

```
B<-B
DtM<-data.frame(B,Grupo)
retval<-list(matriz=B,tabla=DtM)
}
```

Cada función genera los datos con sus respectivas etiquetas de las dos clases a clasificar, `n2` es el tamaño de muestra que se desea para cada etiqueta, `m1` es el vector de media del primer grupo de datos, `m2` es el vector de media del segundo grupo de datos, `S2` es la matriz de varianza-covarianza de los grupos de datos, `A2` es el parámetro de forma para la distribución MSND y `g2` son los grados de libertad para la distribución MTD.
Para los cuatro clasificados utilizados, a continuación se presentan las funciones utlizadas.
Para SVM se tiene

```
SVM<-function(etabla,tabla,v){
m<-v+1
svm.modelo<-svm(Grupo ~ .,data=etabla,
cost=100,gamma=1)
svm.predic<-predict(svm.modelo,tabla[,-m],
type="class")
cont<-table(pred=svm.predic,true=tabla[,m])
ec<-(cont[1,2]+cont[2,1])/sum(cont)
}
```

Para FC se tiene

```
Fuzzy<-function(matriz,tabla,v){
m<-v+1
cl<-cmeans(matriz,2,verbose=FALSE,
method="cmeans",m=2)$cluster
cont<-table(pred=cl,true=tabla[,m])
ec<-(cont[1,2]+cont[2,1])/sum(cont)
}
```

Para LR se tiene

```
Logis<-function(etabla,tabla,v){
m<-v+1
rl<-glm(Grupo ~ .,data=etabla,
family=binomial)
rrl<- predict(rl,tabla[,-m],
type = "response")
n<-nrow(tabla)
vec<-rep(1:0,c(n/2,n/2))
```

```
                                out<- data.frame(vec, rrl, 1-rrl)
                                oout<- cbind(out, predstatus = ifelse
                                (apply(out[,-1], 1,which.max) == 1, 1, 0))
                                cont<-table(factor(oout[,1], levels = 0:1),
                                factor(oout[,4], levels = 0:1))
                                ec<-(cont[2,1] + cont[1,2])/sum(cont)
                               }
```

Para LDA se tiene

```
Disc<-function(matriz){
                        n<-nrow(matriz)
                        vec<-rep(1:0,c(n/2,n/2))
                        AD<-lda(matriz,vec)
                        pre<-predict(AD)$class
                        cont<-table(pred=pre,true=vec)
                        ec<-(cont[1,2]+cont[2,1])/sum(cont)
                        }
```

Las siguientes funciones generan las TCE para cada distribución considerada. Cada función a continuación corresponden a las distribuciones MND, MSND y MTD, respectivamente.

```
ERRNM<-function(v,n,m1,m2,S1,IT){
                                  ZZ<-matrix(rep(0),ncol=4,nrow=IT)
                                  for(i in 1:IT){
                                                 A<-NormalM(n,m1,m2,S1)
                                                 EA<-NormalM(50,m1,m2,S1)
                                                 a<-SVM(EA$tabla,A$tabla,v)
                                                 b1<-Fuzzy(A$matriz,A$tabla,v)
                                                 b2<-1-b1
                                                 b<-min(b1,b2)
                                                 d<-Logis(EA$tabla,A$tabla,v)
                                                 e<-Disc(A$matriz)
                                                 ZZ[i,]<-c(a,b,d,e)
                                                }
                                  ZZ<-ZZ
                                  colMeans(ZZ)
                                 }

ERRNS<-function(v,n,m1,m2,S,P,IT){
                                   ZZ<-matrix(rep(0),ncol=4,nrow=IT)
                                   for(i in 1:IT){
```

```
                                        A<-NormalS(n,m1,m2,S,P)
                                        EA<-NormalS(50,m1,m2,S,P)
                                        a<-SVM(EA$tabla,A$tabla,v)
                                        b1<-Fuzzy(A$matriz,A$tabla,v)
                                        b2<-1-b1
                                        b<-min(b1,b2)
                                        d<-Logis(EA$tabla,A$tabla,v)
                                        e<-Disc(A$matriz)
                                        ZZ[i,]<-c(a,b,d,e)
                                        }
                                 ZZ<-ZZ
                                 colMeans(ZZ)
                                }

ERRTM<-function(v,n,m1,m2,S,g,IT){
                                   ZZ<-matrix(rep(0),ncol=4,nrow=IT)
                                   for(i in 1:IT){
                                                 A<-TM(n,m1,m2,S,g)
                                                 EA<-TM(50,m1,m2,S,g)
                                                 a<-SVM(EA$tabla,A$tabla,v)
                                                 b1<-Fuzzy(A$matriz,A$tabla,v)
                                                 b2<-1-b1
                                                 b<-min(b1,b2)
                                                 d<-Logis(EA$tabla,A$tabla,v)
                                                 e<-Disc(A$matriz)
                                                 ZZ[i,]<-c(a,b,d,e)
                                                 }
                                   ZZ<-ZZ
                                   colMeans(ZZ)
                                   }
```

`IT` corresponde al número de iteraciones que se deseen hacer en la simulación.

Bibliografía

[Adeli and Hung, 1994] Adeli, H. and Hung, S.-L. (1994). *Machine Learning: Neural Networks, Genetic Algorithms, and Fuzzy Systems*. Wiley.

[Anderer et al., 1994] Anderer, P., Saletu, B., Klöppel, B., Semlitsch, H., and Werner, H. (1994). Discrimination between demented patients and normals based on topographic eeg slow wave activity: comparison between statistics, discriminant analysis and artificial neural network classifiers. *Electroencephalography and clinical neurophysiology*, 91(2):108–117.

[Anderson, 1958] Anderson, T. W. (1958). *An introduction to multivariate statistical analysis*, volume 2. Wiley New York.

[Asparoukhov and Krzanowski, 2001] Asparoukhov, O. K. and Krzanowski, W. J. (2001). A comparison of discriminant procedures for binary variables. *Computational Statistics & Data Analysis*, 38(2):139–160.

[Azzalini and Capitanio, 1999] Azzalini, A. and Capitanio, A. (1999). Statistical applications of the multivariate skew normal distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):579–602.

[Azzalini and Dalla Valle, 1996] Azzalini, A. and Dalla Valle, A. (1996). The multivariate skew-normal distribution. *Biometrika*, 83(4):715–726.

[Bezdek et al., 1984] Bezdek, J. C., Ehrlich, R., and Full, W. (1984). Fcm: The fuzzy k-means clustering algorithm. *Computers & Geosciences*, 10(2):191–203.

[Bull, 2012] Bull, L., editor (2012). *Applications of Learning Classifier Systems (Studies in Fuzziness and Soft Computing)*. Springer.

[Chen et al., 2009] Chen, S.-T., Hsiao, Y.-H., Huang, Y.-L., Kuo, S.-J., Tseng, H.-S., Wu, H.-K., and Chen, D.-R. (2009). Comparative analysis of logistic regression, support vector machine and artificial neural network for the differential diagnosis of benign and malignant solid breast tumors by the use of three-dimensional power doppler imaging. *Korean Journal of Radiology*, 10(5):464–471.

[Contreras et al., 2010] Contreras, J. A., Martinez, L. B., and Puerta, Y. V. (2010). Clasificador difuso para diagnóstico de enfermedades. *Revista Tecnológicas*, (25).

[Correa, 2010] Correa, J. C. (2010). Diagnósticos de regresión usando la fdr (tasa de descubrimientos falsos). *Revista Comunicaciones en Estadística*, 3(2):109–118.

[Cortes and Vapnik, 1995] Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

[Crawley, 2012] Crawley, M. J. (2012). *The R book*. John Wiley & Sons.

[de Caluwe, 1997] de Caluwe, R. (1997). *Fuzzy And Uncertain Object-Oriented Databases: Concepts And Models (Advances in Fuzzy Systems: Application and Theory)*. Wspc.

[Dudoit et al., 2002] Dudoit, S., Fridlyand, J., and Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American statistical association*, 97(457):77–87.

[Fisher, 1936] Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188.

[Fisher, 1938] Fisher, R. A. (1938). The statistical utilization of multiple measurements. *Annals of eugenics*, 8(4):376–386.

[Flury, 1997] Flury, B. (1997). *A first course in multivariate statistics*. Springer.

[Fukunaga and Hayes, 1989] Fukunaga, K. and Hayes, R. R. (1989). Effects of sample size in classifier design. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 11(8):873–885.

[Genton, 2004] Genton, M. G. (2004). *Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality*. Chapman and Hall/CRC.

[Hastie et al., 2009] Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Friedman, J., and Tibshirani, R. (2009). *The elements of statistical learning*, volume 2. Springer.

[Hernández and Correa, 2009] Hernández, F. and Correa, J. C. (2009). Comparación entre tres técnicas de clasificación. *Revista Colombiana de Estadística*, 32(2):247–265.

[Hosmer and Lemeshow, 2000] Hosmer, D. W. and Lemeshow, S. (2000). *Applied Logistic Regression (Wiley Series in Probability and Statistics)*. Wiley-Interscience Publication.

[Izenman, 2008] Izenman, A. J. (2008). *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning (Springer Texts in Statistics)*. Springer.

[Jen et al., 2012] Jen, C.-H., Wang, C.-C., Jiang, B. C., Chu, Y.-H., and Chen, M.-S. (2012). Application of classification techniques on development an early-warning system for chronic illnesses. *Expert Systems with Applications*, 39(10):8852–8858.

[Johnson, 2001] Johnson, D. E. (2001). *Metodos Multivariados Aplicados Al Analisis de Datos (Spanish Edition)*. I.T.P. Latin America.

[Johnson, 1987] Johnson, M. E. (1987). *Multivariate Statistical Simulation: A Guide to Selecting and Generating Continuous Multivariate Distributions (Wiley Series in Probability and Statistics)*. Wiley.

[Johnson and Wichern, 2007] Johnson, R. A. and Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis (6th Edition)*. Pearson.

[Jolicoeur and Mosimann, 1960] Jolicoeur, P. and Mosimann, J. E. (1960). Size and shape variation in the painted turtle. a principal component analysis. *Growth*, 24(4):339–354.

[Jovell, 1995] Jovell, A. J. (1995). *Análisis de regresión logítica*. Centro de Investigaciones Sociológicas.

[Koggalage and Halgamuge, 2004] Koggalage, R. and Halgamuge, S. (2004). Reducing the number of training samples for fast support vector machine classification. *Neural Information Processing-Letters and Reviews*, 2(3):57–65.

[Kotz and Nadarajah, 2004] Kotz, S. and Nadarajah, S. (2004). *Multivariate T-Distributions and Their Applications*. Cambridge University Press.

[Kumar et al., 1977] Kumar, R., Niero, M., Barros, M., Lucht, L., and Manso, A. (1977). Effect of the size of training samples on classification accuracy. *The Laboratory for Applications of Remote Sensing (LARS)*.

[Kuncheva, 2000] Kuncheva, L. I. (2000). *Fuzzy classifier design*, volume 49. Springer New York.

[Kuncheva, 2004] Kuncheva, L. I. (2004). *Combining pattern classifiers: methods and algorithms*. John Wiley & Sons.

[Langrand, 2000] Langrand, C. (2000). *Análisis de datos Métodos y ejemplos*. Escuela Colombiana de Ingeniería.

[Lubischew, 1962] Lubischew, A. A. (1962). On the use of discriminant functions in taxonomy. *Biometrics*, pages 455–477.

[Martin del Brio and Sanz, 2008] Martin del Brio, B. and Sanz, A. (2008). *Redes Neuronales y Sistemas Borrosos, 3. Ed. (Spanish Edition)*. Alfaomega.

[Mayorga, 1993] Mayorga, J. H. (1993). Un método de discriminación en dos grupos por medio de variables dicotómicas usando desarrollo binario. *Revista Colombiana de Estadística*, 27:26–38.

[Pedrycz, 1997] Pedrycz, W. (1997). *Fuzzy evolutionary computation*. Springer.

[Ponce Cruz, 2010] Ponce Cruz, P. (2010). *Inteligencia Artificial con Aplicación a la Ingeniería (Spanish Edition)*. Alfa Omega editores.

[Press, 2006] Press (2006). *Face Processing: Advanced Modeling and Methods*. Academic Press.

[R Core Team, 2013] R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[Rao and Govindaraju, 2013] Rao, C. and Govindaraju, V. (2013). *Handbook of Statistics: Machine Learning: Theory and Applications*, volume 31. North Holland.

[Raudys and Jain, 1991] Raudys, S. J. and Jain, A. K. (1991). Small sample size effects in statistical pattern recognition: Recommendations for practitioners. *IEEE Transactions on pattern analysis and machine intelligence*, 13(3):252–264.

[Ravi Kumar and Ravi, 2007] Ravi Kumar, P. and Ravi, V. (2007). Bankruptcy prediction in banks and firms via statistical and intelligent techniques–a review. *European Journal of Operational Research*, 180(1):1–28.

[Salazar et al., 2012] Salazar, D. A., Vélez, J. I., and Salazar, J. C. (2012). Comparison between svm and logistic regression: Which one is better to discriminate? *Revista Colombiana de Estadística*, 35(SPE2):223–237.

[Suykens et al., 2002] Suykens, J. A. K., Gestel, T. V., Brabanter, J. D., Moor, B. D., and Vandewalle, J. (2002). *Least Squares Support Vector Machines*. World Scientific Publishing Company.

[Vapnik, 1998] Vapnik, V. N. (1998). *Statistical Learning Theory (Adaptive and Learning Systems for Signal Processing, Communications and Control Series)*. Wiley-Interscience.

[Zadeh, 1965] Zadeh, L. A. (1965). Fuzzy sets. *Information and control*, 8(3):338–353.