Comparación de Árboles de Regresión y Clasificación y regresión logística

Sandra Carolina Serna Pineda

Director: Juan Carlos Correa Morales
Ph.D University of Kentucky
Profesor Asociado, Escuela de Estadística
Universidad Nacional de Colombia

Trabajo presentado como requisito para optar
al título de Magíster en Estadística

Escuela de Estadística
Facultad de Ciencias
Universidad Nacional de Colombia
Sede Medellín
2009

Resumen

El problema de la clasificación de individuos u objetos en grupos o poblaciones conocidas es de gran interés en estadística, por esta razón se han desarrollado varias técnicas para cumplir éste propósito. En este trabajo se presenta la comparación, mediante simulación Monte Carlo, de dos técnicas estadísticas de clasificicación: Árboles de Regresión y Clasificación (CART) y Regresión Logística. El comportamiento de las técnicas fue medido con la Tasa de Mala Clasificación (TMC). En general, la Regresión Logística presentó una Tasa de Mala Clasificación más baja que los Árboles de Clasificación. Se presenta una aplicación a la Encuesta de Innovación y Desarrollo Tecnológico, utilizando las técnicas estudiadas, para contribuir a un mejor conocimiento del sistema nacional de innovación en Colombia.

Palabras Calve: *Clasificación, CART: Árboles de clasificación y Regresión, Regresión Logística, Simulación, Tasa de Mala Clasificación*

Abstract

The classification problem of individuals or objects in known groups or populations is of great interest in statistics, for this reason it has been developed several techniques for achieving this purpose. This works presents the comparison between two classificaction techinques: Classification and Regression Trees, and Logistic Regression, by using Monte Carlo simulation. The behavior of both techniques was measured with the misclassification rate (MCR). Generally, logistic regression presented lower Misclassification rates than classification and regression trees. We present an application to the Innovation and Technologic Survey, with the mentioned techniques, to contribute to a better understanding of the national system of innovation in Colombia. The data bases were provided by the "Descubrimiento de Conocimiento sobre la Innovación en Colombia" project.

keywords: *Classification, CART: Classification and Regression Trees, Logistic Regression, Simulation, Misclassification Rate*

Índice general

Índice de figuras

CAPÍTULO 1

Introducción

El problema de la clasificación de individuos u objetos en grupos o poblaciones conocidas es de gran interés en estadística, por esta razón se han desarrollado técnicas para cumplir éste objetivo. Algunas de las más conocidas son:

- Análisis discriminante lineal.
- Análisis discriminate cuadrático.
- Análisis discriminate no-métrico.
- Regresión logística.

El análisis discriminante es una de las técnica más utilizadas para clasificación, pero el requerimiento de normalidad y homoscedasticidad no se cumple con frecuencia, como consecuencia de esto es necesario utilizar técnicas que no requieran tal supuesto, como la regresión logística Barajas (2007), Usuga (2006) y Castrillón (1998).

Se han desarrollado otras técnicas de clasificación basadas en árboles de decisión. Una de ellas es Árboles de Regresión y Clasificación, en adelante CART (de sus siglas en inglés, **C**lassification **A**nd **R**egression **T**rees), propuesta por Breiman et˜al. (1984).

Aunque los árboles de regresión y clasificación cada vez se hacen más populares, su desempeño respecto a otras técnicas de clasificación como la Regresión Logística ha sido poco estudiado. Por ello, el objetivo de éste trabajo es observar el desempeño

de los Árboles de Regresión y Clasificación con respecto a la Regresión Logística y determinar bajo que condiciones cuál de las dos pruebas es mejor, en términos de la Tasa de Mala Clasificación (TMC).

1.1. Planteamiento del problema

La clasificación es una actividad inherente al hombre, siempre existe la necesidad de ordenar o poner límites pues esto ayuda a entender fenómenos reales. En la solución de problemas y en la toma de decisiones uno de los primeros pasos consiste en clasificar el problema o la situación, para después aplicar la metodología correspondiente y ésta metodología dependerá en gran medida de la clasificación.
Podemos distinguir dos enfoques del problema de clasificación:

- El primero de ellos es cuando se conocen los grupos o categorías y se pretende ubicar los individuos dentro de estas categorías a partir de los valores de ciertos parámetros, para este caso las técnicas más utilizadas son el Análisis Discriminante y la Regresión Logística. También son conocidas como técnicas supervisadas (Webb, 2002).

- El segundo enfoque, que no es de interés en éste trabajo, ocurre cuando no se conocen los grupos de antemano y lo que se pretende es establecerlos a partir de los datos con los que se cuenta, dentro de estas técnicas se encuentra el Análsis de Clusters. Estas técnicas son conocidas también como no supervisadas (Webb, 2002).

De forma general, el análisis discriminante es una técnica que permite analizar las diferencias entre grupos de objetos a partir de variables medidas sobre los mismos. Algunas extensiones del análisis discriminante son:

- Análisis discriminante lineal, LDA: Está basado en el supuesto de normalidad multivariada e igualdad de las matrices de varianzas y covarianzas de los grupos. En la ecuación 1.1 se observan las relaciones lineales entre las variables x_i observadas, donde q es el número de grupos, p el número de variables medidas y $m = min(q-1, p)$, número de relaciones lineales (Seber, 1938).

$$\begin{aligned} y_1 &= a_{11}x_1 + \cdots + a_{1p}x_p + a_{10} \\ &\cdots \\ y_m &= a_{m1}x_1 + \cdots + a_{mp}x_p + a_{m0} \end{aligned} \tag{1.1}$$

  El objetivo del LDA es maximizar el cociente entre la varianza entre grupos y la varianza intra grupos.

$$Entre = \sum_{j=1}^{q} n_j \left(\bar{x}_{.j} - \bar{x}\right)\left(\bar{x}_{.j} - \bar{x}\right)^T \tag{1.2}$$

$$Intra = \sum_{j=1}^{q} \sum_{i=1}^{n} (\bar{x}_{ij} - \bar{x}_{.j}) (\bar{x}_{ij} - \bar{x}_{.j})^T \tag{1.3}$$

- Análisis discriminante cuadrático, QDA: Tiene como supuesto la normalidad multivariada pero no requiere igualdad de las matrices de varianzas y covarianzas. Marks & Dunn (1974) mostraron mediante simulación que QDA es más eficiente que LDA para muestras grandes. Para muestras pequeñas debe haber una marcada diferencia entre las matrices de varianzas y covarianzas para que QDA sea mejor que LDA (Seber, 1938).

- Análisis discriminante no-métrico, NDA: Propuesto por Raveh (1989) como un procedimiento que busca una función discriminante que maximice un índice de separación entre dos grupos.

Los tres procedimientos del análisis discriminante han sido comparados con la regresión logística en los estudios de Shelley & Donner (1987), Castrillón (1998), Usuga (2006) y Barajas (2007) obteniendo que, en general, la regresión logística produce mejores resultados.

Se han encontrado pocos estudios comparativos entre CART y las demás metodologías de clasificación, y los pocos hallados fueron realizados para un tipo específico de datos.

Puesto que se han desarrollado tantas técnicas para clasificación es necesario saber bajo qué condiciones y cuál de ellas es mejor en términos de la menor tasa de mala clasificación, por ésta razón se han realizado los estudios de comparación antes mencionados.
El objetivo es determinar cuál técnica, entre CART y Regresión Logística obtiene menores tasas de mala clasificación para diferentes conjuntos de datos.

1.2. Marco teórico

Frecuentemente la investigación estadísticas se ve enfrentada a manipular grandes cantidades de datos complejos que incluyen un gran número de variables, de los cuales es necesario obtener información, encontrar patrones y definir tendencias. Con este propósito Sonquist, Baker y Morgan, (1971) propusieron el programa AID (Automatic Interaction Detection), el cual representa uno de los primeros métodos de ajuste de los datos basados en modelos de árboles de clasificación (Hadidi, 2003). En 1980, Kass propone un algoritmo recursivo de clasificación no binaria llamado

CHAID (Chi Square Automatic Interaction Detection). Otros métodos más recientes son: FIRM (Formal Inference-based Recursive Modeling) propuesto por Hawkins (Hadidi, 2003); y MARS (Multivariate Adaptive Regression Splines), propuesto por Friedman en el año 1991. Este capítulo se centra en la metodología CART la cual se usa para la construcción de árboles de regresión y clasificación, y utiliza un algoritmo recursivo de partición binaria en cada nodo.

1.2.1. CART: Classification And Regression Trees

Breiman (1984), desarrolló el algoritmo CART cuyo resultado es en general, un árbol de decisión, las ramas representan conjuntos de decisiones y cada decisión genera reglas sucesivas para continuar la clasificación (partición) formando así grupos homogéneos respecto a la variable que se desea discriminar. Las particiones se hacen en forma recursiva hasta que se alcanza un criterio de parada, el método utiliza datos históricos para construir el árbol de decisión, y este árbol se usa para clasificar nuevos datos.

CART es un método no-paramétrico de segmentación binaria donde el árbol es construído dividiendo repetidamente los datos. En cada división los datos son partidos en dos grupos mutuamente excluyentes. El nodo inicial es llamado nodo raíz o grupo *madre* y se divide en dos grupos *hijos* o nodos, luego el procedimiento de partición es aplicado a cada grupo *hijo* por separado. Las divisiones se seleccionan de modo que "la impureza" de los *hijos* sea menor que la del grupo *madre* y éstas están definidas por un valor de una variable explicativa (Deconinck et˜al., 2006). El objetivo es particionar la respuesta en grupos homogéneos y a la vez mantener el árbol razonablemente pequeño.
Para dividir los datos se requiere un criterio de particionamiento el cual determinará la medida de impureza, esta última establecerá el grado de homogeneidad entre los grupos.

El análisis de árboles de clasificación y regresión (CART) generalmente consiste en tres pasos (Timofeev, 2004):

1. Construcción del árbol máximo.

2. Poda del árbol.

3. Selección del árbol óptimo mediante un procedimiento de validación cruzada ("cross-validation").

Construcción del árbol máximo

El árbol máximo es construido utilizando un procedimiento de partición binario, comenzando en la raíz del árbol, este árbol es un modelo que describe el conjunto de entrenamiento (grupo de datos original) y generalmente es *sobreajustado*, es decir, contiene gran cantidad de niveles y nodos que no producen una mejor clasificación y puede ser demasiado complejo.
Cada grupo es caracterizado por la distribución (respuesta categórica), o por la media (respuesta numérica) de la variable respuesta, el tamaño del grupo y los valores de las variables explicativas que lo definen. Gráficamente, el árbol se representa con el nodo raíz (los datos sin ninguna división), al iniciar y las ramas y hojas debajo (cada hoja es el final de un grupo).

Calidad del Nodo: Función de Impureza

La función de impureza es una medida que permite determinar la calidad de un nodo, esta será denotada por $i(t)$. Existen varias medidas de impureza (criterios de particionamiento) que nos permiten analizar varios tipos de respuesta, las tres medidas más comunes presentadas por Breiman et˜al. (1984), para árboles de clasificación son:

- El índice de información o entropía el cual se define como:

$$i\left(t\right)=\sum_{j}p\left(j\left|t\right.\right)\ln p\left(j\left|t\right.\right) \tag{1.4}$$

  El objetivo es encontrar la partición que maximice $\Delta i(t)$ en la ecuación 1.5

$$\Delta i(t)=-\sum_{j=1}^{k}p(j|t)\ln p(j|t), \tag{1.5}$$

  donde $j=1,...,k$ es el número de clases de la variable respuesta categórica y $p(j|t)$ la probabilidad de clasificación correcta para la clase j en el nodo t.

- El índice Gini tiene la forma

$$i\left(t\right)=\sum_{i\neq j}p\left(j\left|t\right.\right)p\left(i\left|t\right.\right) \tag{1.6}$$

  Encontrar la partición que maximice $\Delta i(t)$ en 1.7

$$\Delta i=-\sum_{j=1}^{k}[p_{j}(t)]^{2}, \tag{1.7}$$

Este índice es el más utilizado. En cada división el índice Gini tiende a separar la categoría más grande en un grupo aparte, mientras que el índice de información tiende a formar grupos con más de una categoría en las primeras decisiones, y por último,

- El índice "Towing". A diferencia del índice Gini, Towing busca las dos clases que juntas formen más del 50 % de los datos, esto define dos "super categorías" en cada división para las cuales la impureza es definida por el índice Gini. Aunque el índice towing produce árboles más balanceados, este algoritmo trabaja más lento que la regla de Gini (Deconinck et˜al., 2006). Para usar el índice towing seleccione la partición s, que maximice

$$\frac{p_L p_R}{4}\left[\sum_j |p(j|t_L) - p(j|t_R)|\right]^2, \tag{1.8}$$

donde t_L y t_R representan los nodos hijos izquierdo y derecho respectivamente, p_L y p_R representan la proporción de observaciones en t que pasaron a t_L y a t_R en cada caso.

Poda del árbol

El árbol obtenido es generalmente sobreajustado por tanto es podado, cortando sucesivamente *ramas* o nodos terminales hasta encontrar el tamaño "adecuado" del árbol.
Breiman et˜al. (1984) introducen algunas ideas básicas para resolver el problema de seleccionar el mejor árbol. Computacionalmente el procedimiento descrito es complejo. Una forma es buscar una serie de árboles anidados de tamaños decrecientes (De'ath & Fabricius, 2000), cada uno de los cuales es el mejor de todos los árboles de su tamaño.
Estos árboles pequeños son comparados para determinar el óptimo. Esta comparación esta basada en una función de costo complejidad , $R_\alpha(T)$.
Para cada árbol T , la función costo - complejidad se define como (Deconinck et˜al., 2006):

$$R_\alpha(T) = R(T) + \alpha|\tilde{T}| \tag{1.9}$$

donde $R(T)$ es el promedio de la suma de cuadrados entre los nodos, puede ser la tasa de mala clasificación total o la suma de cuadrados de residuales total dependiendo del tipo de árbol, $|\tilde{T}|$ es la complejidad del árbol, definida como el número total de nodos del sub-árbol y α es el parámetro de complejidad.
El parámetro α es un número real mayor o igual a cero, Cuando $\alpha = 0$ se tiene el árbol más grande y a medida que α se incrementa, se reduce el tamaño del árbol. La función $R_\alpha(T)$ siempre será minimizado por el árbol más grande, por tanto se

necesitan mejores estimaciones del error, para esto Breiman et˜al. (1984) proponen obtener estimadores "honestos" del error por "validación cruzada". Computacionalmente el procedimiento es exigente pero viable, pues solo es necesario considerar un árbol de cada tamaño, es decir, los árboles de la secuencia anidada.

Selección del árbol óptimo

De la secuencia de árboles anidados es necesario seleccionar el árbol óptimo y para esto no es efectivo utilizar comparación o penalización de la complejidad (De'ath & Fabricius, 2000), por tanto se requiere estimar con precisión el error de predicción y en general esta estimación se hace utilizando un procedimiento de validación cruzada. El objetivo es encontrar la proporción óptima entre la tasa de mala clasificación y la complejidad del árbol, siendo la tasa de mala clasificación el cociente entre las observaciones mal clasificadas y el número total de observaciones.
El procedimiento de validación cruzada puede implementarse de dos formas:

- Si se cuenta con suficientes datos se parte la muestra, sacando la mitad o menos de los datos y se construye la secuencia de árboles utilizando los datos que permanecen, luego predecir, para cada árbol, la respuesta de los datos que se sacaron al iniciar el proceso; obtener el error de las predicciones; seleccionar el árbol con el menor error de predicción.

En general no se cuenta con suficientes datos como para utilizar el procedimiento anterior, de modo que otra forma sería:

- Validación cruzada con partición en V, (v-fold cross validation, se menciona más adelante).

La idea básica de la "Validación cruzada" es sacar de la muestra de aprendizaje una muestra de prueba, con los datos de la muestra de aprendizaje se calculan los estimadores y el subconjunto sacado es usado para verificar el desempeño de los estimadores obtenidos utilizandolos como "datos nuevos". El desempeño entendido como el error de predicción, es acumulado para obtener el error medio absoluto del conjunto de prueba.

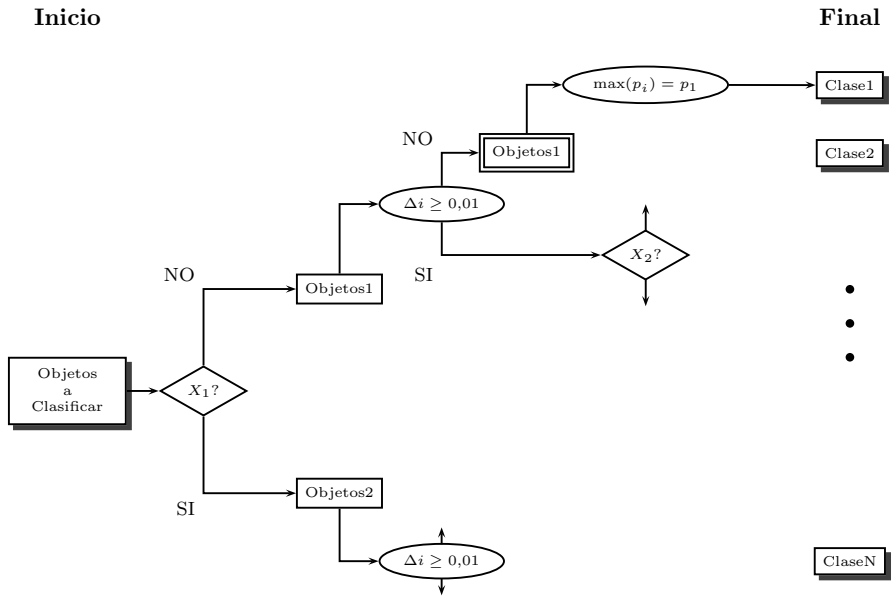
Como se mencionó anteriormente, para la metodolgía CART generalmente se utiliza Validación Cruzada con partición en V (v-fold cross validation), tamando $V = 10$ y el procedimiento es el siguiente:

- Dividir la muestra en diez grupos mutuamente excluyentes y de aproximadamente igual tamaño.

- Sacar un conjunto por vez y construír el árbol con los datos de los grupos restantes. El árbol es usado para predecir la respuesta del conjunto eliminado.
- Calcular el error estimado para cada subconjunto.
  Repetir los "ítems" dos y tres para cada tamaño de árbol.
- Seleccionar el árbol con la menor tasa de mala clasificación.

Al llegar a este punto se procede a analizar el árbol obtenido.

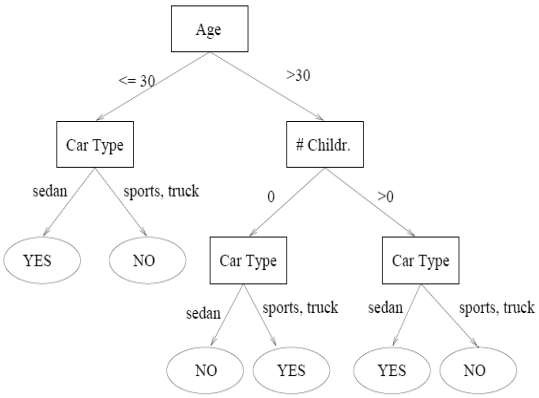
La siguiente figura es el diagrama de flujo del algoritmo CART.



Como ejemplo suponga el árbol y los datos en la Figura 1.1, donde se quiere determinar un conjunto de reglas que indiquen si un conductor vive o no en los suburbios.

Se concluye:

- Si $Age \leq 30$ y $CarType = Sedan$ entonces Si
- Si $Age \leq 30$ y $CarType = truck/Sports$ entonces No
- Si $Age > 30$, $Children = 0$ y $CarType = Sedan$ entonces No
- Si $Age > 30$, $Children = 0$ y $CarType = truck/Sports$ entonces Si



Car type	Age	Chd	Sub
sedan	23	0	yes
sports	31	1	no
sedan	36	1	no
truck	25	2	no
sports	30	0	no
sedan	36	0	no
sedan	25	0	yes
truck	36	1	no
sedan	30	2	yes
sedan	31	1	yes
sports	25	0	no
sedan	45	1	yes
sports	23	2	no
truck	45	0	yes

Figura 1.1: *Ejemplo árbol de clasificación. Fuente: Dobra (2002)*

- Si $Age > 30$, $Children > 0$ y $CarType = Sedan$ entonces Si
- Si $Age > 30$, $Children > 0$ y $CarType = truck/Sports$ entonces No

1.2.2. Regresión logística o Discriminante logístico

Cuando se desea clasificar un sujeto dentro de uno o más grupos previamente determinados a partir de un conjunto de características observadas del sujeto, es razonable pensar en la utilización de una medida probabilística.
La regresión logística estima la probabilidad de un suceso en función de un conjunto de variables explicativas y en la construcción del modelo no hay ningún supuesto en cuanto a la distribución de probabilidad de las variables por lo que puede incluirse cualquier tipo de variable.
El modelo de regresión logística puede considerarse como una formula para calcular la probabilidad de pertenencia a uno de los grupos, de manera que este estima la probabilidad de que una observación pertenezca a uno de los grupos.
La interpretación del resultado de la aplicación de esta metodología es sencilla por tratarse en términos de probabilidad.

El modelo de regresión logística se formula matemáticamente relacionando la probabilidad de ocurrencia de algún evento, E, condicionado a un vector, x, de variables explicativas, a través de la forma funcional de la $c.d.f$ logística (Press &

Wilson, 1978). Así,

$$p(x) = P(E|x) = \frac{1}{1 + e^{-\alpha - \beta^T x}}, \tag{1.10}$$

donde (α y β) son parámetros desconocidos que se estiman de los datos.

Este modelo puede usarse para clasificar un objeto en una de dos poblaciones, siendo E el evento que el objeto pertenezca a la primera población, y x denote un vector de atributos del objeto que será clasificado.
Una medida útil para verificar la calidad en las clasificaciones obtenidas por el modelo puede ser la tasa de mala clasificación (tasa de desaciertos), que es la proporción de observaciones mal clasificadas. El modelo de regresión logística tiene como ventaja que es claro y pueden usarse todos los tipos de variables.

Regresión Logística Multinomial

Ahora consideremos que se tiene más de una variable regresora y, por lo menos una es de tipo cuantitativo. La técnica de regresión logística multinomial consiste en la estimación de la probabilidad de que una observación x pertenezca a uno de los grupos, dados valores de las p variables que la conforman.
El modelo compara G-1 categorías contra una categoría de referencia. Dadas n observaciones (y_i, x_i) donde x_i es un vector con p variables y y_i es una variable aleatoria independiente Multinomial con valores $1, 2, ..., G$ la cual indica el grupo al cual pertence cada observación, la probabilidad condicional de pertenencia de x_i a cada grupo está dada por:

$$P(y = j|x_i) = \frac{\exp\left(\alpha_{1j} + \beta'_{1j} x_i\right)}{1 + \sum_{k=2}^{G} \exp\left(\alpha_{1j} + \beta'_{1k} x_i\right)} \tag{1.11}$$

Donde $\alpha_{11} = \beta_{11} = 0$. Para clasificar la observación p-variada, en un grupo, se calcula la probabilidad de pertenencia a cada uno de los G grupos y se asigna la mayor probabilidad (Hosmer & Lemeshow, 1989).

Durante el desarrollo de este trabajo se presentó un problema en la convergencia de los estimadores del modelo de regresión logística; éste es denominado "problema de la separación". En la siguiente sección se presenta una breve descripción de esta situación solo para efectos aclaratorios.

El problema de la Separación

El modelo de regresión logística es uno de los más utilizados y aplicados, pero cuando los datos no están bien estructurados o hay muy pocos, se puede presentar el problema de la separación, donde el proceso de estimación de los estimadores

por máxima verosimilitud de los parámetros, no converge (el algoritmo de Newton-Raphson crece infinitamente).

La principal consecuencia de la no convergencia es que no se puede realizar inferencias sobre los estimadores del modelo. El modelo resultante puede servir para clasificar observaciones, pero debe evitarse realizar inferencias (Allison, 1999).

Albert & Anderson (1986) estudian las posibles configuraciones de los datos en el espacio $\mathcal{R}^p$ que caen en tres categorías principales o tres posibles formas de separación en la estructura de datos (Valencia, 2002):

- **Separación Completa:** Es una condición dónde la variable explicativa o una combinación lineal de ellas predicen la respuesta perfectamente. En este caso es imposible calcular los estimadores de máxima verosimilitud para los parámetros β (Ver Ecuación 1.10) porque el algoritmo iterativo necesario para el cálculo de los mismos, no converge.
- **Separación Cuasicompleta:** Ocurre cuando valores de la variable respuesta se traslapan o empatan con valores de la variable explicativa. El análisis no verifica la separación *cuasi-completa*, pero los síntomas son los valores calculados sumamente grandes por los parámetros β o los errores grandes. El análisis también puede no converger.
- **Sobreposición:** Cuando no se presenta separación completa o cuasicompleta, la variable respuesta ocurre en cualquier parte del rango de la(s) variable(s) explicativa(s). Los estimadores de máxima verosimilitud existen.

Las posibles causas de la *separación* son: problemas de diseño, mala planeación del experimento o escasez de datos (puede ocurrir cuando el evento es raro), es decir, tamaños de muestra pequeños. En la Figura 1.2 se observa las tres situaciones descritas.

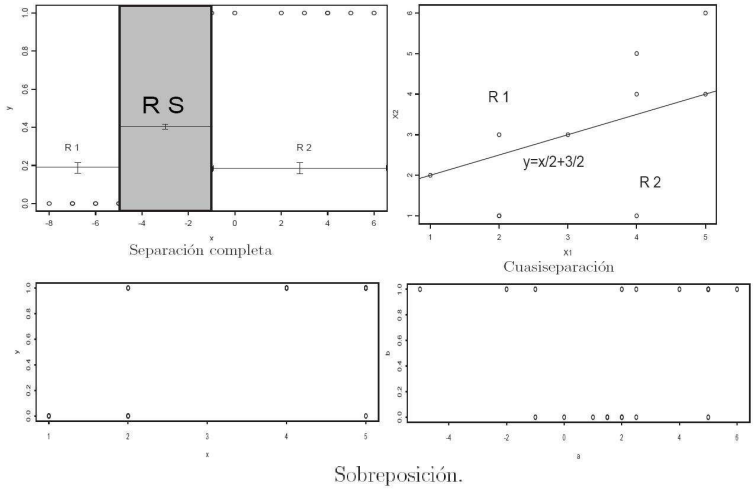


Figura 1.2: *El problema de la separación en Regresión Logística.* *Fuente: Valencia (2002)*

CAPÍTULO 2

Estudios comparativos realizados y propuesta de índice de clasificación

Como mencioó anteriormente, se han desarrollado múltiples técnicas para cumplir el objetivo de clasificación de objetos bajo diferentes supuestos y, principalmente, desde dos puntos de vista: la estadística y la minería de datos[1]. En ambos casos se pueden reconocer dos enfoques: el análisis supervisado y el no supervisado. En éste proyecto de investigación, se está interesado en el primero de ellos, es decir, se conocen de antemano los grupos a los cuales pueden pertenecer las observaciones a clasificar y lo que se quiere es ubicarlas en uno de los grupos.

La regresión logística es uno de los primeros procedimientos utilizados para clasificación. El modelo de regresión logística utiliza la estimación por máxima verosimilitud y estima la probabilidad de que un evento dado ocurra. Para la regresión logística la respuesta debe ser binaria y las covariables pueden ser categóricas o continuas.

La regresión logística, debido a sus ventajas, ha sido comparada con varios métodos de clasificación y se prefiere por ser una metodología sencilla de implementar y fácil de interpretar.
Algunas comparaciones que se han desarrollado son:

[1]El término "Minería de datos" (Data Mining), aparece en la década de los 90 en el ámbito empresarial, a raíz de la gran cantidad de datos almacenados por las organizaciones, su objetivo es la generación de conocimiento para la toma de decisiones. La minería de datos se realiza a partir de técnicas pertenecientes a la Inteligencia Artificial y/o la Estadística. La diferencia entre estos dos enfoques no es tema de esta investigación, para mayor información ver Banet (2001)

- Press & Wilson (1978) realizan la comparación entre la Regresión Logística y el análisis discriminante.
- Shelley & Donner (1987) presentan la comparación entre la Regresión Logística Multinomial y el análisis discriminante con múltiples grupos.
- Castrillón (1998) presenta una comparación del análisis discriminante lineal y cuadrático con la regresión logística para clasificar individuos en dos poblaciones.
- Usuga (2006) presenta una comparación entre Análisis de Discriminante nométrico y Regresión Logística.
- Barajas (2007) presenta una comparación entre el análisis discrimínate nométrico y la regresión logística multinomial.

En general, las conclusiones de estos estudios han sido favorables para la regresión logística, incluso cuando se cumplen los supuestos de la metodología contra la cual es comparada. Por esta razón en este trabajo se considerará únicamente la regresión logística.

Por otro lado, aparecen nuevas técnicas de clasificación basadas en árboles de decisión y, bajo esta estructura de clasificación, se han desarrollado varios algoritmos como: CART (Classification And Regression Trees), CHAID (Chi-squared Automatic Interaction Detection), y algunas variaciones sobre éstos.

El uso de árboles de decisión en la comunidad estadística proviene de AID (Automatic Interaction Detection), propuesto por Morgan y Sonquist en 1963, y del trabajo posterior llamado THAID propuesto por Morgan y Messenger en la década de los 70.

En la década de los 80 aparece CART, un procedimiento propuesto por Breiman et˜al. (1984) como un algoritmo recursivo de partición binaria que divide la muestra en dos nodos hijos cada vez, basado en una medida de impureza. La medida de impureza esta relacionada con la homogeneidad de los nodos hijos y el método de particionamiento busca maximizar la homogeneidad de los mismos.

Los estudios comparativos de árboles de regresión y clasificación se han realizado contra otras metodologías más conocidas dentro del área de "minería de datos", algunos de estos trabajos son:

- Rudolfer et˜al. (1999) presentan una comparación entre la regresión logística y la inducción de árboles de decisión en el diagnóstico del síndrome del túnel carpiano.

- Caruana & Niculescu-Mizil (Caruana & Niculescu-Mizil) presentan una comparación empírica de algoritmos de aprendizaje supervisado, entre los cuales se encuentran la regresión logística, los árboles de decisión y los bosques aleatorios, entre otros.

- Kurt et˜al. (2008) presentan una comparación de la regresión logística, los árboles de regresión y clasificación y las redes neuronales para predecir enfermedad coronaria.

La mayoría de las comparaciones han sido realizadas para datos específicos en general en el área de la medicina.

En este trabajo se desarrolla la comparación entre Árboles de Clasificación y la Regresión Logística para diferentes estructuras de datos generados vía Monte Carlo, utilizando como función de impureza o criterio de partición de los nodos, el índice de Gini.

2.1. Propuesta de Índice de Clasificación

El coeficiente de determinación (R^2) que aparece en los modelos de regresión es una medida de la calidad del ajuste del modelo propuesto, mide la proporción de variabilidad total de la variable dependiente respecto a su media.
Kvalseth (1985) propone ciertas propiedades necesarias que debe tener un "buen" R^2. Algunas de estas propiedades son:

1. Debe ser útil como medida de bondad de ajuste y tener interpretación sencilla.

2. No debe tener dimensión, es decir, debe ser independiente de la unidad de medida de las variables del modelo.

3. El rango de variación debe estar bien definido.

Se ha encontrado que este indicador no es adecuado para modelos de regresión donde se tiene como variable respuesta una variable categórica, por ello se han propuesto varios medidas análogas al R^2 (Menard, 2000), (Mittlböck & Schemper, 1996). Entre las medidas propuestas se encuentran:

- El R^2 por mínimos cuadrados ordinarios

$$R^2_O = 1 - \frac{\sum (y - \hat{y})^2}{\sum (y - \overline{y})^2} \tag{2.1}$$

- El R^2 logaritmo de la razón de verosimilitud

$$R_L^2 = \frac{ln(L_0) - ln(L_M)}{ln(L_0)} = 1 - \frac{ln(L_M)}{ln(L_0)} \tag{2.2}$$

- La mejora del promedio geométrico cuadrado por observación R^2

$$R_M^2 = 1 - \left(\frac{(L_0)}{(L_M)}\right)^{\frac{2}{n}} \tag{2.3}$$

- La mejora del promedio geométrico ajustado al cuadrado R^2

$$R_N^2 = \frac{1 - (\frac{L_0}{L_M})^{\frac{2}{n}}}{1 - (L_0)^{\frac{2}{n}}} \tag{2.4}$$

- El Coeficiente de Contingencia

$$R_C^2 = \frac{G_M}{G_M + n} \tag{2.5}$$

Donde L_0 es la función de verosimilitud del modelo que contiene solo la media, L_M es la función de verosimilitud que contiene todos los predictores, el estadístico chi-cuadrado del modelo $G_M = -2\left[\ln\left(L_0\right) - \ln\left(L_M\right)\right]$, $\hat{y}$ el valor predicho de la variable dependiente Y obtenida del modelo, una probabilidad continua entre cero y uno, y el valor observado de la variable dependiente, y $\overline{y}$ el valor promedio de la variable dependiente Y, una probabilidad continua.

Estas medidas proponen comparar las probabilidades predichas del modelo de regresión logística con los datos observados, comparación que no parece ser adecuada pues de esta manera nunca habría un ajuste apropiado. Por ello se propone una medida análoga al coeficiente de determinación, pero utilizando para la comparación tanto las probabilidades predichas del modelo de regresión logística como las probabilidades predichas por un modelo de regresión no paramétrica.
La regresión no paramétrica es un procedimiento que requiere un número mínimo de supuestos, donde el ajuste es realizado únicamente apartir de los datos; por esta razón el modelo ajustado por regresión no paramétrica podría considerarse como la verdadera curva de los datos (Cleveland, 1979).

Se propone el siguiente indicador:

$$R_{jcc}^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_{i,(NP)} - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_{i,(NP)} - \bar{y}_{(NP)}\right)^2}, \tag{2.6}$$

donde $y_{(NP)}$ es la probabilidad predicha con el modelo no paramétrico, y_i es la probabilidad predicha con el modelo de regresión logística o con los Árboles de Clasificación, y $\overline{y}_{(NP)}$ es el promedio de las probabilidades predichas con el modelo de regresión no paramétrica.

En la figura 2.1 se observa el ajuste por regresión logística y regresión no paramétrica. En la construcción de la nueva propuesta, como ejemplo ilustrativo, las lineas azules representan las distancias entre los dos ajustes y cada ajuste con la media. Para mostrar que R^2_{jcc} tiene rango de variación definido, considere las

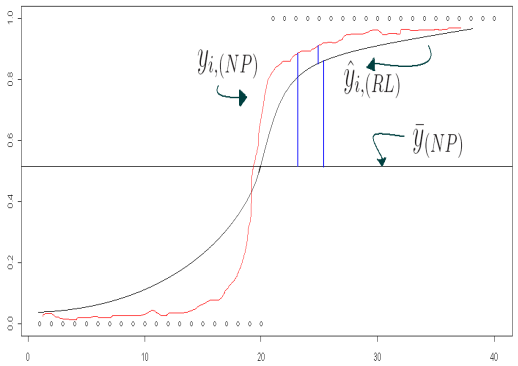


Figura 2.1: *Construcción del* R^2_{jcc}

distancias en la figura 2.1 para el caso de la regresión logística.

$$\left|y_{i,(NP)} - \hat{y}_{i,(RL)}\right| \leq \left|y_{i,(NP)} - \bar{y}_{(NP)}\right| \tag{2.7}$$

El término de la izquierda en la ecuación 2.7 representa el error de ajuste. Cuando $\hat{y}_{j,(RL)} = \bar{y}_{(NP)}$, $\left|y_{i,(NP)} - \hat{y}_{i,(RL)}\right| = \left|y_{i,(NP)} - \bar{y}_{(NP)}\right|$, usando la identidad $\sqrt{x^2} = |x|$ se obtiene

$$\left(y_{i,(NP)} - \hat{y}_{i,(RL)}\right)^2 \leq \left(y_{i,(NP)} - \bar{y}_{(NP)}\right)^2 \tag{2.8}$$

esta relación de orden entre números positivos (distancias), conduce a la siguiente expresión

$$\sum_i \left(y_{i,(NP)} - \hat{y}_{i,(RL)}\right)^2 \leq \sum_i \left(y_{i,(NP)} - \bar{y}_{(NP)}\right)^2. \tag{2.9}$$

Observe que la relación previa es aproximada ya que en la región de inflexión de las curvas, ésta no se satisface; argüimos al respecto que siempre que se garanticen grandes pendientes entre las regresiones no paramétrica y logística, con el fin de que el área entre éstas se minimize, es válida la expresión 2.9. Este supuesto se basa en

el hecho que la mayoría de los datos se encuentran en los extremos de las curvas. Como $\sum_i \left(y_{i,(NP)} - \bar{y}_{(NP)}\right)^2 \neq 0$, entonces

$$0 \leq \frac{\sum_i \left(y_{i,NP} - \hat{y}_{i,RL}\right)^2}{\sum_i \left(y_{i,NP} - \bar{y}_{NP}\right)^2} \leq 1 \tag{2.10}$$

este indicador es independiente de la unidad de medida de las variables implicadas en el modelo, tiene rango $0 \leq R^2_{jcc} \leq 1$ y tiene residuales positivos y negativos igualmente pesados.

CAPÍTULO 3

Estudio de Simulación

En este estudio de simulación se compararon dos métodos estadísticos de clasificación: la regresión logística para dos o más poblaciones y los árboles de regresión y clasificación, más conocidos por sus siglas en inglés, CART. Se utilizaron variables provenientes de la distribución normal multivariada, la distribución Lognormal y la distribución normal sesgada, para diferentes parámetros de sesgamiento. Detalles sobre estas tres distribuciones se pueden ver en el apéndice. Detalles sobre las distribuciones se pueden encontrar en el apéndice A.3

3.1. Metodología

El procedimiento de comparación para los dos procedimientos se presenta en los siguientes pasos:

1. Generación de las muestras para cada uno de los casos de simulación. Se utilizaron diferentes poblaciones, con diferentes parámetros para la distribución generadora de las muestras (media y matriz de varianzas y covarianzas). Adicionalmente, se variaron los tamaños de muestra de los grupos. Se definieron tres tamaños de muestra: 20, 50 y 100, tomados de la metodología de simulación realizada por Usuga (2006) y Barajas (2007).
   Para las matrices de varianzas y covarianzas se utilizaron cinco valores de correlación: 0.1, 0.3, 0.5, 0.7 y 0.9.

2. La clasificación se lleva a cabo mediante un procedimiento de validación cruzada conocido como *Leave one out*, que consiste en eliminar una observación

completa de los datos, ajustar el modelo de interés y luego predecir para el dato eliminado.

3. Se obtiene la tasa de mala clasificación (TMC), así:

$$TMC = \frac{\text{Observaciones mal clasificadas}}{\text{Total de observaciones en el Grupo}}$$

4. Repetición de los pasos de dos y tres el número de simulaciones determinadas, es decir, 1000 veces.

Para la solución del problema se utilizaron datos simulados de:

⋆ La distribución Normal bivariada, para el caso de clasificación en dos grupos y multivariada para clasificar en más de dos grupos.

⋆ La distribución Lognormal bivariada, para el caso de clasificación en dos grupos y multivariada para clasificar en más de dos grupos.

⋆ La distribución Normal Sesgada bivariada, para el caso de clasificación en dos grupos y multivariada para clasificar en más de dos grupos.

Los pasos anteriores se aplicaron sobre los siguientes casos de simulación

3.1.1. Casos de Simulación para Clasificación en dos Grupos

Caso1, Distribución normal bivariada con estructura de varianza y covarianza igual para los dos grupos y tres diferentes vectores de medias.

Caso1A) $\mathcal{N}\left(\binom{0}{0}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right), N\left(\binom{0}{1}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right)$

Caso1B) $\mathcal{N}\left(\binom{0}{0}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right), N\left(\binom{0}{2}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right)$

Caso1C) $\mathcal{N}\left(\binom{0}{0}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right), N\left(\binom{0}{10}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right)$

Caso2, Distribución normal bivariada con estructura de varianza y covarianza diferente para cada grupo y tres diferentes vectores de medias.

Caso2A) $\mathcal{N}\left(\binom{0}{0}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right), N\left(\binom{0}{1}, 2\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right)$

Caso2B) $\mathcal{N}\left(\binom{0}{0}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right), N\left(\binom{0}{2}, 2\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right)$

Caso2C) $\mathcal{N}\left(\binom{0}{0}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right), N\left(\binom{0}{10}, 2\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right)$

Caso3, Datos de la distribución lognormal, generados transformando los generados a partir de las siguientes normales bivariadas, (Ver apéndice A.2).

Caso3A) $\mathcal{N}\left(\binom{0}{0}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right), N\left(\binom{0}{1}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right)$

Caso3B) $\mathcal{N}\left(\binom{0}{0}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right), N\left(\binom{0}{2}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right)$

Caso3C) $\mathcal{N}\left(\binom{0}{0}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right), N\left(\binom{0}{10}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right)$

Por último, el Caso4, donde se generan datos de la distribución normal sesgada (Ver apéndice A.3)
La distribución normal sesgada está dada por, (Azzalini & Dalla_Valle, 1996):

$$\phi\left(z;\lambda\right) := 2\phi\left(z\right)\Phi\left(\lambda z\right)$$

donde $z \in R$, y los valores de λ seleccionados (para el caso bivariado) son:

- $\lambda = (1, 1)$
- $\lambda = (1, 10)$
- $\lambda = (1, 5)$
- $\lambda = (1, 20)$

Se consideraron, además, diferentes tamaños de muestra.

3.2. Resultados

En las Figuras de esta sección, las letras *a, c* y *e* que aparecen en los paneles corresponden a la separación entre los grupos (Ver Figura 3.1), es decir, al vector de medias de la distribución de la cual se generaron los datos, siendo $a = (0, 0), (0, 1)$, $c = (0, 0), (0, 2)$ y $e = (0, 0), (0, 10)$, éste último con el objetivo de observar qué pasa cuando los grupos están muy separados; los números *20, 50* y *100* corresponden a los tamaños de muestra de ambos grupos. Cuando aparecen dos números juntos significa que los grupos tiene tamaños de muestra diferentes (*muestras desbalanceadas*).

Para todos los casos de simulación se presentan dos situaciones diferentes, en la primera los dos grupos son generados con igual tamaño de muestra, en la segunda situación los dos grupos son generados con tamaños de muestra diferentes (desbalanceados).

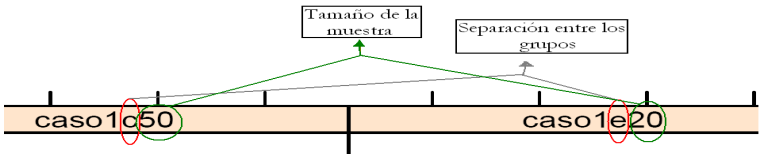


Figura 3.1: *Identificación de los paneles en las gráficas*

3.2.1. Clasificación en dos Grupos

Caso1, $\Sigma_1 = \Sigma_2$

Para el **Caso1**, cuando se tienen muestras balanceadas, ver Figura 3.2, se observa que al incrementar el tamaño de muestra y mantener la misma separación entre los grupos, la Tasa de Mala Clasificación no cambia, principalmente para la Regresión Logística donde además, se nota que al aumentar la correlación entre las variables explicativas la Tasa de Mala Clasificación (en adelante TMC), se reduce. Para los árboles de clasificación, al incrementar la correlación entre las variables se nota una reducción en la TMC que se hace más evidente al aumentar el tamaño de las muestras y la separación entre los grupos, pero la TMC siempre es más alta en ésta metodología de clasificación. Cuando la separación entre los grupos es grande (letra *e* en el panel) se obtiene una clasificación perfecta para ambas metodologías, pero se presenta el problema de la separación completa de la Regresión Logística (Ver sección 1.2.2), debido a la separación de los grupos, luego los estimadores de máxima verosimilitud no convergen y los resultados obtenidos no son válidos. En esta situación los árboles de clasificación presentan una ventaja respecto a la Regresión Logística.

Cuando los grupos son desbalanceados, ver Figura 3.3,se observa que a medida que el desbalance entre los grupos es mayor la TMC se reduce, la razón por la que esto ocurre esta determinada por la *Probabilidad de mala clasificación*, es decir, cuando los tamaños de muestra son muy similares en los grupos la probabilidad de mala clasificación está alrededor de 0,5 mientras que, a medida que aumenta el desbalance, la probabilidad de mala clasificación en el grupo mayor se reduce notablemente. También se nota que la TMC se reduce al aumentar la correlación entre las variables. Al incrementar la separación entre los grupos se nota el mismo comportamiento que se describió en la Figura 3.2 y, en este caso, para la máxima separación entre los grupos, se obtiene clasificación perfecta para los Árboles de Clasificación y el problema de convergencia de la Regresión Logística se hace mas evidente.

Las gráficas presentadas en este capítulo representan el comportamiento general observado para todos los casos de simulación, los demás resultados se encuentran en el apéndice B.

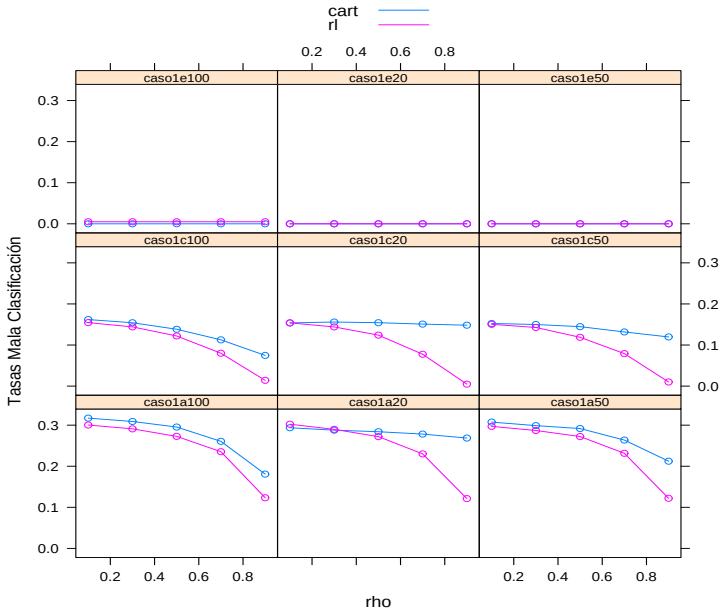


Figura 3.2: *Caso1,* $\Sigma_1 = \Sigma_2$

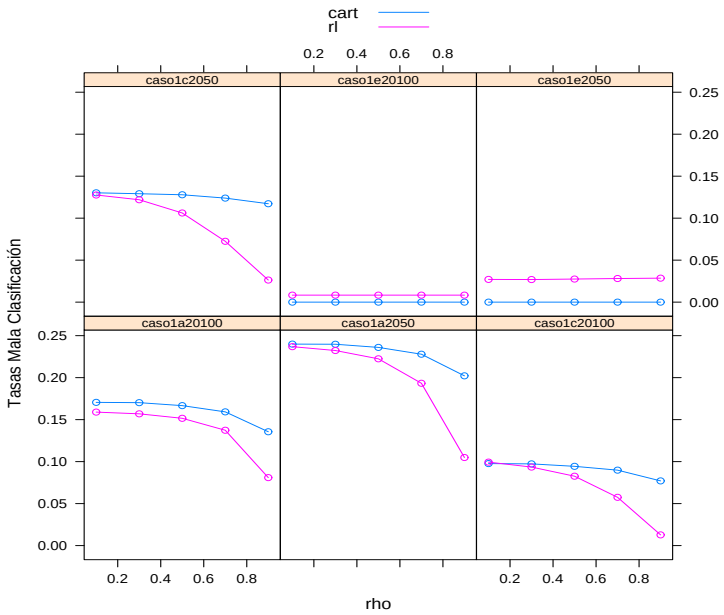


Figura 3.3: *Caso1,* $\Sigma_1 = \Sigma_2$ *muestras desbalanceadas*

CAPÍTULO 4

Aplicación: Encuesta sobre Desarrollo tecnológico en el establecimiento Industrial Colombiano, 1995

Las dos técnicas estudiadas serán aplicadas a los datos de la Encuesta sobre Desarrollo tecnológico en el establecimiento Industrial Colombiano, 1995, realizada en el año 1996.

4.1. Encuesta sobre Desarrollo tecnológico en el establecimiento Industrial Colombiano

Información obtenida por muestreo aplicado por entrevista directa a 885 establecimientos industriales de todo el país. La muestra tiene cobertura geográfica de ámbito nacional y es representativa a nivel de agrupaciones industriales según la clasificación CIIU (Código Industrial Internacional Uniforme). Se adoptó como marco muestral el directorio de establecimientos de la Encuesta Anual Manufacturera del DANE (Departamento Administrativo Nacional de Estadística), de 1991, por ser la última de la cual se tiene información disponible. Se decidió excluir de la muestra las empresas con menos de 10 empleados, teniendo en cuenta que a este nivel el marco presenta mayor grado de subregistro.
La encuesta fue diseñada por el Departamento de Planeación, por la división de desarrollo tecnológico. Los documentos (manuales, levantamiento de la información, y las bases de datos), estuvieron a cargo de la firma consultora Sistemas Especializados de Información, S.E.I. S.A.
El mayor inconveniente que se presentó en la realización de esta encuesta fue el nivel

de desagregación de la información en los establecimientos, por esto esta está muy incompleta.
La unidad de selección y observación es el establecimiento industrial, definido como: "la unidad económica que, bajo una forma jurídica única o un solo propietario y en una sola ubicación física, se dedica a la producción del grupo más homogéneo posible de bienes facturados"[1]

4.1.1. Contenido de la Encuesta

La encuesta consta de 140 preguntas agrupadas en nueve capítulos, divididos en secciones. Los capítulos se observan en la Tabla 4.1. Cuando el establecimiento se niegue a responder o no conozca la información se anotará el código "999999" como respuesta, (también puede encontrarse "9", "999",...)

Tabla 4.1: *Contenido de la Encuesta*

Capítulo	Periodo de referencia	A quien se aplica el capítulo
I– Identificación del Establecimiento	Fecha de la entrevista	Todos los establecimientos de la muestra
II - Desempeño Económico del Establecimiento durante 1995	01/01/1995 a 31/12/1995	Todos los establecimientos de la muestra
III - Caracterización de la Dinámica Tecnológica	Enero 1/1993 a la fecha	Todos los establecimientos de la muestra
IV - Tipificación de la Innovación Tecnológica en el Establecimiento	Enero 1/1989 a la fecha	Establecimientos con actividades Innovativas en período 89-96
V - Actividades Innovativas y de Desarrollo Tecnológico	Enero 1/1989 a la fecha	Establecimientos con actividades Innovativas en período 89-96
VI - Proyectos de Investigación y Desarrollo	Enero 1/1989 a la fecha	Establecimientos con proyectos de I+D en período 89-96
VII - Capacitación Tecnológica	1993 a 1995	
VIII - limitaciones de la Innovación y Perspectivas Futuras	Fecha de la entrevista	Todos los establecimientos de la muestra
IX - Sistemas Nacionales de Propiedad Industrial y Metrología. Normalización y Calidad	Fecha de la entrevista	Establecimientos que contestaron positivamente a la pregunta 504

4.2. Encuestas de Innovación

La competitividad de las empresas está dada por la capacidad de mantener ventajas que le permitan alcanzar y mantenerse en el mercado, cualquiera sea su razón de ser. Tal ventaja se da en la medida que las empresas ofrezcan productos o servicios los cuales escaseen en sus competidores. La empresa debe orientarse, por completo, en la búsqueda de la competitividad, diseñando estrategias encaminadas a este objetivo. La Competitividad es el resultado de una mejora de calidad constante y de innovación, por ende, esta última ha sido un amplio objeto de estudio.

[1]Tomado de (Colciencias et˜al., 1996)

El concepto de innovación es atribuido a Joseph Schumpeter (1939) quien en su propuesta de desarrollo económico determinó que las empresas pueden estar en dos estados, un estado de no crecimiento (circuito) o un estado de crecimiento (evolución) y que para pasar del primer estado al otro es necesario realizar innovaciones (Suárez, 2004). Schumpeter define la innovación como el arte de convertir las ideas y el conocimiento en productos, procesos o servicios nuevos, o mejorados que el mercado reconozca y valore. Luego, la innovación consiste no solo en nuevos productos y procesos, sino también, en nuevas formas de organización, nuevos mercados, nuevas estrategias de comercialización, (Currie & Harris, 2005).

Posterior a la definición de innovación propuesta por Schumpeter y muchas otra definiciones, aparece el Manual de Oslo (1992), cuyo objetivo es proporcionar directrices para la recogida e interpretación de información relativa a innovación, con el ánimo de recolectar datos internacionalmente comparables (Sánchez & Castrillo, 2006), la tercera edición de este manual fue publicada en Octubre de 2005 y se realizó dada la necesidad de incorporar a esta medición el sector servicios, para lo que las ediciones anteriores no estaban preparadas. El manual de Oslo utiliza una definición de innovación más amplia que incluye Innovación tecnológica: innovación en tecnologías de productos y procesos (TPP), innovaciones organizacionales y de marketing.

> **Definición:** Innovación es la implementación de un producto (bien o servicio) o proceso nuevo o con un alto grado de mejora, o un método de comercialización u organización (OCDE, 2005).

De donde se tiene que[2]:

- Una innovación de producto es la introducción de un bien o servicio nuevo o con un alto grado de mejora, respecto a sus características o su uso deseado. Esta incluye mejoras importantes en especificaciones técnicas, componentes y materiales, software incorporado, ergonomía u otras características funcionales.
- Una innovación de proceso es la implementación de un método de producción o distribución nuevo o con un alto grado de mejora. Esta incluye mejoras importantes en técnicas, equipo y/o software.
- Una innovación de marketing es la implementación de un nuevo método de comercialización que entraña importantes mejoras en el diseño del producto o en su presentación, o en su política de emplazamiento (posicionamiento), promoción o precio.

[2]Tomado de OCDE (2005)

- Una innovación organizacional es la implementación de un nuevo método de organización aplicado a las prácticas de negocio, al lugar de trabajo o a las relaciones externas de la empresa.

Otras dos definiciones más simples que la propuesta por el manual de Oslo son:

- **Livingstone C.** La innovación es un proceso mediante el cual las ideas son transformadas a través de actividades económicas en resultados generadores de valor.
- **Conference Board of Canada** se encuentra que la innovación es un proceso mediante el cual se extrae valor económico del conocimiento a través de la generación, desarrollo y aplicación de ideas en la producción de nuevos productos, procesos y Servicios[3].

4.2.1. Innovación en Colombia

La necesidad de conocer el estado actual de un país con respecto a indicadores de innovación condujo a construir encuestas de innovación que ayudaran a determinar tal estado y la dirección en que se estaba encaminado. La importancia de tales encuestas está fundamentada en la escasez de datos sobre innovación y el insuficiente monitoreo y evaluación de políticas. Su objetivo es estimular la investigación del comportamiento de las empresas innovadoras.

La primera encuesta nacional de Innovación y Desarrollo tecnológico en el establecimiento industrial colombiano, elaborada por el Departamento Nacional de Planeación (DNP) y el Instituto Colombiano para el desarrollo de la Ciencia y la Tecnología "Francisco José de Caldas", Colciencias, en el año 1996, fue la primera realizada en Colombia. La ejecución de esta encuesta permitió realizar, por primera vez, una clasificación de las empresas del sector manufacturero.

El objetivo de este capítulo es determinar las caracterésiticas de los establecimientos Colombianos en cuanto a innovación utilizando para ello los datos de la primera encuesta de Innovación y Desarrollo Tecnológico y dos técnicas de clasificación, los árboles de clasificación y la regresión logística, para ello se seleccionó un grupo de variables[4] a ser tomadas como explicativas para cada tipo de innovación, estas variables están relacionadas con:

⋆ Código del tamaño del establecimiento, CIIU3, 28 en total.

[3]Ambas definiciones fueron tomadas de Salazar & Holbrook (2004)

[4]Las variables seleccionadas fueron avaladas por el profesor Jorge Robledo Velásquez, Doctor en Estudios de Política Científica y Tecnológica y Director del proyecto Descubrimiento de Conocimiento de la innovación en Colombia

- ⋆ Fuentes internas de la innovación (5 variables) y fuentes externas de la innovación (10 variables).
- ⋆ Naturaleza jurídica del establecimiento (12 niveles).
- ⋆ Inversión bruta en maquinaria, Inversión bruta total, valor de las exportaciones, valor de las utilidades (o pérdidas) sobre las ventas del establecimiento en 1995.
- ⋆ Porcentaje de empleo calificado en producción. Encargados de la ejecución de las actividades innovativas al interior de la empresa.

Con las variables mencionadas, utilizadas como variables explicativas se construyeron los modelos de clasificación, utilizando como variables respuesta cuatro diferentes tipos de innovación tecnológica, definidos en la EDTI (Colciencias et˜al., 1996).

- **Innovación de productos:** la adquisición, asimilación o imitación de nuevas tecnologías para mejorar tecnológicamente productos, para comenzar a producir productos que no existían en la empresa, y/o innovar productos no existentes en el mercado.
- **Innovación de procesos:** la adquisición, asimilación o imitación de nuevas tecnologías para mejorar tecnológicamente procesos productivos existentes en la empresa, para comenzar a utilizar procesos que no existían en la empresa, y/o innovar procesos inexistentes en el mercado.
  Nota: Se entiende por mejora tecnológica de un producto o proceso el desarrollo de un producto o proceso existente con mejoras sustanciales en los beneficios generados o en su desempeño.
- **Cambios en las formas de organización y administración:** esto incluye cambios tanto en la organización del proceso productivo, como en la organización y gestión del establecimiento en general que implican cambios radicales en las estrategias corporativas, basados fundamentalmente en la posibilidades abiertas por las nuevas tecnologías informáticas.
- **Cambios en el empaque y embalaje:** se refiere a todo tipo de mejora que se introduzca en el embalaje o envoltorio del producto final de la firma, que no altera sustancialmente las propiedades del mismo. Por ejemplo, pasar de una presentación de leche en caja sin troquel a otra con troquel.

4.3. Resultados

4.3.1. Innovación de Producto

En la Figura 4.1 se muestra el árbol de clasificación obtenido a partir de las variables mencionadas anteriormente, el árbol clasifica los establecimientos en

innovadores de producto o no, se utiliza el paquete estadístico R (2007). Se puede notar que según esta clasificación un establecimiento es innovador de producto si cumple con dos condiciones:

- La **fuente de la innovación son los directivos** del establecimiento: Si la fuente interna de la innovación son los directivos del establecimiento, se puede afirmar que el establecimiento es innovador de producto, en este grupo se ubicaron 458 de los 747 establecimientos de la muestra, de los cuales ninguno quedó mal clasificado.
- La **fuente de la innovación son los clientes** del establecimiento: Para el caso en que la innovación no proviene de los directivos del establecimiento, entonces hay que revisar si la fuente externa de la innovación son los clientes, de modo que si la fuente de la innovación proviene de los clientes entonces se tiene establecimientos innovadores de producto. Si la Fuente de la innovación no son los directivos del establecimiento y tampoco son los clientes, entonces se tiene establecimientos que no son innovadores.

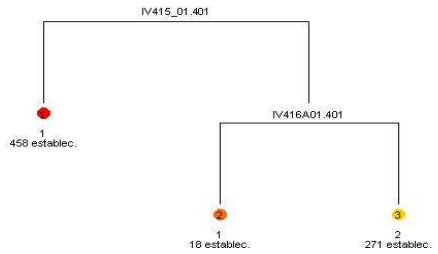


Figura 4.1: *Árbol de clasificación Innovación de producto, organizacional y, empaque y embalaje*

4.3.2. Innovación de Proceso

La innovación de proceso consiste en verificar si el establecimiento a realizado o no mejoras tecnológicas en los procesos. En la Figura 4.2 se observa el árbol obtenido, de donde se concluye que la innovación de proceso en los establecimientos colombianos depende de cuatro variables:

- La **fuente de la innovación son los directivos o los clientes** del establecimiento: Si la fuente interna de la innovación son los directivos del establecimiento, este último es innovador de proceso, de lo contrario la fuente de la innovación son los clientes. Pero si la innovación no proviene de los clientes o los directivos entonces,

- El encargado de la **ejecución de actividades innovativas** es un grupo de trabajo creado para la solución de un problema específico Si el establecimiento pertenece al sector Bebidas, textiles, papel o derivados, imprentas y editoriales, productos químicos, metales no ferrosos, maquinaria y aparatos eléctricos o, equipo profesional y científico y además la innovación es llevada a cabo por un grupo específico de trabajo entonces hay innovación de proceso, de lo contrario, el establecimiento no es innovador.

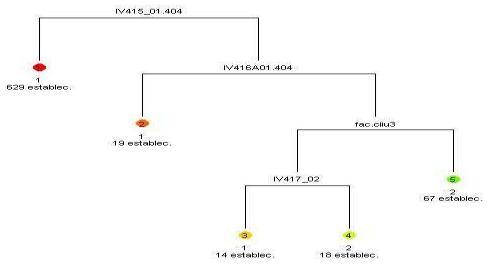


Figura 4.2: *Árbol de clasificación Innovación de proceso*

4.3.3. Innovación Organizacional

Por último, se muestra en la Figura 4.1 la clasificación para los establecimientos que han implementado cambios en la gestión y administración del negocio. Nuevamente se obtiene que la innovación debe provenir de los directivos del establecimiento, de lo contrario de los clientes.

4.3.4. Innovación en empaque y embalaje

Se obtiene el mismo árbol de clasificación de la Figura 4.1, se observa la misma estructura luego, se concluye que para que cualquier establecimiento sea innovador, la fuente de la innovación ha de provenir de los directivos o los clientes del establecimiento.

Los resultados encontrados son coherentes con estudios previos realizados sobre la encuesta, es decir, en general los establecimientos innovadores dependen en un 95 % de sus directivos. Mayor información sobre los resultados de la EDT1 pueden encontrarse en Vargas & Malaver (2004), Durán et˜al. (1998), Durán et˜al. (2000).

4.3.5. Regresión Logística

El modelo de regresión logística es uno de los más aplicados y uno de sus mayores problemas es el de la separación que trae como consecuencia la no existencia de los estimadores de máxima verosimilitud, pues el proceso iterativo para la obtención de los mismos no converge y por tanto, no se pueden realizar inferencias. Al aplicar la regresión logística para el caso de las variables seleccionadas en la Primera Encuesta de desarrollo tecnológico en el establecimiento industrial colombiano se presenta éste problema, conocido como "Separación completa" o "Separación Cuasicompleta" (Prieto˜Castellanos, 2005) y el modelo de regresión logística no converge, luego no se puede obtener resultados para ninguno de los modelos presentados anteriormente.

Para el caso de la *Primera Encuesta*, los árboles de clasificación obtienen un resultado de clasificación que es consistente con análisis realizadas anterioremente a la encuesta, mientras que por el problema de la *Separación Completa* los estimadores de máxima verosimilitud de la regresión logística no son confiables, luego la clasificación no se puede obtener.

CAPÍTULO 5

Conclusiones y Recomendaciones

En general, se observó que cuando se tiene igual separación entre los grupos las Tasas de Mala Clasificación (TMC), de los Árboles de Clasificación y la Regresión Logística, cambian muy poco al incrementar la correlación entre las variables explicativas, además se nota que la TMC se reduce al incrementar el tamaño de la muestra.

La regresión logística presenta siempre una TMC más baja que los Árboles de Clasificación, exceptuando el caso donde la matriz de varianzas y covarianzas poblacional de uno de los grupos es cuatro veces mayor que la del otro.

Al incrementar la separación entre los grupos la regresión logística evidencia el problema de separación completa al no converger el algoritmo de estimación, mientras que los Árboles de Clasificación presentan una clasificación perfecta.

Cuando se compara la clasificación para grupos con igual tamaño de muestra pero incrementando la separación entre los mismos, se nota una clara reducción en las TMC, pero de igual manera, al incrementar la correlación, la TMC para la Regresión Logística es menor que para los Árboles de Clasificación.

Cuando se consideran grupos desbalanceados, el problema de la separación completa en la Regresión Logística se hace más evidente, al obtener TMC que no varían bajo ningún cambio en correlaciones o tamaños de muestras. Para éste caso, nuevamente los Árboles de Clasificación presentan clasificación perfecta.

Al comparar las TMC en grupos provenientes de la distribución normal sesgada

con los demás casos (grupos normales con igual estructura de covarianzas, grupos normales con diferentes estructura de covarianza y grupos lognormales), se nota una reducción en las TMC, que se hace más evidente a medida que se aumenta el sesgo.

la Regresión Logística presentó una Tasa de Mala Clasificación más baja que los Árboles de Clasificación, situación aún más evidente al tener grupos donde las variables tienen correlaciones altas (0,7 y 0,9) y los tamaños de muestra son pequeños (20 observaciones). Las tasas de Mala Clasificación de los Árboles disminuyen y se acercan más a las de la Regresión Logística cuando las variables tienen correlación alta y los tamaños de muestra son mayores (50 y 100 observaciones).

APÉNDICE A

Distribuciones de los datos simulados

A.1. Distribución Normal

Distribución Normal Univariada, (Casella & Berger, 2001)

$$f\left(x;\mu,\sigma\right)=\frac{1}{\sqrt{2\pi\sigma}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Distribución Normal Multivariable, (Seber, 1938). La variable $Y\sim N_p\left(\mu,\Sigma\right)$ si

$$f\left(y;\mu,\Sigma\right)=\frac{1}{\left(2\pi\right)^{p/2}\left|\Sigma\right|^{1/2}}e^{-\frac{1}{2}(y-\mu)^T\Sigma^{-1}(y-\mu)}$$

Para el caso bivariado se tiene

$$f(y_1,y_2;\mu,\Sigma)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\times\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{y_1-\mu_1}{\sigma_1}\right)^2-2\rho\frac{(y_1-\mu_1)(y_2-\mu_2)}{\sigma_1\sigma_2}+\left(\frac{y_2-\mu_2}{\sigma_2}\right)^2\right]\right\}$$

A.2. Distribución Log-normal

Distribución Lognormal Univariada,

$$f\left(x;\mu,\sigma\right)=\frac{1}{x\sigma\sqrt{2\pi}}e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}}$$

Transformación - Lognormal

Sea $X = [X_1, X_2, \cdots, X_p]$ un vector de p componentes distribuidos normal multivariable con media μ y matriz de covarianzas Σ, usando la transformación $Y_i = \exp(X_i)$, defina $Y = [Y_1, Y_2, \cdots, Y_p]$. La densidad de Y es una distribución lognormal multivariada, (Tarmast, 199).

Distribución Lognormal bivariable

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} \log(X_1) \\ \log(X_2) \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} \right)$$

y en general, la distribución Lognormal multivariable es:

$$f(x_1, ..., x_p; \mu, \Sigma) = (2\pi)^{-p/2} |\Sigma|^{-\frac{p}{2}} \left(\prod_{i=1}^{p} \frac{1}{x_i}\right) \exp\left\{ -\frac{1}{2} \left[(\log x_1, \ldots, \log x_p) - \mu\right]' \Sigma^{-1} \left[(\log x_1, \ldots, \log x_p) - \mu\right] \right\}$$

A.3. Distribución Normal Sesgada

Azzalini & Dalla_Valle (1996) trabajan en la llamada Distribución Normal Sesgada, introducen la familia paramétrica multivariada tal que las densidades marginales son normal-sesgada escaladas, y estudian sus propiedades, con especial énfasis en el caso bivariado. La variable $Z \sim SN(\lambda)$ si su función de densidad es:

$$\phi(z; \lambda) := 2\phi(z)\,\Phi(\lambda z), \quad (z \in \mathcal{R})$$

donde $\phi(z)$ y $\Phi(z)$ denotan la función de densidad y la distribución normal, respectivamente. El parámetro λ regula el sesgo y está definido en $(-\infty, \infty)$.
En el paquete estadístico R (2007), Azzalini desarrolló la librería *sn* (Azzalini, 2008) donde presenta el desarrollo computacional de la distribución, funciones para el caso univariado y multivariado e incluye funciones de graficación (Pérez, 2008).
En la figura A.1 se observan los contornos de la normal sesgada para los valores del parámetro de sesgo seleccionados (1,1), (1,5), (1,10) y (1,20).
Es claro que cuando el parámetro de sesgo es $(0, 0)$ la distribución corresponde a una normal bivariada, como se muestra en la Figura A.2.

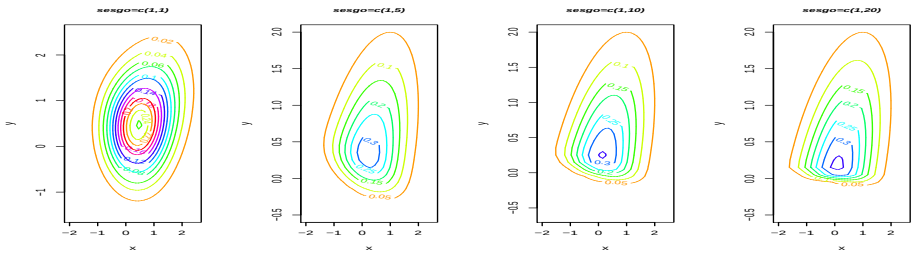


Figura A.1: *Contornos de la distribución normal sesgada bivariada, para diferentes parámetros de sesgo*

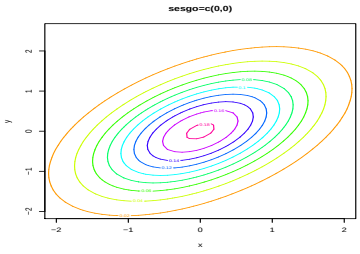


Figura A.2: *Contorno de la distribución Normal bivariada*

APÉNDICE B

Resultados adicionales

B.1. Caso2, $2\Sigma_1 = \Sigma_2$

Para el **Caso2** donde la varianza de uno de los grupos es mayor ($\Sigma_2 = 2\Sigma_1$), el comportamiento de la TMC en la Figura B.1 es muy similar al **Caso1**, al aumentar el tamaño de muestra la tasa de mala clasificación es, en general la misma, de igual manera al incrementar la correlación de las variables explicativas y la separación entre los grupos, especialmente para la Regresión Logística. Para los Árboles de Clasificación, la reducción en la TMC al incrementar la correlación, es mas notoria para muestras grandes.

Al tomar grupos desbalanceados para el **Caso2** como se muestra en la Figura B.2, nuevamente se observa el problema en los estimadores para el caso de separación completa en la Regresión Logística. Al aumentar el desbalance en los grupos y la correlación entre las variables, la TMC se reduce.

B.2. Caso2, $4\Sigma_1 = \Sigma_2$

En la Figura B.3 se observa la Tasa de Mala Clasificación para datos generados de una distribución normal donde la matriz de varianzas y covarianzas de uno de los grupos es cuatro veces más que la del otro grupo ($\Sigma_2 = 4\Sigma_1$). En ésta Figura se nota que para tamaños de muestra de 20 y 50, y poca separación entre los grupos la TMC tiende a disminuir a medida que se incrementa la correlación entre los grupos y en general, se nota una TMC menor para los árboles. Cuando se tiene una muestra de tamaño 100 en cada grupo, la TMC se mantiene constante, observándose más alta

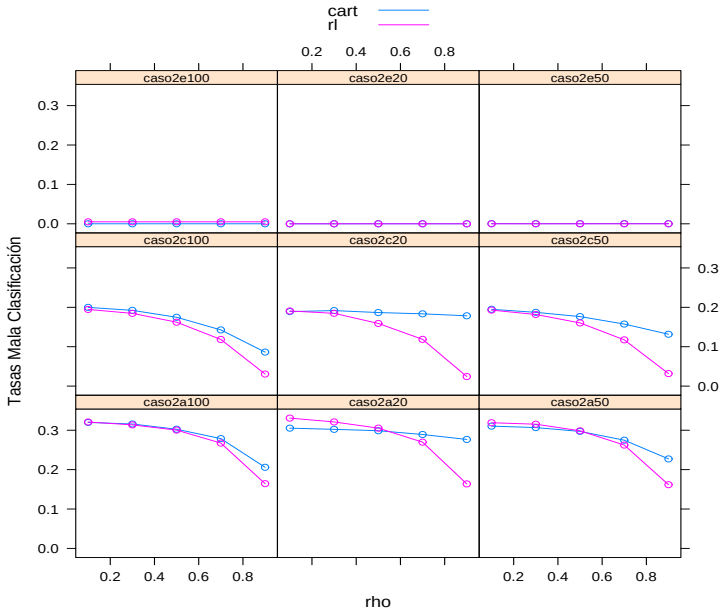


Figura B.1: *Caso2,* $2\Sigma_1 = \Sigma_2$

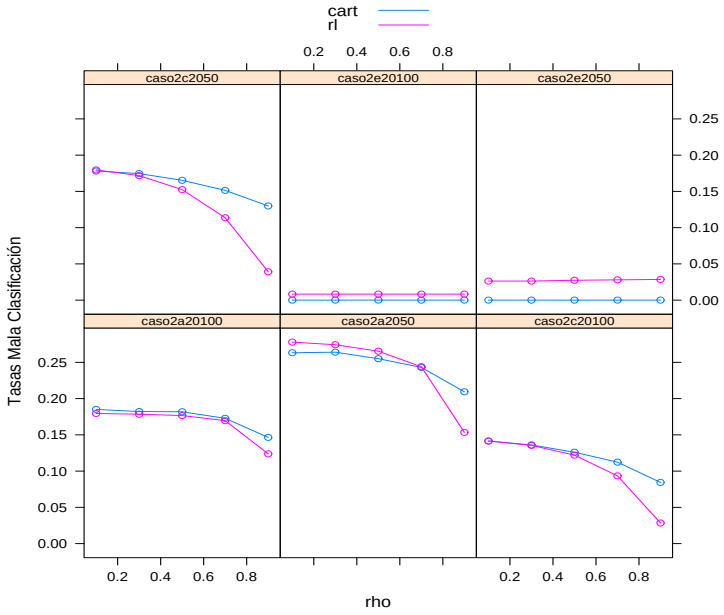


Figura B.2: *Caso2,* $2\Sigma_1 = \Sigma_2$, *muestras desbalanceadas*

para la clasificación logística. De igual manera al incrementar la separación entre los grupos la TMC se reduce.
En la Figura B.4 se observa un comportamiento bastante particular, pues al tener muestras desbalanceadas pareciera que la correlación entre las variables explicativas no influyera sobre la clasificación. Al observar la Figura B.5 donde se presentan los contornos de las dos distribuciones de las cuales se obtuvieron los datos, en los dos primeros paneles las distribuciones se solapan y en el último los grupos están completamente separados (los tres casos de separación para el caso presentado en las Figuras B.3 y B.4 para una correlación de 0,5).

B.3. Caso3, Distribución Lognormal

El **Caso3**, donde se toman grupos generados a partir de datos de la distribución *lognormal*, las TMC obtenidas se observan en la Figura B.6 donde el comportamiento es parecido al de los casos anteriores, cuando se tiene tamaño de muestra $n = 20$, para los Árboles de Clasificación, la TMC cambia muy lentamente al incrementar la correlación entre las variables o la separación entre los grupos. Al incrementar el tamaño de muestra,la correlación y la separación entre los grupos la TMC se reduce. Cuando la separación entre los grupos es grande, la clasificación es perfecta.
En la Figura B.7, donde se consideró muestras desbalanceadas para la distribución lognormal, se observa nuevamente mejores tasas de clasificación para la regresión logística a medida que se incrementa la correlación entre las variables. Cuando se aumenta la separación entre los grupos y el desbalance entre las muestras, la TMC se reduce. Cuando la separación de los grupos es grande (panel con la letra e) se presenta clasificación completa en la Regresión Logística y para los árboles la clasificación es casi perfecta.

B.4. Caso4, Distribución Normal Sesgada

En las figuras de esta sección: Figura B.8, B.9,B.10, B.11, B.12, B.13, B.14 y B.15, se observan los resultados de la simulaciones para la distribución normal sesgada y los diferentes parámetros de sesgo considerados, SN(1,1), SN(1,5), SN(1,10) y SN(1,20). En general se observa el mismo comportamiento descrito en los tres escenarios anteriores.

Al incrementar el sesgo de la distribución se nota una reducción alrededor del 10 % en las Tasas de Mala Clasificación, de igual manera al incrementar el tamaño

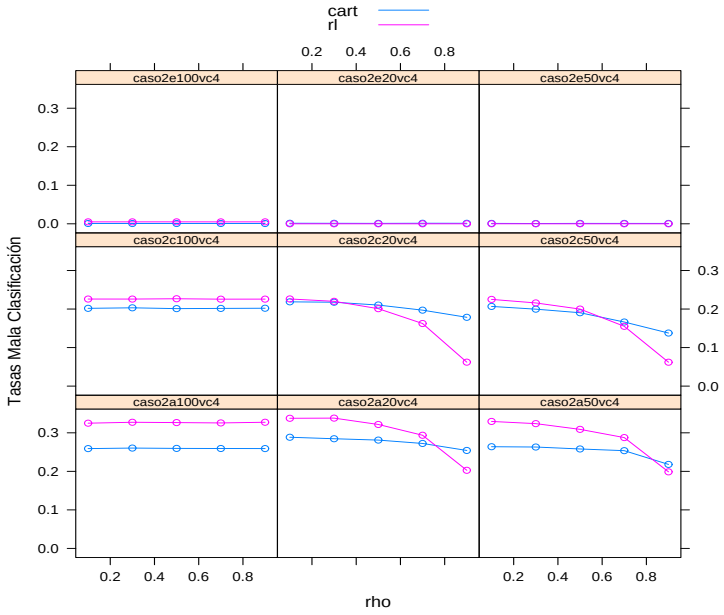


Figura B.3: *Caso2,* $4\Sigma_1 = \Sigma_2$

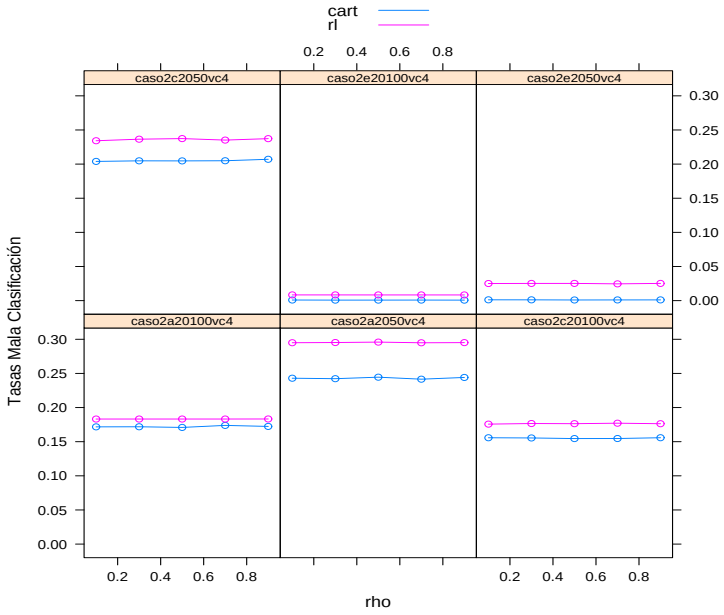


Figura B.4: *Caso2,* $4\Sigma_1 = \Sigma_2$, *muestras desbalanceadas*

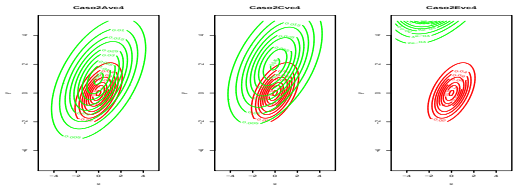


Figura B.5: *caso2, Contornos de la distribución normal para* $4\Sigma_1 = \Sigma_2$

de la muestra y/o la correlación, las Tasas de Mala Clasificación se reducen.

Las Tasas de Mala Clasificación son, en general, más bajas que en cualquiera de los casos anteriores y este reducción es mayor a medida que se incrementa el sesgo de la distribución, para muestras desbalanceadas se nota, nuevamente una reducción en la TMC al incrementar el desbalance de las muestras.

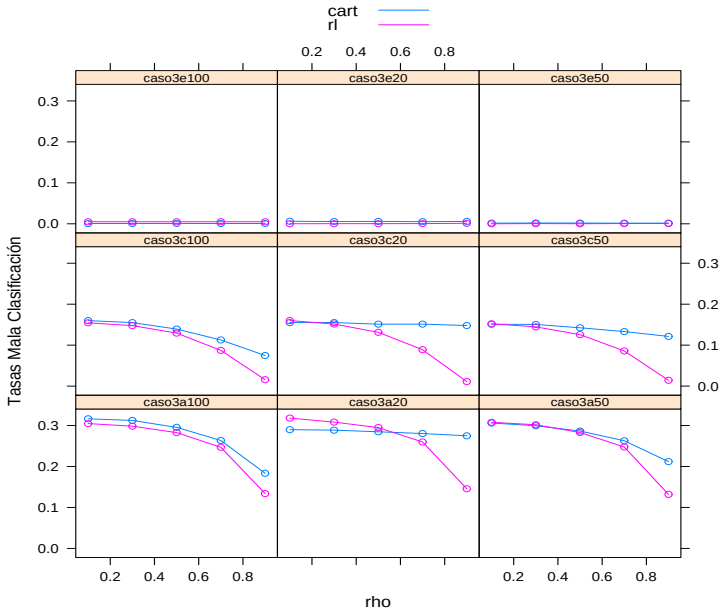


Figura B.6: *Caso3, distribución lognormal*

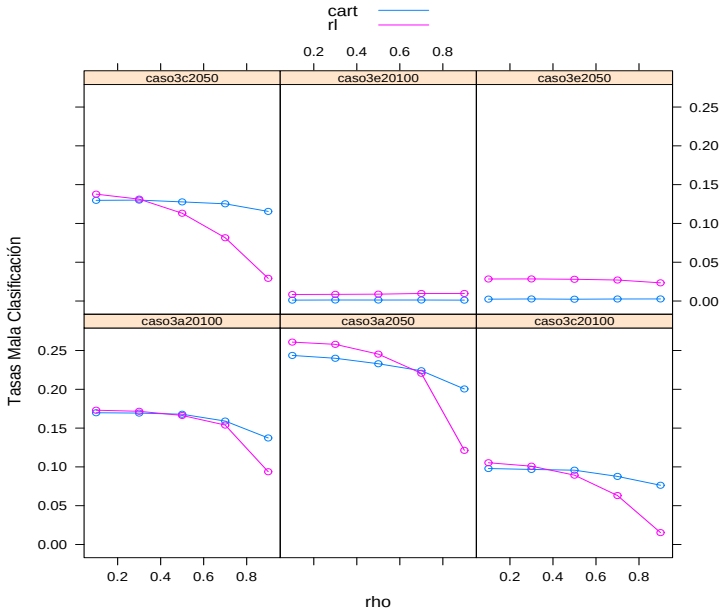


Figura B.7: *Caso3, distribución lognormal, muestras desbalanceadas*

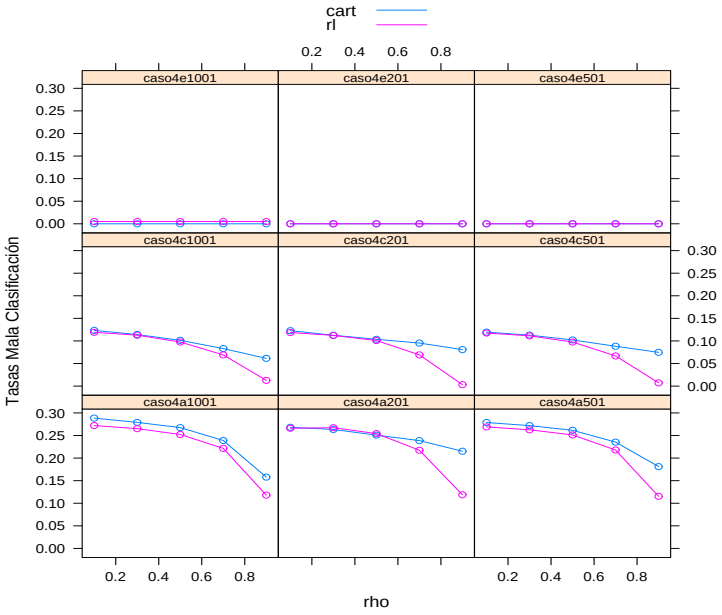


Figura B.8: *Caso4, Distribución normal sesgada,* $SN(1,1)$

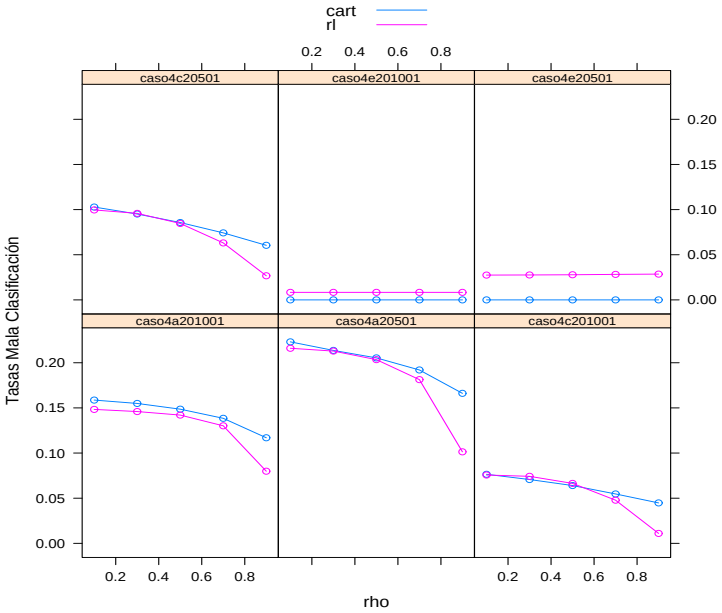


Figura B.9: *Caso4, Distribución normal sesgada,* $SN(1,1)$*, muestras desbalanceadas*

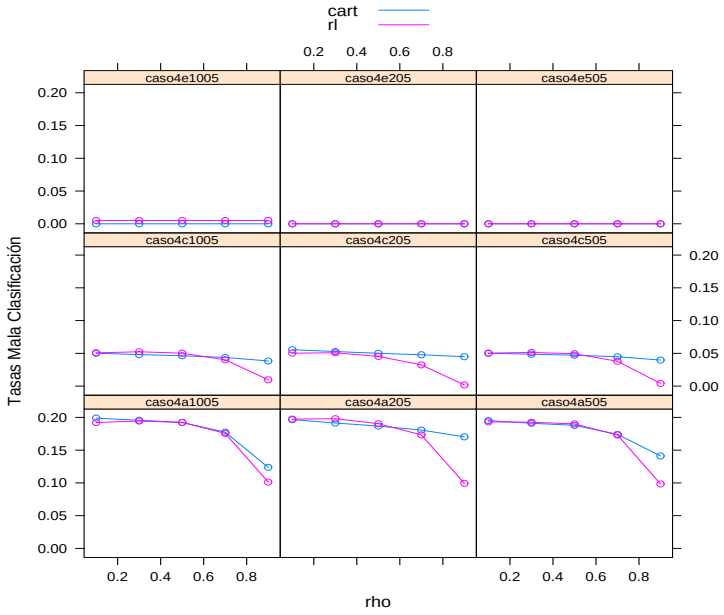


Figura B.10: *Caso4, Distribución normal sesgada,* $SN(1, 5)$

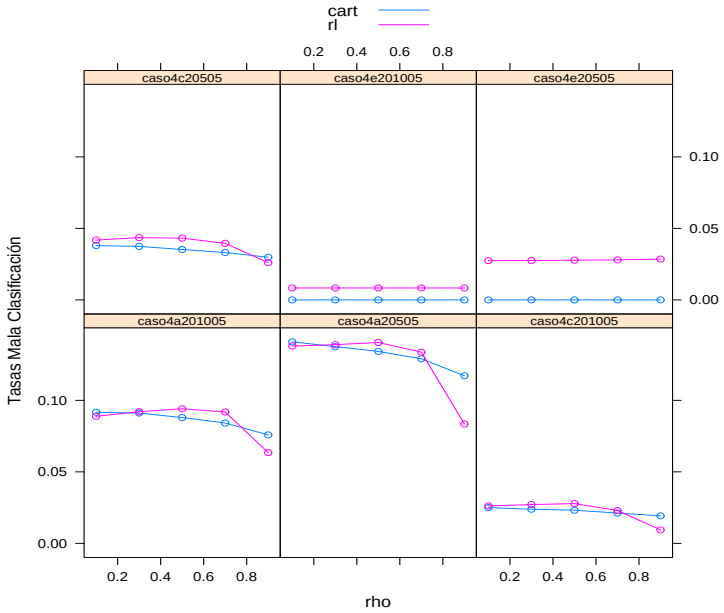


Figura B.11: *Caso4, Distribución normal sesgada,* $SN(1, 5)$ *muestras desbalanceadas*

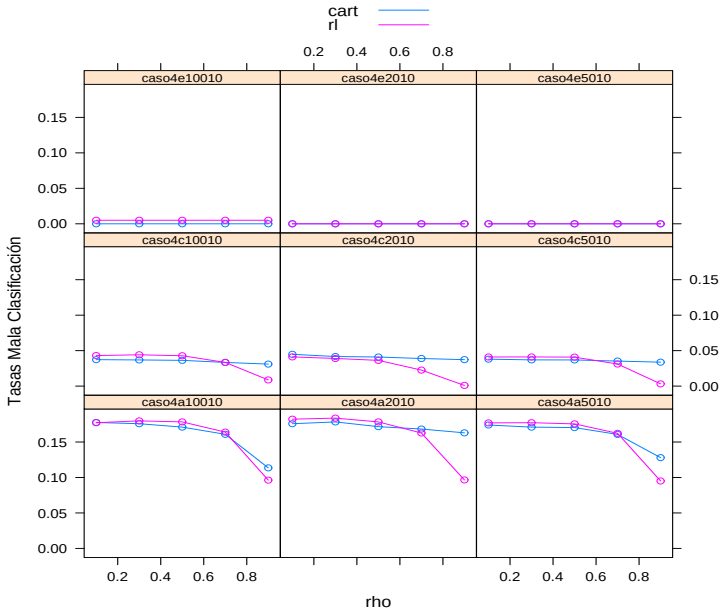


Figura B.12: *Caso4, Distribución normal sesgada,* $SN(1, 10)$

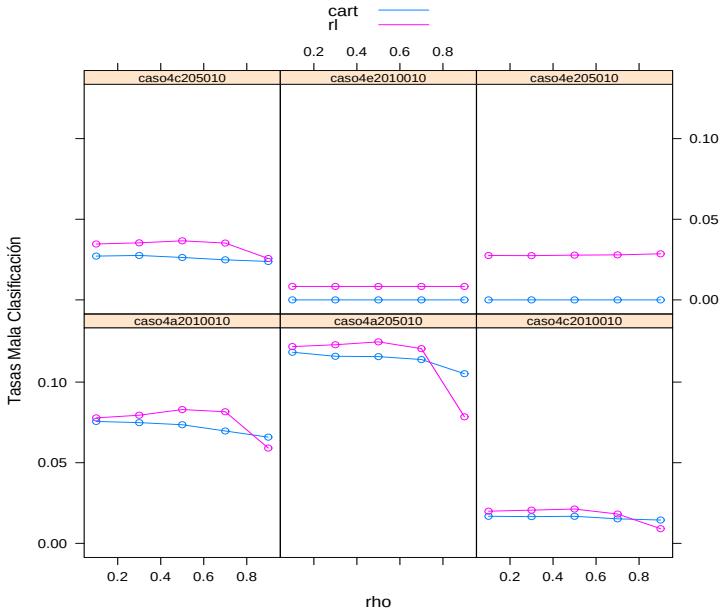


Figura B.13: *Caso4, Distribución normal sesgada,* $SN(1, 10)$*, muestras desbalanceadas*

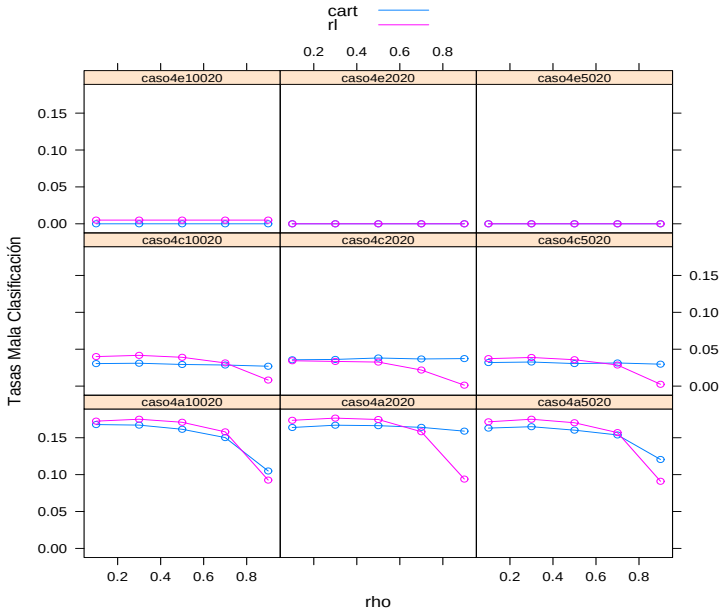


Figura B.14: *Caso4, Distribución normal sesgada,* $SN(1, 20)$

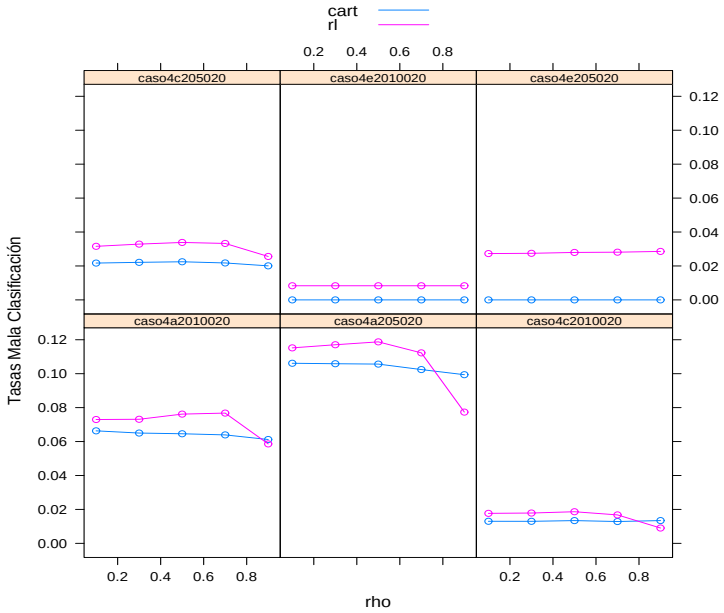


Figura B.15: *Caso4, Distribución normal sesgada,* $SN(1, 20)$*, muestras desbalanceadas*

APÉNDICE C

Programa R

```
######################### SIMULACIONES PARA LAS DOS METODOLOGIAS
require(MASS)
require(rpart)
##
simulaciones<-1000
## variables finales
tasas.final<-matrix(NA,simulaciones,10)
##ciclo
for (Nsim in 1:simulaciones){
## Funcion para obtener la muestra de cada correlacion
rho<-c(0.1,0.3,0.5,0.7,0.9)
muestras<-function(rho){
tot.muestra<-matrix(NA,ncol=15,nrow=40)
for(k in 1:length(rho)){#para rocorrer las correlaciones
muestra1<-mvrnorm(n=20,rep(0,2),matrix(c(1,rho[k],rho[k],1),2,2))
muestra2<-mvrnorm(n=20,c(0,10),matrix(c(1,rho[k],rho[k],1),2,2))
poblacion1<-rbind((cbind(muestra1,c(0))),(cbind(muestra2,c(1))))
muestra<-poblacion1
if(k==1) tot.muestra<-muestra
else tot.muestra<-cbind(tot.muestra,muestra)
tot.muestra
}
return(tot.muestra)
}
```

```
poblas<-muestras(rho)
### tot.muestra es una matriz de 40 x 15 y cada poblacion son tres columnas
##
poblacion1<-poblas[,1:3]
poblacion2<-poblas[,4:6]
poblacion3<-poblas[,7:9]
poblacion4<-poblas[,10:12]
poblacion5<-poblas[,13:15]
poblaciones<-list(poblacion1,poblacion2,poblacion3,poblacion4,poblacion5)
##
for(j in 1:length(poblaciones)){
#### separacion de las muestras
pobla<-poblaciones[j]
poblacion1<-matrix(unlist(pobla),ncol=3)
## son argumentos de la funcion
x1<-poblacion1[,1]#muestra 1
x2<-poblacion1[,2]#muestra 2
y.1<-poblacion1[,3]#variable respuesta
#### CALCULO DE LOS VALORES PREDICHOS POR VALIDACION CRUZADA
############ LEAVE ONE OUT (LOO) ############
loo.predichos<-function(i,y,x1,x2){
mod.logi2<-glm(y~x1+x2, family="binomial")
beta0<-mod.logi2$coefficients[1]
beta1<-mod.logi2$coefficients[2]
beta2<-mod.logi2$coefficients[3]
y.temp<-y.1[-i]
x1.temp<-x1[-i]
x2.temp<-x2[-i]
# ensayar dando los valores iniciales
mod.logi<-glm(y.temp~x1.temp+x2.temp, family="binomial",start=c(beta0,beta1,beta2)
mod.cart<-rpart(y.temp~x1.temp + x2.temp,parms=list(split="gini"))
dato.nuevo<-data.frame(x1.temp=x1[i], x2.temp=x2[i])
pred.logi<-predict(mod.logi, dato.nuevo, type='response')
pred.cart<-predict(mod.cart,newdata=dato.nuevo) #
predicho<-cbind(pred.logi,pred.cart)
return(predicho)
}
pred<-apply(matrix(1:length(y.1),ncol=1),1,loo.predichos,y.1,x1,x2)
predichos<-t(unlist(pred))# ;class(predichos);dim(predichos)
#### POBLACION SIMULADA 2 PARA VALIDACI'ON
####### JUNTO CON LOS PREDICHOS PARA LA CLASIFICACI'ON
```

```
clas.pto.corte<-function(corte,predi.logi,predi.cart,observado){
punto.corte.rl<-quantile(predi.logi,probs=1-corte)##
punto.corte.ct<-quantile(predi.cart,probs=1-corte)
y.predi.logi<-ifelse(punto.corte.rl<predi.logi,1,0)
y.predi.cart<-ifelse(punto.corte.ct<predi.cart,1,0)
tabla.logi<-table(observado,y.predi.logi)
adic<-matrix(c(0,0),ncol=1)
tabla.logi<-cbind(tabla.logi,adic)#dim de la tabla siempre 2 2
tabla.cart<-table(observado,y.predi.cart)
tabla.cart<-cbind(tabla.cart,adic)
clasifi.logi<-(tabla.logi[1,1]+tabla.logi[2,2])/sum(tabla.logi)
clasifi.cart<-(tabla.cart[1,1]+tabla.cart[2,2])/sum(tabla.cart)
clasifi<-cbind(clasifi.logi,clasifi.cart)
return(clasifi)#tasas de clasificacion correcta
}
corte<-as.matrix(seq(0.05,0.99,length=20))

cortes.clasi<-apply(corte,1,clas.pto.corte,predichos[,1],predichos[,2],
poblacion1[,3])
cortes.clas<-t(unlist(cortes.clasi))##
t.cla.logi<-cortes.clas[,1]
t.cla.cart<-cortes.clas[,2]
t.claerr.logi<-1-t.cla.logi[which.max(t.cla.logi)]
t.claerr.cart<-1-t.cla.cart[which.max(t.cla.cart)]
tasa.err.rl.ct<-cbind(t.claerr.logi,t.claerr.cart)
tasa.err.rl.ct1<-tasa.err.rl.ct#desde la flecha no va
#para guardar los res de cada poblacion
if(j==1) tasas.todas<-tasa.err.rl.ct1
else tasas.todas<-cbind(tasas.todas,tasa.err.rl.ct1)
tasas.todas
}
tasas.final[Nsim,]<-tasas.todas
tasas.final
}
### las tasas queden intercaladas primero rl y luego cart
sink('Caso 1E20.txt')
tasa.promedio<-colMeans(tasas.final)->tp
tasas.rl<-c(tp[1],tp[3],tp[5],tp[7],tp[9])
tasas.ct<-c(tp[2],tp[4],tp[6],tp[8],tp[10])
print(tasas.rl)
print(tasas.ct)
```

```
nombre<-paste('Caso1E20','.bmp',sep='')
  bmp(file=nombre) #este lo lee el latex2e
plot(tasas.rl,ylim=c(min(tp),max(tp)),type='b',ylab='Misclassification Rate',
xlab='Correlaciones',axes=F)
axis(1,1:5,c(0.1,0.3,0.5,0.7,0.9))
axis(2);box()
lines(tasas.ct,ylim=c(min(tp),max(tp)),type='b',col='red',)
dev.off()
sink()
```

Referencias

Albert, A. & Anderson, J. (1986). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, (71), 1–10.

Allison, P. (1999). Logistic regression using the sas system: Theory and application. Cary, NC: SAS Institute Inc.

Azzalini, A. (2008). *R package* `sn`*: The skew-normal and skew-t distributions (version 0.4-6)*. Università di Padova, Italia.

Azzalini, A. & Dalla_Valle, A. (1996). The multivariate skew-normal distribution. *Biometrika*, (83), 715–726.

Banet, T. A. (2001). La minería de datos, entre la estadística y la inteligencia artificial. *Questiio: Quaderns d'Estadistica, Sistemes, Informatica i Investigació Operativa*, *25*(3), 479–498.

Barajas, F. H. (2007). Comparación entre análisis discriminante no-métrico y regresión logística multinomial. Tesis de Maestría, Facultad de ciencias, Universidad Nacional de Colombia.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. G. (1984). *Classification and Regression Trees*. Wadsworth International Group, Belmont, California, USA.

Caruana, R. & Niculescu-Mizil, A. An empirical comparison of supervised learning algorithms.

Casella, G. & Berger, R. L. (2001). *Statistical Inference* (2nd ed.). Duxbury Pr.

Castrillón, F. (1998). Comparación de la discriminación normal lineal y cuadrática con la regresión logística para clasificar vectores en dos poblaciones. Tesis de Maestría, Facultad de ciencias, Universidad Nacional de Colombia.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, *74*, 859–836.

Colciencias, DNP, & S.E.I.sa. (1996). *Encuesta sobre Desarrollo tecnológico en el establecimiento Industrial Colombiano*. Manual del Encuestador.

Currie, W. & Harris, J. (2005). Estrategia regional en ciencia y tecnología. Región la Araucania, Chile, Informe Final.

De'ath, G. & Fabricius, K. E. (2000). Classification and regression trees: A powerful yet simple technique for ecological data analysis. *Ecology*, *81*(11), 3178–3192.

Deconinck, E., Zhang, M. H., Coomans, D., & Heyden, Y. V. (2006). Classification tree models for the prediction of blood-brain barrier passage of drugs. *Journal of Quemical Information and Modeling*, *46*(3), 1410–1419.

Dobra, A. (2002). Classification and regression tree construction. Thesis proposal, Departament of Computer Science, Cornell University, Ithaca NY.

Durán, X., Ibáñez, R., Salazar, M., & Vargas, M. (1998). La innovación tecnológica en colombia: características por tamaño y tipo de empresa. *OCyT, Observatorio Colombiano de Ciencia y Tecnología*.

Durán, X., Ibáñez, R., Salazar, M., & Vargas, M. (2000). La innovación tecnológica en colombia: características por sector industrial y región geográfica. *OCyT, Observatorio Colombiano de Ciencia y Tecnología*.

Hadidi, N. (2003). Classification ratemaking using decision trees. CAS Forum.

Hosmer, D. & Lemeshow, S. (1989). *Applied Logistic Regression*. Wiley & Sons.

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

Kvalseth, T. O. (1985). Cautionary note about r2. *The American Statistician*, *39*(4), 279–285.

Marks, S. & Dunn, O. J. (1974). Discriminant functions when covariance matrices are unequal. *Journal of the American Statistical Association*, *69*(346), 555–559.

Menard, S. (2000). Coefficients of determination for multiple logistic regression analysis. *The American Statistician*, *54*(1), 17–24.

Mittlböck, M. & Schemper, M. (1996). explained variation in logistic regression models. *Statistics in Medicine, 15, 1987-1997.*, *15*(1987-1997).

OCDE (2005). *Oslo Manual: Guidelines for Collecting and Interpreting Innovation* (3rd ed.). OCDE publications.

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

Pérez, D. M. (2008). Estudio de la robustez del estadístico t^2 de hotelling para el caso de una y dos poblaciones cuando los datos provienen de una distribución normal sesgada. Tesis de Maestría en desarrollo.

Prieto Castellanos, K. A. (2005). Regresión logística con penalidad ridge aplicada a datos de expresión genética. Tesis de Maestría, Recinto universiatrio de Mayagüez, Universidad de Puerto Rico.

R (2007). R development core team. a language and environment for statistical computing. ISBN 3-900051-07-0.

Raveh, A. (1989). A nonmetric approach to linear discriminant analysis. *Journal of the American Statistical Association*, *84*(405), 176–183.

Rudolfer, S. M., Paliouras, G., & Peers, I. S. (1999). A comparison of logistic regression to decision tree induction in the diagnosis of carpal tunnel syndrome. *Computers and Biomedical Research*, *32*(5), 391–414.

Salazar, M. & Holbrook, J. A. (2004). A debate on innovation surveys. *Science and Public Policy*, *31*(4).

Seber, G. A. F. (1938). *Multivariate Observations*. Wiley series in probability and mathematical statistics. Jhon Wiley and Sons, Inc.

Shelley, B. & Donner, A. (1987). The efficiency of multinomial logistic regression compared with multiple group discriminant analysis. *Journal of American Statistical Association*, (82), 1118–1122.

Sánchez, M. P. & Castrillo, R. (2006). La tercera edición del manual de oslo: cambios e implicaciones. una perspectiva de capital intelectual. *madri+d*, (35), 1–16.

Suárez, O. M. (2004). Schumpeter, innovación y determinismo tecnológico. *Scientia et Technica Año X*, (25).

Tarmast, G. (199?). Multivariate log - normal distribution.

Timofeev, R. (2004). Classification and regression trees (cart). theory and applications. Master thesis, CASE - Center of Applied Statistics and Economics. Humboldt University, Berlin.

Usuga, O. (2006). Comparación entre análisis de discriminante no-métrico y regresión logística. *Proceedings of the Federal American Society of Experimental Biology*, *31*, 58–61.

Valencia, M. (2002). El problema de la separación en regresión logística. Escuela de Estadística, Universidad Nacional de Colombia - Sede Medellín.

Vargas, M. & Malaver, F. (2004). Los avances en la medición del desarrollo tecnológico en la industria colombiana. *Revista CTS*, *2*(1), 137–166.

Webb, A. R. (2002). *Statistical Pattern Recognition*. John Wiley & Sons.