CAPITULO IV

4. RESULTADOS DE LA PÁGINA

En este capítulo se expondrán los diferentes resultados de la página interactuando con los distintos métodos y opciones. Aquí se detallarán algunas características de cada método en relación al conjunto de datos al cual se apliquen. Para resaltar mejor alguna característica, en algunos casos se hará comparaciones visuales entre 2 o más métodos o entre 2 o más versiones del mismo método haciendo variar su parámetro correspondiente.

4.1 Método de los k-vecinos más cercanos

Expresividad:

Este método a diferencia de otros tiene un grado de expresividad bastante alto. Este grado de expresividad depende del valor de k, el cual disminuye cuando k aumenta. Tal como se aprecia en la figura 4.1 cuando k=1 la curva es tan flexible que no permite ninguna mala clasificación en los datos de entrenamiento (Error de entrenamiento Ee=0). En general cuando k=1 el error de entrenamiento siempre será cero.

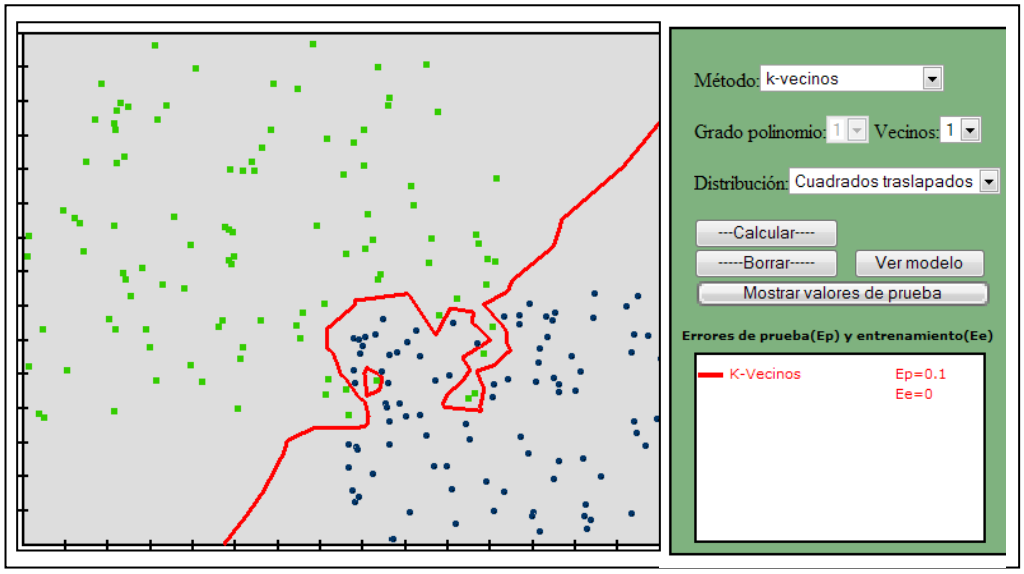

FIGURA 4.1 MUESTRA DE LA EXPRESIVIDAD DEL MÉTODO 1-VECINO MÁS CERCANO

La figura 4.2 muestra la aplicación de los k-vecinos más cercanos para k=5 al mismo conjunto de datos de la figura 4.1, como se mencionó la curva discriminante pierde expresividad pero esta pérdida de expresividad se compensa con menos sensibilidad a puntos ruidosos o anormales y menor posibilidad de sobreajuste a los datos de entrenamiento.

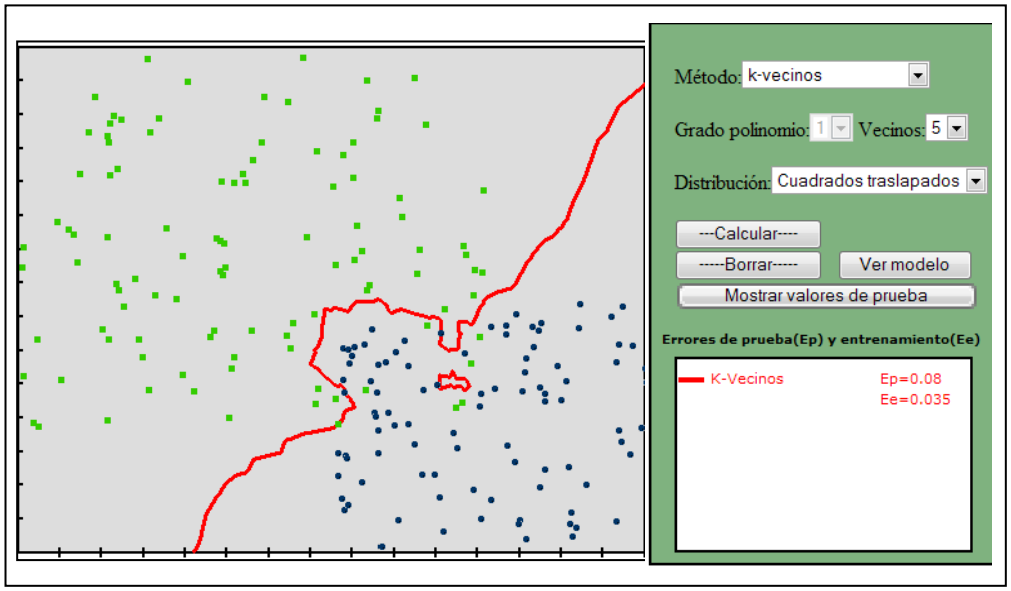

FIGURA 4.2 MUESTRA DE LA EXPRESIVIDAD DEL MÉTODO 5-VECINOS MÁS CERCANOS

A medida que k aumenta la curva discriminante tiende a regularizarse. La figura 4.3 muestra la secuencia de aplicar este método al mismo conjunto de datos con k=1, 7 y 15 (de izquierda a

derecha). En este caso la curva discriminante a medida que k aumenta comienza a tener una tendencia lineal.

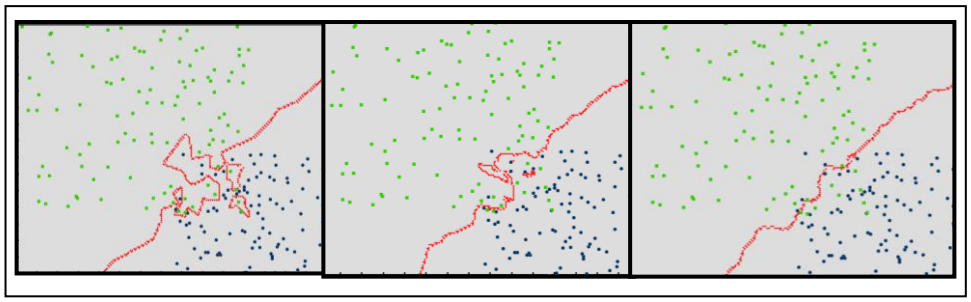

FIGURA 4.3 CAMBIOS EN LA EXPRESIVIDAD DE LOS K-VECINOS PARA K=1, 7 Y 15

La alta expresividad de este método es una de sus mayores ventajas en relación a otros. Al compararse este método con la regresión lineal de grado 1, por ejemplo, se nota que se adapta mucho mejor a distribuciones de datos irregulares dada su naturaleza local.

Precisión:

A continuación se examinará si para un conjunto de datos particular generados bajo la distribución de “cuadrados traslapados” existe alguna ventaja en incrementar el valor de k en relación a su precisión. La tabla II muestra los resultados del experimento para k=1, k=7, k=15 y N=200 datos.

TABLA II

ERRORES DEL MÉTODO K-VECINOS PARA K=1, 7 Y 15 APLICADOS A LA DISTRIBUCIÓN DE CUADRADOS TRASLAPADOS

K	Error de entrenamiento	Error de prueba
1	0	0.09
7	0.055	0.08
15	0.065	0.09

Como se observa en la tabla II el valor de k=15 no produjo el error de prueba más bajo, (*como talvez podría haberse esperado*). Según esta tabla no existe una diferencia notable en los errores de prueba respecto al valor de k seleccionado para este conjunto de datos. En general, la determinación del adecuado valor de k es una de las decisiones más importantes en este método. Si se escoge un valor muy bajo para k la curva discriminante no será muy estable mientras que si se escoge valores muy altos de k aumentará el número de malas clasificaciones (error de prueba).

La herramienta gráfica que aquí se presenta permite comparar curvas discriminantes para diferentes valores de su parámetro en

una misma pantalla. En este caso se ha graficado el método de los k-vecinos para distintos valores de k para considerar el error de prueba. La figura 4.4 muestra el resultado.

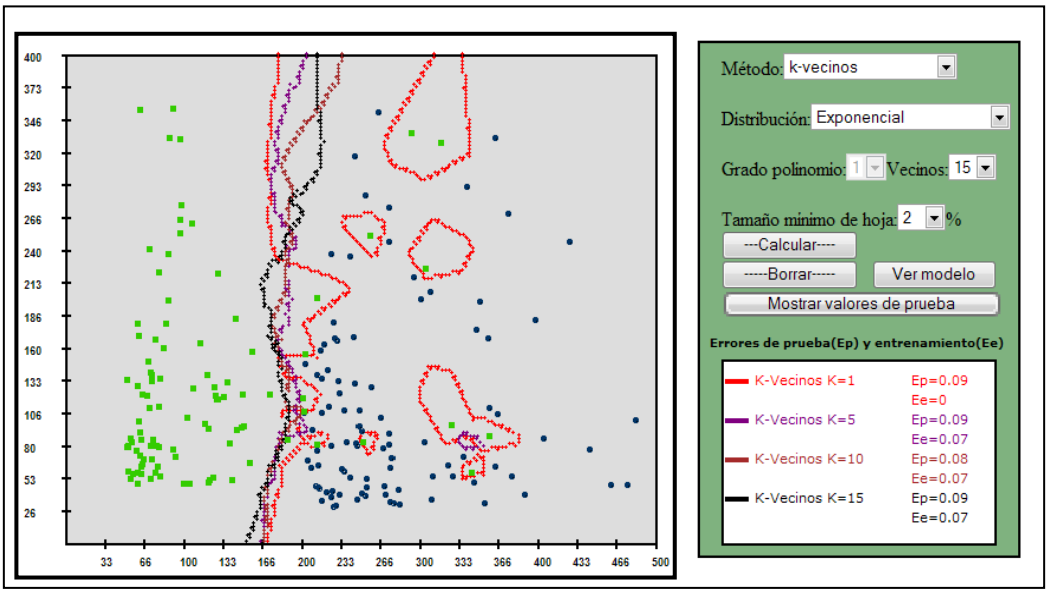

FIGURA 4.4 MÉTODO DE LOS K-VECINOS PARA K=1,5,10 Y 15 PARA CLASES DISTRIBUIDAS EXPONENCIALMENTE.

Se observa en la figura 4.4 que la precisión del método no ha variado mucho para distintos valores de k, siendo la mayor para k=10. La precisión del método es bastante buena con un máximo error de prueba de 0.9.

Costo computacional:

Puesto que el método de los k-vecinos es retardado, el tiempo de respuesta será elevado si el número de datos N es grande. Si k aumenta también lo hará el tiempo de respuesta. Si bien el costo computacional para k=1 y N=200 datos no es alto, sin embargo para k=50 y N=600000, por ejemplo, este costo es un limitante.

Robustez a datos anormales o ruidosos:

Este programa tiene la ventaja de permitir realizar pequeños cambios en el conjunto de puntos de la muestra para observar como reacciona la curva discriminante. A continuación se analizará experimentalmente si este método es sensible a datos anormales desplazando uno de sus puntos.

La figura 4.5 realiza el experimento para k=1 en donde se nota que el método es sensible, la presencia de un dato anormal (*punto azul en la parte derecha de la figura*) produjo el aumento de una frontera adicional en la curva discriminante. Mientras que en la figura 4.6 donde k=5 se observa que la curva permanece insensible a ese mismo dato anómala.

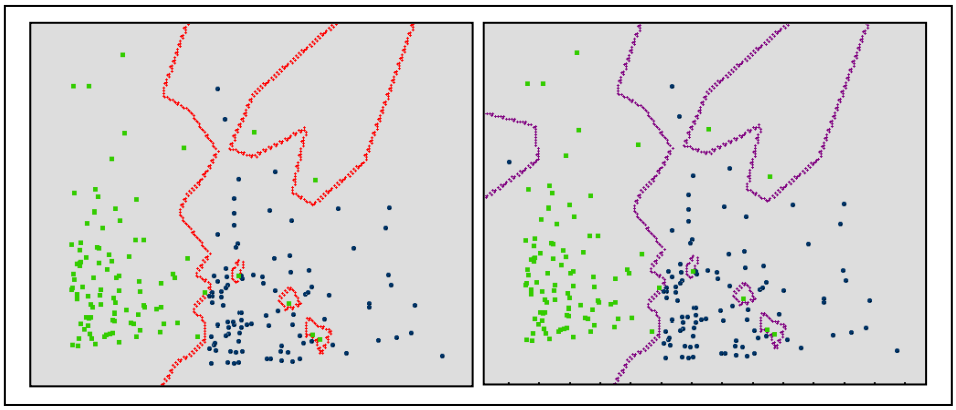

FIGURA 4.5 EJEMPLO DE LA SENSIBILIDAD DE 1-VECINO MÁS CERCANO A UN DATO ANÓMALA

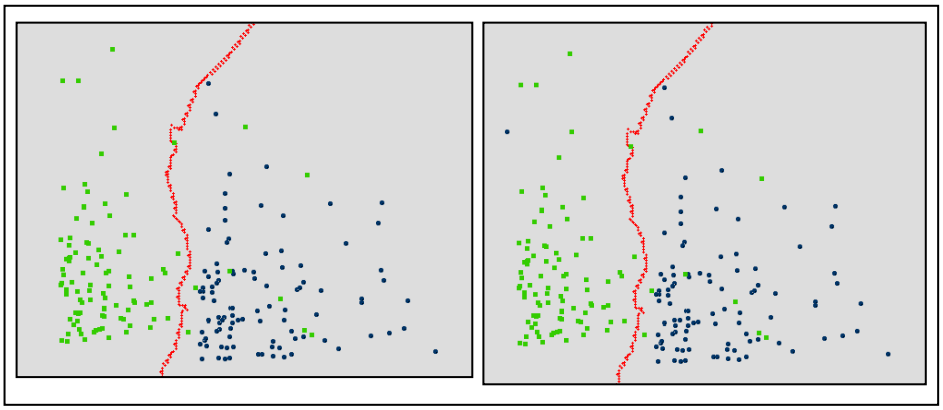

FIGURA 4.6 EJEMPLO DE LA SENSIBILIDAD DE 5-VECINOS MÁS CERCANOS A UN DATO ANÓMALA

Observando las figuras 4.5 y 4.6 se aprecia que el método de los kvecinos para k=5 es más general puesto que asume que los elementos “verdes” que contaminan la clase “azul” no son una ley general de la población sino mas bien sólo casos particulares de la

muestra. Esto se debe a que cada punto (x,y) del plano bidimensional de la figura 4.6 espera dentro de sus 5 vecinos más cercanos tener al menos 3 vecinos “verdes” para considerarse que está dentro de una región “verde”. Posiblemente por este grado de generalidad 5-vecinos más cercanos tenga mejor capacidad para clasificar acertadamente nuevos elementos con la misma distribución que 1-vecino más cercano. Este hecho lo refleja la tabla III que muestra los resultados de aplicar el método kvecinos a varias muestras de la misma población de donde proviene la muestra de las figuras 4.5 y 4.6. En esta tabla se observa que *5-vecinos más cercanos* clasificó mejor la mayoría de las muestras.

TABLA III

ERROR DE PRUEBA DE 1-VECINO VS. 5-VECINOS PARA MUESTRAS CON CLASES EXPONENCIALES

Muestra	k = 1	k = 5
1	0.15	0.1
2	0.13	0.05
3	0.08	0.06
4	0.14	0.07
5	0.2	0.08
6	0.07	0.07
7	0.14	0.06
8	0.07	0.08
9	0.08	0.07
10	0.16	0.12
11	0.11	0.1

4.2 Método de Naive Bayes Kernel

Como se mencionó anteriormente este método aplica explícitamente el teorema de Bayes para la tarea de clasificación. Si se conociesen las funciones de probabilidad de los atributos de cada nuevo elemento, éste se lo clasificaría utilizando estas funciones, de manera casi instantánea se hallarían las probabilidades a priori y a posteriori y la clase a la que pertenece este nuevo elemento. Si este fuese el caso, este método sería anticipado tal como lo es el método de regresión que invierte un considerable tiempo generando el modelo pero luego realiza la clasificación casi inmediatamente. Lamentablemente, éste no es el caso.

Puesto que se desconoce las funciones de probabilidad del conjunto de datos, éstas se estiman a través de las funciones núcleo que utilizan todo o una parte de este conjunto para encontrar el valor aproximado de la funciones.

La siguiente expresión, que es la utilizada para estimar una función de densidad en el caso univariado, muestra que para encontrar el

valor aproximado de f(x) se debe recurrir a los n valores de la variable presentes en el conjunto de datos:

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

Si bien hay funciones núcleo que solamente utilizan para su estimación las observaciones presentes en una vecindad de longitud 2, cada vez que se requiere clasificar un nuevo elemento es necesario, al igual que en el método de los k-vecinos más cercanos, recurrir a observaciones anteriores. Por lo tanto, este método de Naive Bayes es retardado. Puesto que este método es retardado su costo computacional será elevado para un número grande de datos.

Expresividad:

Una de las ventajas de los métodos no paramétricos como éste, es que son más expresivos que los métodos paramétricos. Aunque la curva discriminante de Naive Bayes puede ser no lineal ésta no es tan expresiva como la de los k-vecinos.

Precisión:

A continuación se examinará la precisión del método aplicado a muestras de diferentes poblaciones. En la figura 4.7 se muestra el resultado de haber aplicado este método a muestras de poblaciones distribuidas exponencialmente. La precisión del método es buena con un error de prueba para este caso de 0.06.

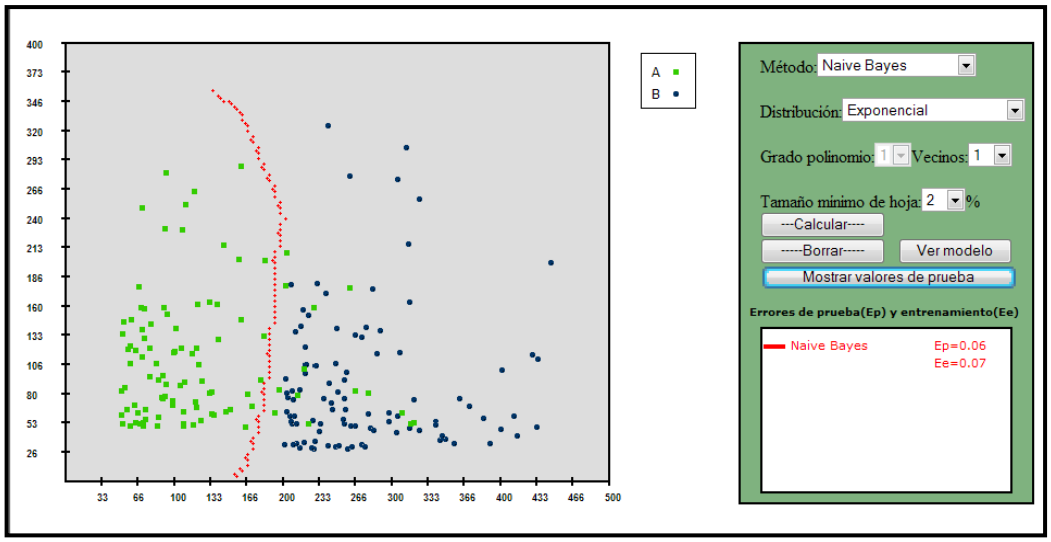

FIGURA 4.7 NAIVE BAYES APLICADO A MUESTRAS DE POBLACIONES EXPONENCIALES.

La figura 4.8 muestra la aplicación del método de Bayes a una distribución de cuadrados traslapados, en este caso se nota que este método no paramétrico local supera en precisión al método paramétrico global de regresión de grado 2. Mientras que el método

de Naive Bayes obtiene un error de prueba de 0.09 el de regresión obtiene 0.11.

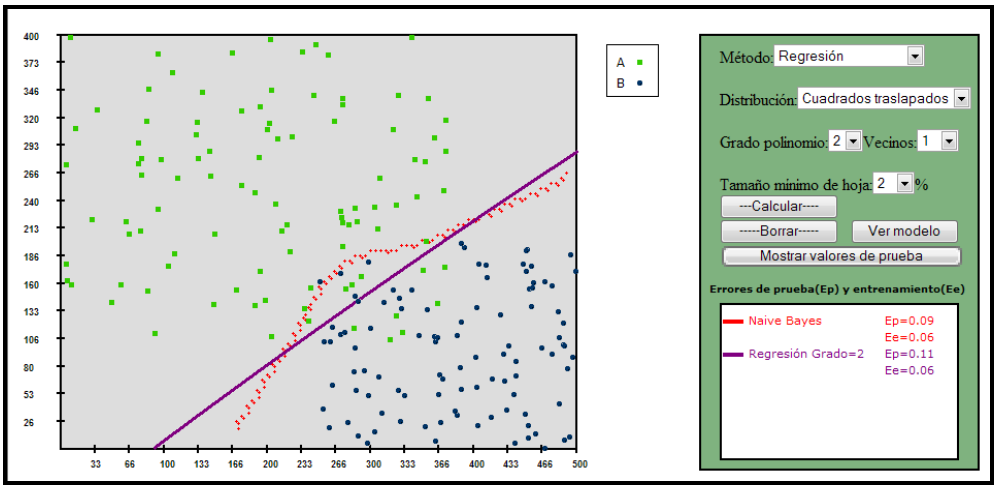

FIGURA 4.8 COMPARACION ENTRE EL MÉTODO DE NAIVE BAYES Y REGRESIÓN PARA DISTRIBUCION DE CUADRADOS TRASLAPADOS

Como se mencionó anteriormente este método se basa en la suposición (frecuentemente no muy realista) de que los atributos de cada elemento son independientes, para este caso particular se tienen 2 atributos o variables X y Y que pueden ser independientes o no.

En la configuración de rectángulos traslapados que ofrece la página web los puntos que pertenecen a cada una de las clases se

encuentran restringidos por los límites de cada rectángulo, es decir para cualquiera de los 2 rectángulos se cumple que: $L1<Xc<L2$ y $L3<Yc<L4$ donde Xc y Yc son las variables X y Y correspondientes a la clase c. Aunque X y Y están restringidas son independientes. En la configuración "diagrama de Venn" los miembros de cada clase se encuentran restringidos a los límites de cada círculo, de tal manera, se tiene que en ambos círculos se cumple que $(Xc-h)^2+(Yc-k)^2=r^2$ donde así mismo Xc y Yc son las variables X y Y correspondientes a la clase c; r el radio del círculo y (h,k) el centro del mismo. Se observa entonces que para esta configuración X y Y no son independientes.

Aunque Naive Bayes tiene la restricción de trabajar con variables o atributos independientes en la práctica se puede comparar el uso de este método con métodos que no tienen esta restricción. En la figura 4.9 se observa la aplicación de este método sobre la configuración "diagrama de Venn" en la cual X y Y son dependientes. En la misma figura se observa también la aplicación de 6-vecinos más cercanos, este último método no tiene la restricción de independencia.

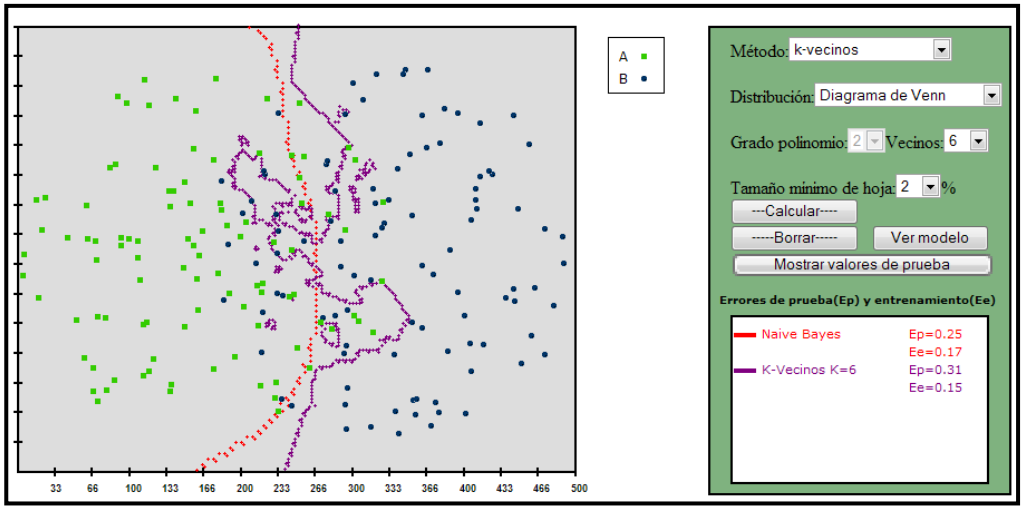

FIGURA 4.9 COMPARACIÓN ENTRE 6-VECINOS Y NAIVE BAYES

Aunque el método de los k-vecinos no tiene restricción de independencia y Naive Bayes sí la tiene, el error de prueba de los kvecinos para k=6 es más alto que el de Bayes (0.25 vs. 0.31). Sin embargo, ambos errores son considerablemente altos.

Aunque en la figura 4.9 no se muestra, cuando se aplica al mismo conjunto de datos 12-vecinos más cercanos el error de prueba baja a 0.2 con lo cual se supera a Naive Bayes. Si se conoce que las variables de un conjunto de datos no son independientes sería aconsejable contrastar los resultados ofrecidos por Naive Bayes con otros métodos como en este caso.

Robustez a datos anómalas:

A continuación se examinará si este método Naive Bayes es robusto a datos ruidosos o anómalas. Se examinará esto en 2 conjuntos de datos diferentes. Las técnicas bayesianas son robustas al ruido y estables a la muestra, sin embargo, el método que aquí se trata es mas bien una combinación: Naive Bayes con estimaciones núcleo, éste es un método bayesiano no paramétrico.

La figura 4.10 muestra la comparación del método aplicado al mismo conjunto de datos sin y con un dato anómala para la distribución de cuadrados traslapados, en la figura este dato se encuentra encerrado en un círculo.

La figura 4.10 muestra que la curva discriminante de este método puede ser sensible a pequeños cambios. El desplazamiento de apenas un punto ha aumentado fronteras a la curva discriminante. Sin embargo, para esta distribución de puntos el desplazamiento realizado aumenta fronteras a la curva discriminante que prácticamente no tienen influencia.

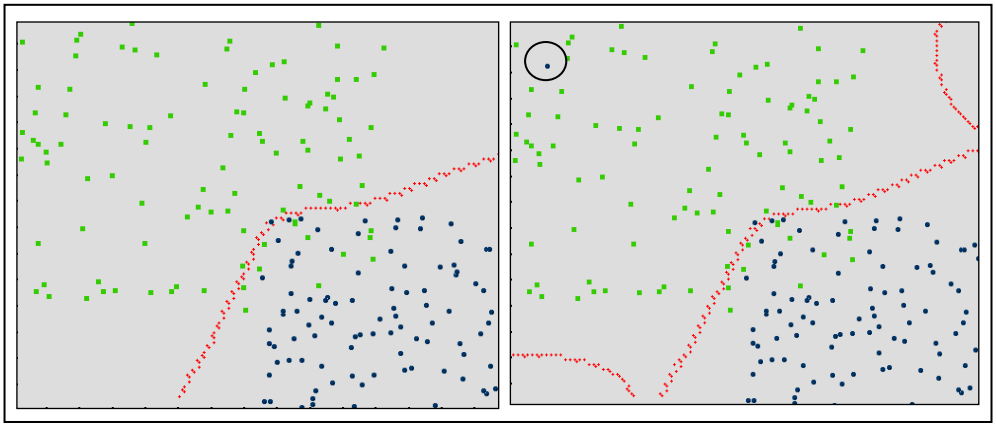

FIGURA 4.10 SENSIBILIDAD DE NAIVE BAYES KERNEL ANTE UN DATO ANÓMALA, DISTRIBUCIÓN DE CUADRADOS TRASLAPADOS

Como se observó anteriormente el hecho de que la curva discriminante de un método se dibuje diferente en el espacio de datos al convertir mediante desplazamiento a un dato normal en dato anómala, no siempre es razón para creer que la precisión del método ha sido influenciada por este cambio. Puede ser que los cambios en la curva discriminante sean muy poco relevantes para la discriminación de las clases. La figura 4.11 muestra otra vez este caso.

En la parte derecha de la figura 4.11 se observa que el desplazamiento de un punto (*encerrado en un círculo en la figura*) ha hecho que la curva complete su frontera. Puesto que es muy poco

probable que un punto de la distribución exponencial de la clase verde sobrepase esa nueva frontera, el cambio en la graficación de la curva es *prácticamente* irrelevante. Esto se corrobora numéricamente teniendo ambos casos el mismo error de prueba de 0.09.

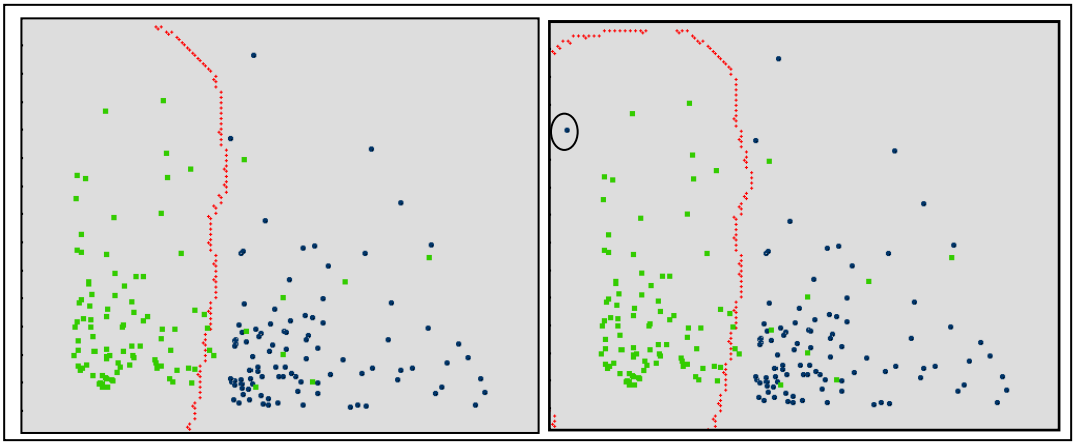

FIGURA 4.11 SENSIBILIDAD DE NAIVE BAYES KERNEL ANTE UN DATO ANÓMALA, DISTRIBUCIÓN EXPONENCIAL

4.3 Método de árbol de decisión

Aunque para N datos el método de árbol de decisión puede continuar segmentando cada nodo hasta que absolutamente todas las hojas sean puras, esto no sería conveniente puesto que las reglas que se desprenderían de este árbol serían demasiado específicas para los datos de entrenamiento (*sobreajustadas*) y

podrían fallar al tratar de clasificar correctamente un nuevo elemento. Para obtener un árbol que produzca un modelo más general se suele eliminar algunas hojas (*condiciones*) del árbol. Este procedimiento se conoce como poda. La poda puede ocurrir antes de terminar de construir todo el árbol (prepoda) o cuando éste ya esté construido (pospoda). Aquí se empleará el primer tipo de poda.

La prepoda consiste en emplear un criterio que controle hasta cuando seguir particionando un nodo aunque éste no sea puro. Un criterio podría ser particionar los nodos del árbol hasta que la cardinalidad de éstos sea igual o menor a un valor s. Este parámetro de parada s que puede ser un porcentaje le dirá al algoritmo que continúe particionando un nodo T hasta $n(T) \leq sN$ donde n(T) es el número de observaciones en el nodo. En el programa que aquí se presenta este parámetro de parada s puede llegar hasta 26%.

A continuación se analiza el comportamiento de este método al ser aplicado a los conjuntos de datos que aquí se proponen.

Precisión:

La herramienta gráfica que aquí se presenta permite observar la precisión de este método al variar su parámetro de parada s, la figura 4.12 muestra un ejemplo de esto.

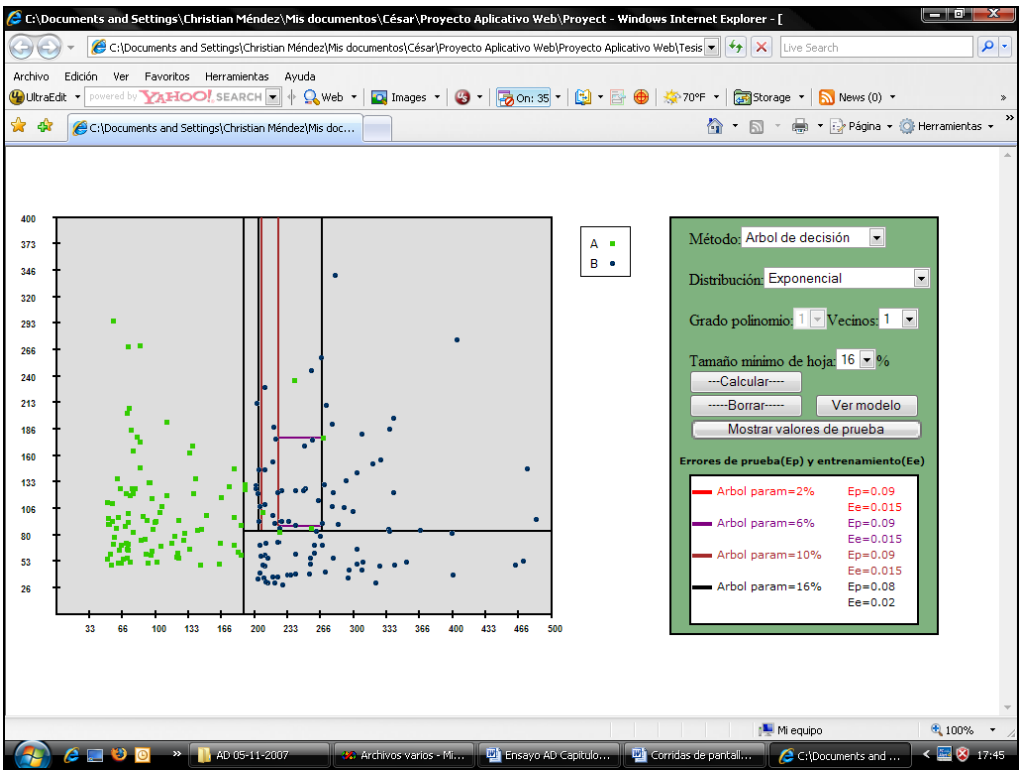

FIGURA 4.12 APLICACIÓN DEL ÁRBOL DE DECISIÓN VARIANDO S PARA LA MISMA MUESTRA

Al analizar la precisión de este método para las distribuciones de diagrama de Venn y cuadrados traslapados en las tablas IV y V, en principio se nota que esta precisión no se afectó demasiado por el

parámetro de parada s. Sin embargo, en el segundo conjunto de datos esta precisión tuvo una ligera mejoría al aumentar s. Aunque cada problema es diferente, en general se puede incrementar el grado de poda (parámetro de parada en este caso) convenientemente para obtener mejorías en la precisión del método. Sin embargo, este grado de poda no puede ser muy elevado.

TABLA IV

RESULTADOS DEL ÁRBOL DE DECISIÓN EN LA DISTRIBUCIÓN DE DIAGRAMA DE VENN

Parámetro	Error entrenamiento	Error de prueba
2%	0.015	0.22
6%	0.05	0.21
10%	0.07	0.22

TABLA V

RESULTADOS DEL ÁRBOL DE DECISIÓN EN LA DISTRIBUCIÓN DE CUADRADOS TRASLAPADOS

Parámetro	Error entrenamiento	Error de prueba
2%	0.005	0.06
6%	0.02	0.05
10%	0.03	0.04

La figuras 4.13 y 4.14 muestran gráficamente las clasificaciones realizadas por el método a los 2 conjuntos de datos de las tablas anteriores.

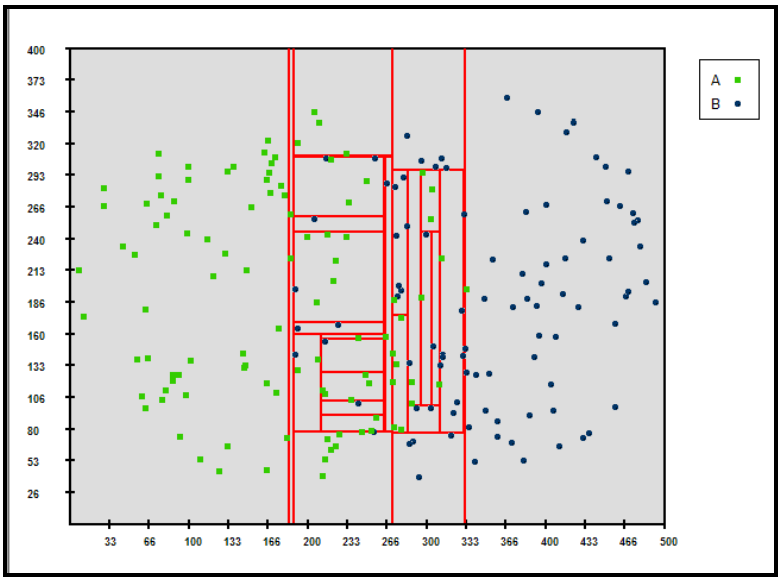

FIGURA 4.13 APLICACIÓN DE ÁRBOL DE DECISIÓN A LA DISTRIBUCIÓN DE DIAGRAMA DE VENN

Se observa que la clasificación hecha por el método en la figura 4.14 es mucho más sencilla que la de la figura 4.13, esto es porque el conjunto de datos de la distribución de diagrama de Venn contiene un mayor región de traslape. Esto en parte explica porque el error de prueba en el caso de esta distribución es más alto. Al ser más amplia la región de traslape el método realiza más segmentaciones y puede sobreajustarse a los datos de entrenamiento.

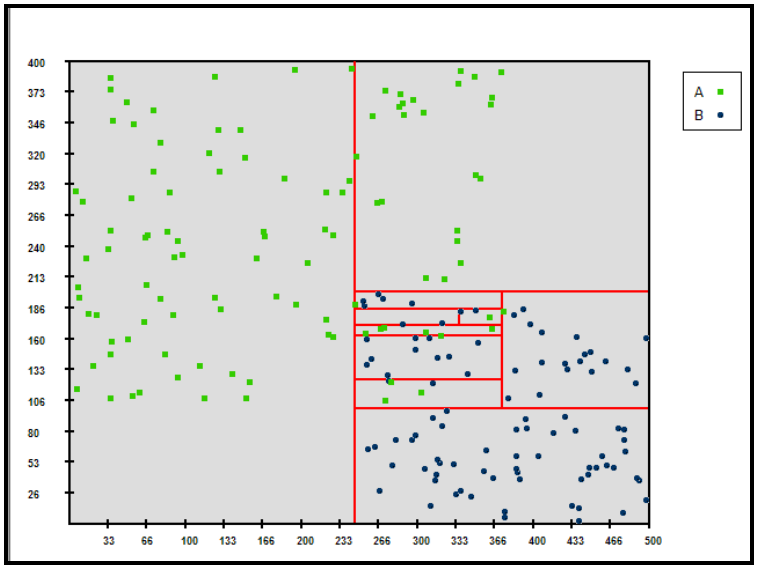

FIGURA 4.14 APLICACIÓN DE ÁRBOL DE DECISIÓN A LA DISTRIBUCIÓN DE CUADRADOS TRASLAPADOS

Estabilidad:

A continuación se experimentará como responde el algoritmo a diferentes muestras de la misma población. Se experimenta con la distribución de diagramas de Venn para 2 parámetros de parada de 2 y 10%.

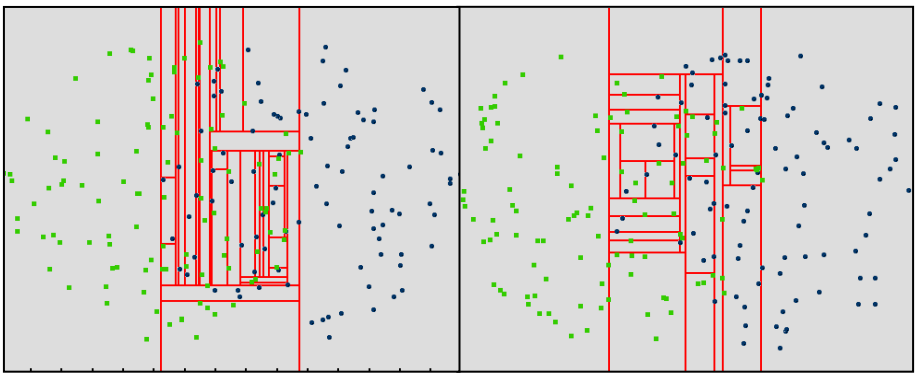

FIGURA 4.15 VARIACIÓN DEL ARBOL DE DECISIÓN A DIFERENTES MUESTRAS CON S=2%

La figura 4.16 muestra las fronteras discriminantes para 2 muestras diferentes con un parámetro de parada s de 10%.

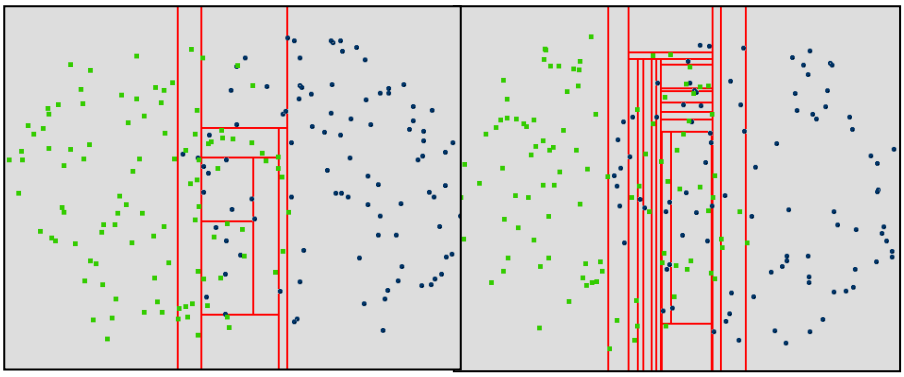

FIGURA 4.16 VARIACIÓN DEL ÁRBOL DE DECISIÓN A DIFERENTES MUESTRAS CON S=10%

La inestabilidad del método para la población de las figuras 4.15 y 4.16 a diferentes parámetros de parada es notoria. Es útil considerar en este punto que son muestras de una población donde la región de traslape carece de alguna regularidad, se podría pensar que talvez por esto se justifica la inestabilidad de este método, después de todo el único patrón que tienen los datos de la distribución de Diagrama de Venn es que los datos de cada clase se encuentran encerrados en un círculo. Sería útil analizar cuál es la estabilidad del método si se lo aplica a una población menos irregular.

A continuación se prueba si este método es sensible a diferentes muestras de poblaciones cuyas distribuciones son exponenciales.

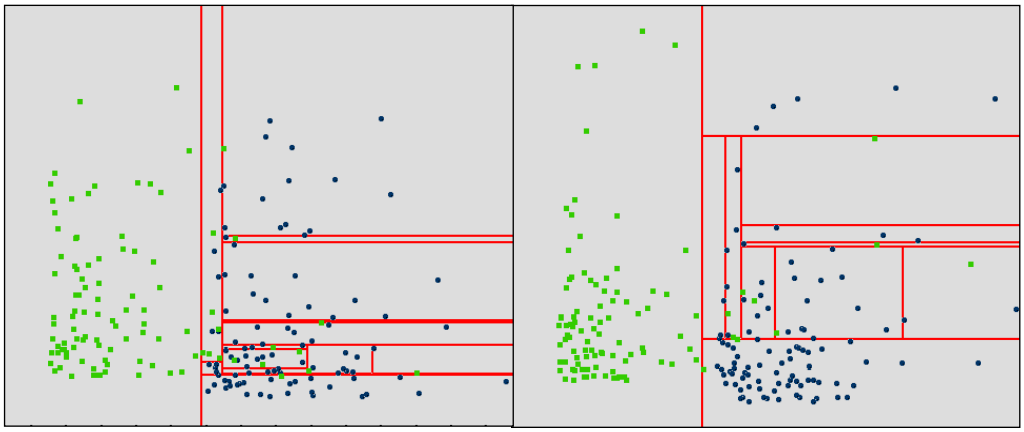

FIGURA 4.17 ÁRBOL DE DECISIÓN PARA 2 MUESTRAS DE POBLACIONES EXPONENCIALES CON S=6%

La figura 4.17 muestra la aplicación del método a 2 muestras distintas de poblaciones exponenciales utilizando un parámetro de parada s=6%, como se observa los gráficos para cada muestra son diferentes. Las fronteras discriminantes reaccionan a cambios en la muestra. A continuación se mostrarán los resultados de la estabilidad en un experimento más largo.

Si varían las fronteras discriminantes con una muestra diferente parecería razonable que los errores de prueba también varíen. La tabla VI muestra los resultados de haber aplicado el algoritmo a 6

muestras distintas de las mismas poblaciones exponenciales. Para las 6 muestras se ha variado el parámetro de parada s a niveles de 2, 6, 10 y 16%.

TABLA VI

ÁRBOL DE DECISIÓN APLICADO A 6 MUESTRAS DE LA MISMA POBLACION CON CLASES DISTRIBUIDAS EXPONENCIALMENTE

	MUESTRAS					
	# 1	# 2	# 3	# 4	# 5	# 6
S	Ep	Ep	Ep	Ep	Ep	Ep
2%	0,06	0,06	0,1	0,14	0,06	0,09
6%	0,06	0,04	0,1	0,13	0,06	0,09
10%	0,05	0,04	0,08	0,13	0,06	0,09
16%	0,05	0,04	0,08	0,13	0,06	0,08

La tabla VI condensa 2 características del método de árbol de decisión para los conjuntos de datos distribuidos exponencialmente que ofrece el programa, estas características son precisión y estabilidad.

Aunque al analizar anteriormente la precisión para la discriminación de las distribuciones de diagrama de Venn se observó que ésta no variaba demasiado al variar el parámetro de parada, una primera observación de la tabla VI es que para esta otra distribución del conjunto de datos (*distribución exponencial*) se logra una mejoría en

la precisión al incrementar el parámetro de parada. A excepción de la muestra # 5 en la cual la precisión se mantiene. Sin embargo, esta precisión varía de una muestra a otra, los mejores (más bajos) errores de prueba cambian de 0.06 a 0.13. No es de esperarse que el error de prueba de un método tenga variación cero, después de todo es una variable aleatoria, pero si la varianza del error de prueba es baja se podrá decir que el método es estable. En general, el método de árbol de decisión es inestable ante variaciones de la muestra.

Expresividad:

Dado que este algoritmo realiza segmentaciones paralelas a los ejes su expresividad es limitada. Existen patrones que mientras otros métodos los capturan con mayor facilidad, este método lo hace utilizando demasiadas particiones mostrando una discriminación de los puntos no sencilla. Esto ocurre en general, cuando las regiones que determinan cada clase tienen fronteras no paralelas a los ejes. En la figura 4.18 se muestra que el método de regresión lineal de grado 1 captura mejor el patrón de datos que el método de árbol.

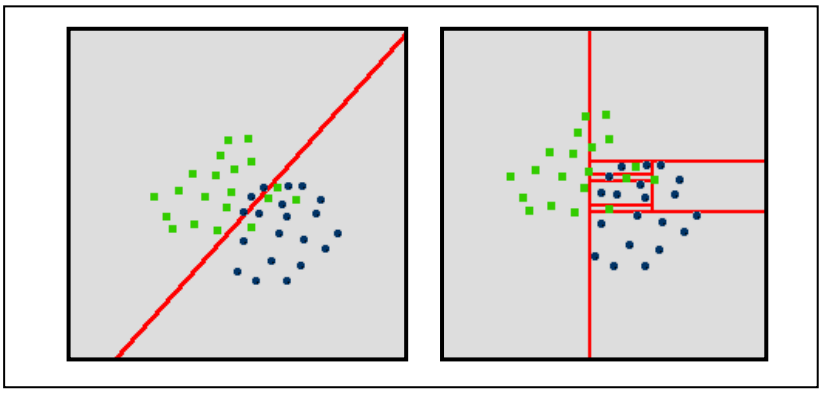

FIGURA 4.18 REGRESIÓN LINEAL DE GRADO 1 Y ÁRBOL DE DECISIÓN: DIFERENCIAS AL CAPTAR EL PATRÓN DE DATOS

Comprensibilidad:

La comprensibilidad de este método es una de sus mayores ventajas, el sistema de reglas que genera es bastante sencillo de seguir desde el punto de vista humano.

La figura 4.19 muestra el conjunto de reglas generado por la página web para la muestra de la figura 4.20. Este modelo es más sencillo que el generado por el método de regresión pues no se requiere realizar ningún cálculo para determinar la clase de un nuevo elemento con atributos (x,y).

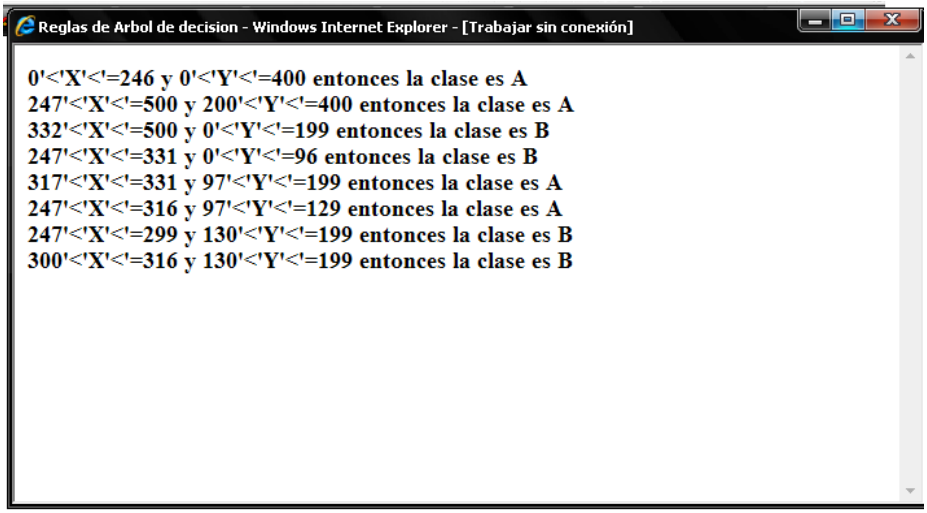

FIGURA 4.19 MODELO DE UN ÁRBOL DE DECISIÓN

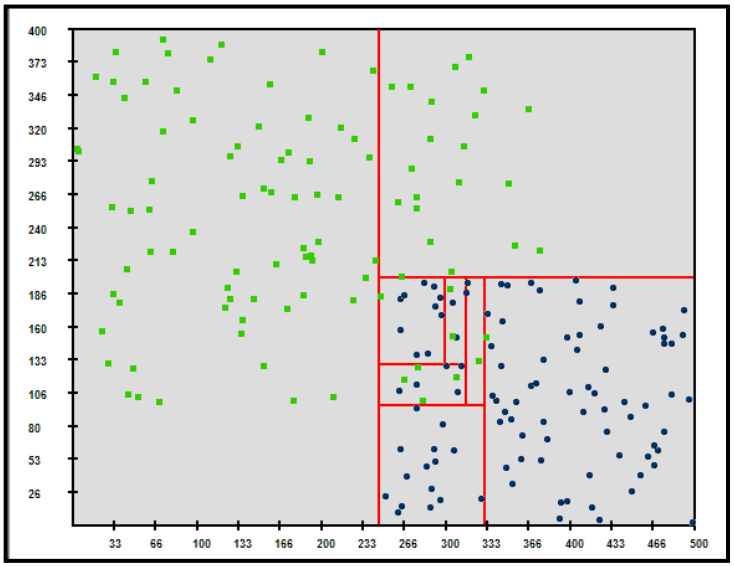

FIGURA 4.20 REPRESENTACIÓN GRÁFICA DEL SISTEMA DE REGLAS DE LA FIGURA 4.19

4.4 Métodos de regresión

Los métodos de regresión son métodos paramétricos que dentro de sus características tienen las de ser anticipativos, comprensibles, de bajo costo computacional pero de limitada expresividad dependiendo ésta de la complejidad del modelo propuesto.

La figura 4.21 muestra el resultado de haber aplicado regresión lineal de grado 1 a un conjunto de datos exponencial. En esta figura la recta parece ser un discriminante aceptable para el conjunto de datos. El error de prueba que ofrece el método es de 0.14

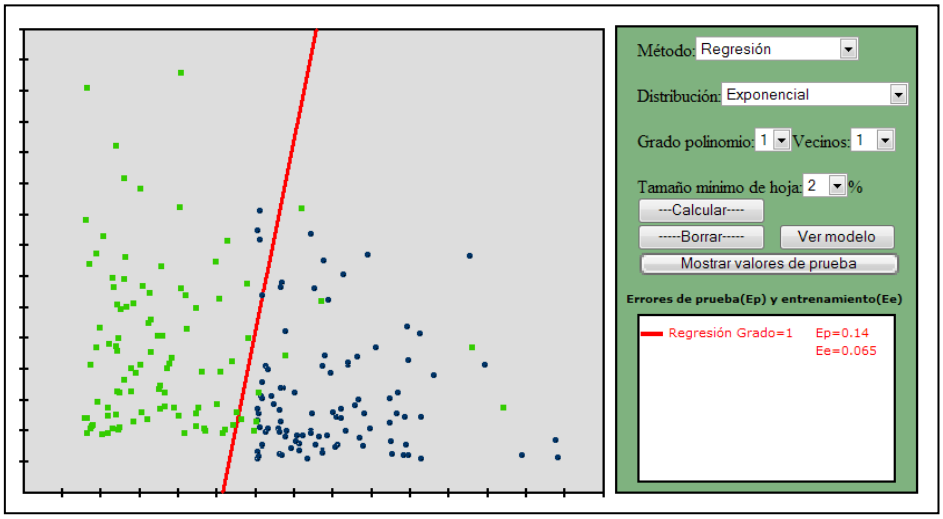

FIGURA 4.21 REGRESION GRADO 1 APLICADO A UN CONJUNTO DE DATOS EXPONENCIAL

Aunque hubiera sido deseable que el error de prueba anterior sea menor, talvez éste podría ser sólo fruto de la incertidumbre. Talvez sería conveniente ejecutar el experimento varias veces y observar como responde el método.

En la tabla VII se muestra el resultado de haber aplicado el método a 10 muestras de la misma población para tener una mejor idea de su precisión. El resultado es un error de prueba medio de 0,079 el cual es mejor. Sin embargo, nada se ha dicho todavía acerca de la variabilidad de este error de una muestra a otra.

Es posible que el error de prueba medio sea aceptable pero que exista una diferencia notable en el error de prueba al considerar 2 muestras particulares. Se observa, por ejemplo, que el error de prueba de la muestra # 2 es 0.03 sin embargo, el error de prueba que se obtuvo al aplicar por primera vez el método fue de 0.14.

TABLA VII
APLICACIÓN DE REGRESION GRADO 1 A 10
MUESTRAS PROVENIENTES DE UNA POBLACIÓN
CON CLASES DISTRIBUIDAS EXPONENCIALMENTE

# Muestra	Ep
1	0,11
2	0,03
3	0,11
4	0,05
5	0,04
6	0,1
7	0,08
8	0,11
9	0,08
10	0,08

En la figura 4.22 se muestra la aplicación del método para 2 muestras de la misma población. Se nota que la sensibilidad de la recta discriminante es considerable al cambiar de muestra.

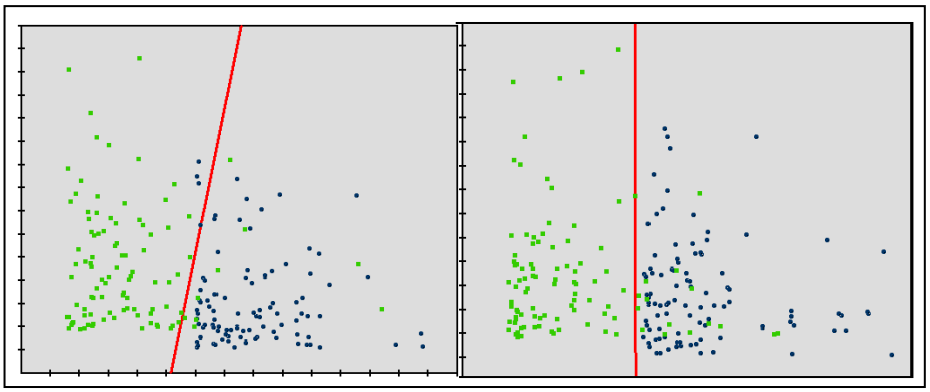

FIGURA 4.22 SENSIBILIDAD DE LA REGRESIÓN LINEAL AL CAMBIAR LA MUESTRA DE UNA POBLACIÓN CON CLASES EXPONENCIALES

Por una parte, existen conjuntos de datos para los cuales es más conveniente aplicar un método antes que otro. La minería de datos ofrece algunas posibilidades en cuanto a métodos. Por otro lado, existen regiones de conjuntos de datos donde podría ser muy difícil obtener una regla de discriminación.

Se podría intentar complicar el modelo para lograr una mejoría en la discriminación, pero tal como lo muestra la figura 4.23, esto no siempre resulta. En esta figura se observa que al aumentar la complejidad del modelo el error de prueba se incrementó ligeramente de 0.11 a 0.12

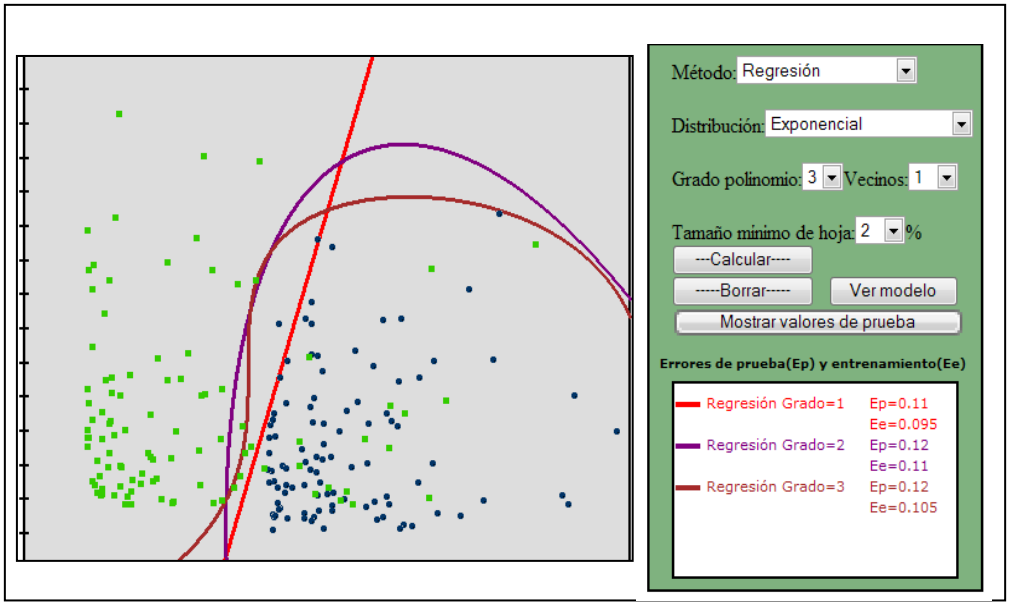

FIGURA 4.23 REGRESION GRADO 1,2 Y 3 PARA DISCRIMINAR CLASES CON DISTRIBUCIÓN EXPONENCIAL

Aunque la expresividad de la recta de regresión es rígida, al aumentar la complejidad del modelo se puede lograr mayor expresividad.

La figura 4.24 muestra como la regresión de grado 2 es mucho más expresiva que la recta de regresión. La parábola logra discriminar bastante bien los datos de entrenamiento.

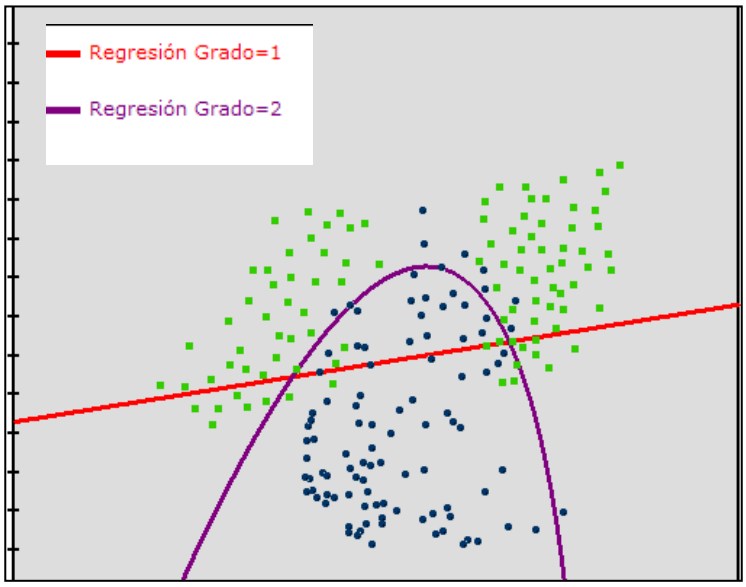

FIGURA 4.24 DIFERENCIAS EN LA EXPRESIVIDAD DE LA REGRESION DE GRADO 1 Y 2

El programa permite graficar curvas discriminantes del método de regresión lineal de hasta grado 4. Si la regresión de grado 2 discrimina bastante bien el conjunto de la figura 4.24 la de grado 4

lo hace casi perfectamente pero sobre los datos de entrenamiento. El problema es que al aumentar la complejidad de la regresión la curva discriminante tiende a ajustarse a los datos de entrenamiento produciendo errores en clasificaciones futuras.

4.4.1 Regresión lineal y regresión logística

Una de las ventajas de esta herramienta es que permite al usuario experimentar con diferentes conjuntos de datos que puede diseñar desplazando los puntos que desee una vez que éstos ya están graficados. En esta parte se experimentará con un conjunto de datos preparado que permite ver una diferencia entre la regresión lineal y la logística.

Como se vio en el capítulo 2, la técnica de regresión es empleada para explicar una variable dependiente o de salida en función de un conjunto de variables independientes o de entrada. Cuando se trata de análisis discriminante esta variable dependiente es categórica. Considérese el simple caso en el que se pretende discriminar un conjunto de elementos entre 2 clases donde cada elemento (*o instancia*)

del conjunto de datos está determinado por sólo una variable. Si se emplea regresión lineal de grado 1 para abordar este problema, se estaría intentando explicar una variable binaria Y (*0 ó 1*) en función de sólo una variable numérica X. Geométricamente se estaría tratando de ajustar una recta $Y=\alpha+\beta X$ a un conjunto de puntos del tipo (0,X) y (1,X). La parte izquierda de la figura 4.25 muestra un esbozo de esta situación. Ante la presencia de un valor extremo (*punto rojo*) como el que se muestra en la parte derecha de la misma figura 4.25 se observa que la recta se ve obligada a desplazarse hacia la derecha para que el ajuste siga siendo óptimo.

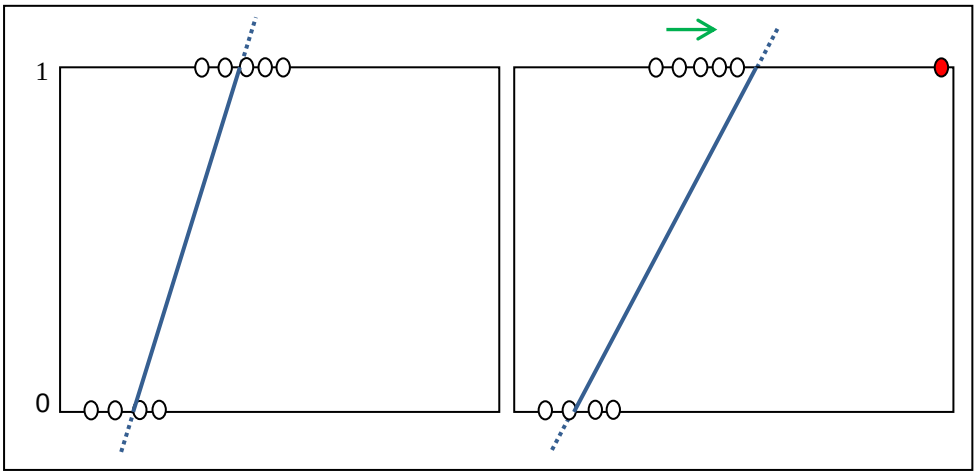

FIGURA 4.25 ESBOZO DE LA SENSIBILIDAD DE LA REGRESIÓN LINEAL ANTE OUTLIERS

A diferencia de la regresión lineal, la curva de ajuste generada con el método de regresión logística no siempre se verá afectada por este fenómeno.

Considérese un caso similar al de la figura 4.25 en el que se intenta ajustar una curva al conjunto de puntos determinados por sólo una variable numérica X. En el caso de regresión logística se intenta ajustar el modelo $p=1/(1+e^{-(\alpha+\beta X)})$ donde p es la probabilidad de que la variable Y sea igual a 1. El tipo de ajuste de esta regresión se lo muestra en la figura 4.26 donde se observa que el punto extremo (*punto rojo*) en la parte derecha de la figura no afecta el comportamiento del modelo. Si se considera, por ejemplo, el modelo $p=f(x)=1/(1+e^{-(-5.1+1.1X)})$, se tiene que f(10)=0.9972680, mientras que f(18)= 0.9999995 y f(22)= 0.99999999. Lo que nos indica que valores mayores y relativamente alejados del conjunto de puntos (*en este caso alejados de X=10*) no afectan significativamente la respuesta del modelo. Aunque no se incluya los valores X=18 y X=22 en el ajuste del modelo inicial, este modelo está ya *implícitamente* ajustado a estos valores alejados.

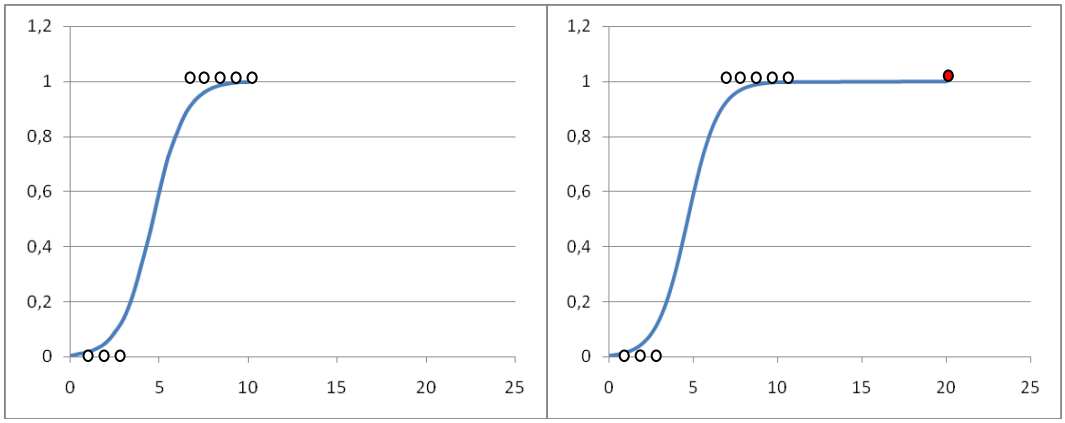

FIGURA 4.26 ESBOZO DE LA ROBUSTEZ DE LA REGRESIÓN LOGÍSTICA ANTE OUTLIERS

A continuación se mostrará esta diferencia entre regresión lineal y logística utilizando la presente herramienta computacional. En la figura 4.27 se muestra un conjunto de datos en el cual la tarea de discriminación es muy sencilla. Se observa en la figura que la recta generada con regresión logística discrimina muy bien las 2 clases mientras que la recta generada con regresión lineal compromete la buena discriminación debido al valor extremo presente en la parte extrema superior (*punto azul en la figura*). Se nota entonces que en este caso la regresión logística es más precisa que la regresión lineal.

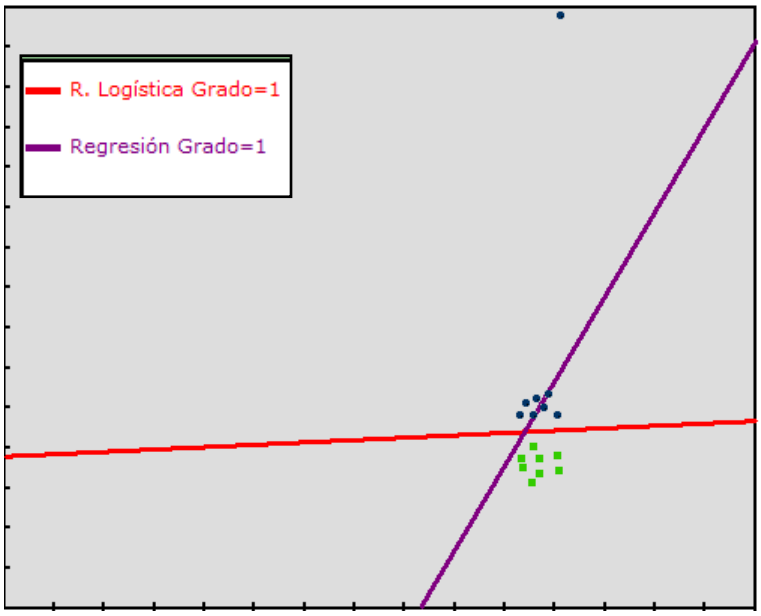

FIGURA 4.27 EJEMPLO DE UNA DIFERENCIA ENTRE REGRESIÓN LINEAL Y REGRESIÓN LOGÍSTICA

Aunque la exposición anterior presentó una diferencia entre la regresión lineal y la logística, hay casos en el que sus errores de prueba son iguales o similares. La figura 4.28 muestra el resultado de aplicar los 2 métodos con polinomios de segundo grado al mismo conjunto de datos. Se observa que aunque el error de prueba es el mismo (Ep=0.07) sus formas de discriminar son diferentes. La precisión de 93% que tienen los 2 métodos en este caso es bastante buena.

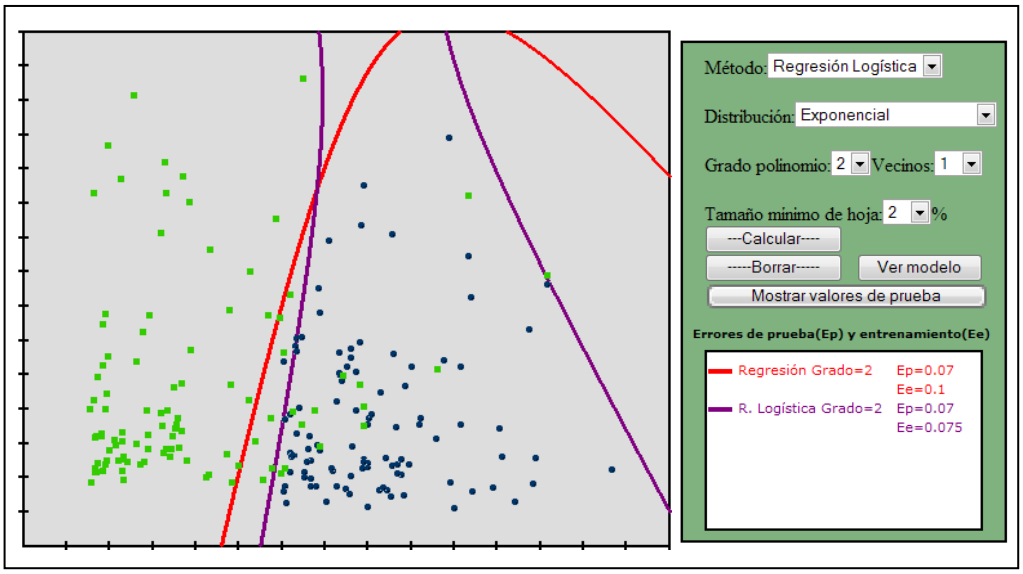

FIGURA 4.28 MÉTODOS DE REGRESIÓN GRADO=2 APLICADOS A UN CONJUNTO CON CLASES DISTRIBUIDAS EXPONENCIALMENTE.

4.5 Experimentos de comparación entre métodos

En esta sección se presentará los resultados de 2 experimentos, cada uno de estos consistirá en la aplicación de los 5 métodos vistos anteriormente a un conjunto determinado de datos. En el primer experimento se aplicarán los métodos a la ya antes vista distribución exponencial; en el segundo se utilizará un conjunto de datos especialmente preparado para resaltar algunas características de los métodos.

Experimento 1

La tabla VIII donde se detallan los resultados del experimento indica cada método utilizado con el valor de su parámetro correspondiente.

TABLA VIII
RESULTADOS DE UN EXPERIMENTO CON CLASES DISTRIBUIDAS EXPONENCIALMENTE

Método	Error de prueba
K-Vecinos K=1	Ep=0.13
K-Vecinos K=5	Ep=0.09
K-Vecinos K=10	Ep=0.09
K-Vecinos K=13	Ep=0.09
K-Vecinos K=17	Ep=0.1
Naive Bayes Kernel	Ep=0.09
Regresión Grado=1	Ep=0.1
Regresión Grado=2	Ep=0.1
Regresión Grado=3	Ep=0.09
Regresión Grado=4	Ep=0.09
R. Logística Grado=1	Ep=0.09
R. Logística Grado=2	Ep=0.09
R. Logística Grado=3	Ep=0.09
Arbol param=2%	Ep=0.11
Arbol param=4%	Ep=0.07
Arbol param=8%	Ep=0.06
Arbol param=12%	Ep=0.07
Arbol param=16%	Ep=0.07
Arbol param=18%	Ep=0.08
Arbol param=20%	Ep=0.08
Arbol param=22%	Ep=0.08
Arbol param=24%	Ep=0.08
Arbol param=26%	Ep=0.08

En general todas las versiones de los métodos presentadas en la tabla brindan buenos resultados. Sin embargo, para este conjunto

específico de datos de entrenamiento y datos de prueba la eficacia aumenta en algunas versiones de los métodos.

Según la tabla del experimento el método de los k-vecinos aumenta su eficacia al pasar de k=1 a k=5, esta eficacia se mantiene para k=10 y k=13 pero desmejora para k=17. De igual manera ocurre con el método de regresión lineal, utilizando los 2 últimos grados este método aumenta su precisión. También se observa que la regresión logística (*en este caso*) no muestra gran diferencia en relación a la regresión lineal.

El método que brinda mejores resultados en el experimento es el árbol de decisión. Si se excluye la versión de este método con parámetro de parada de 2%, todas las otras versiones superan en eficacia al resto de métodos, obteniéndose el mejor resultado al emplear el árbol de decisión con un parámetro de parada de 8% (*línea resaltada en la tabla*). La tabla muestra también el hecho de que el parámetro de parada o grado de poda no siempre puede ser incrementado demasiado sin perder precisión. Aumentando el parámetro de parada de 2 a 8% se logra una buena mejoría. Sin embargo, al incrementarlo de 8 a 12 y a 20% esta eficacia

desmejora. Finalmente, se observa que el método de Naive Bayes basado en funciones núcleo (*kernel*) en este caso es lo suficientemente bueno como para compararse con los métodos de regresión y los de k-vecinos más cercanos.

Experimento 2

En este caso se utiliza un conjunto de datos peculiar pero interesante, el cual es similar al ejemplo que aparece en la página 449 del libro ***"Introducción a la minería de datos"*** que se describe en las referencias bibliográficas.

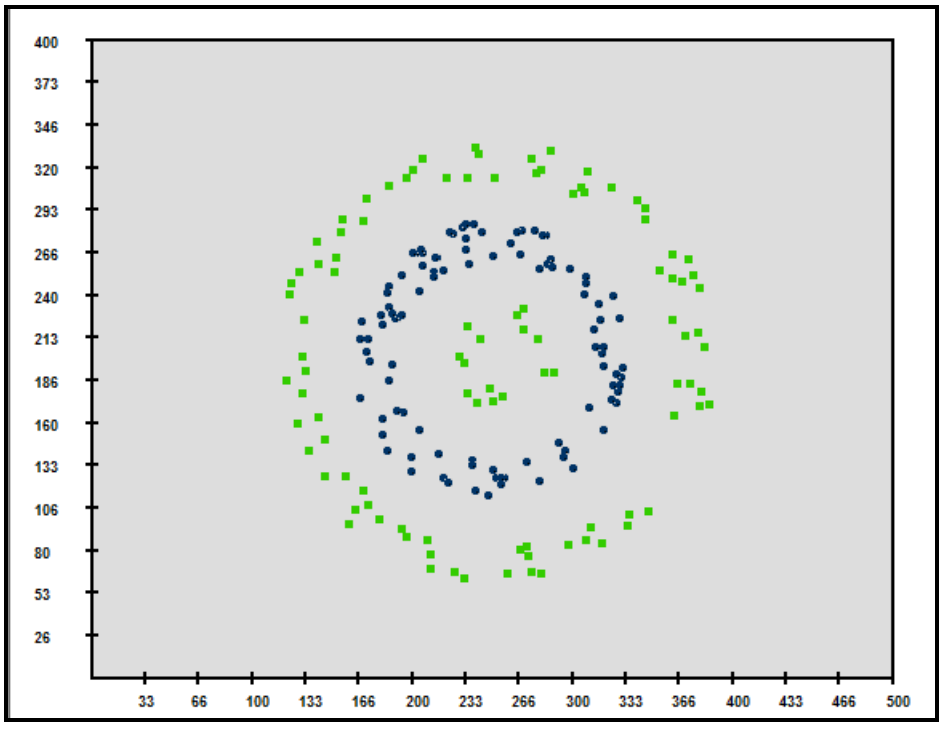

FIGURA 4.29 DISTRIBUCIÓN DE CÍRCULOS CONCÉNTRICOS

La figura 4.29 muestra este conjunto de datos el cual tiene forma de círculos concéntricos con elementos de sólo una clase en cada corona.

Aunque la distribución de círculos concéntricos a simple vista no presenta ninguna dificultad para su discriminación, la utilización de este conjunto tiene un fin pedagógico y se lo utiliza para examinar cómo responde cada método al tratar de captar este patrón de datos.

Es claro que el patrón de la distribución de este experimento requiere una gran expresividad para ser captado; las regresiones lineales de grado 1 y 2 fracasan en este intento tal como lo muestra la figura 4.30, siendo la regresión de grado 1 (*como es de esperarse*) demasiado pobre para este fin con un error de prueba bastante alto. En cambio, tanto la regresión lineal de grado 4 como el método de k vecinos captan bastante bien el patrón, siendo el métodos de los kvecinos el que lo hace de forma perfecta debido a su gran expresividad. La figura 4.31 muestra la captación de la regresión polinomial de grado 4.

A continuación se aplicará los métodos con diferentes valores en sus parámetros a un mismo conjunto de datos según la distribución aquí mencionada. Los resultados del experimento se muestran en la tabla IX.

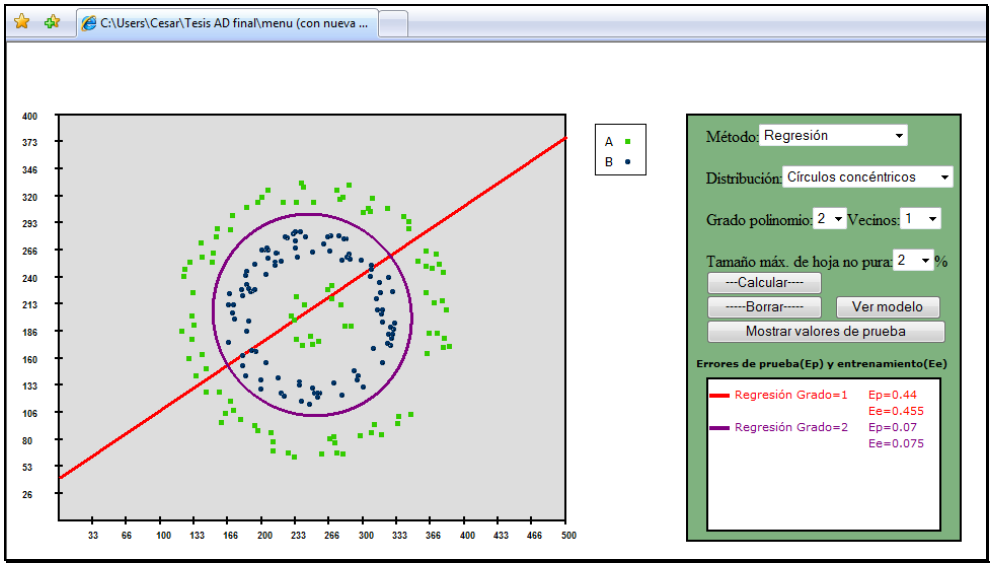

FIGURA 4.30 REGRESIÓN DE GRADO 1 Y 2 APLICADOS A LA DISTRIBUCIÓN DE CÍRCULOS CONCÉNTRICOS

Aunque se mencionó que el método de regresión lineal de grado 2 no podía captar bien el patrón del conjunto de puntos, el error de prueba de este método no es malo, tal como lo muestra la tabla de este experimento. Puesto que existen métodos que se comportan mejor que otros para determinados conjuntos de datos, se observa en la tabla que el árbol de decisión no obtiene los mejores resultados en este caso. La figura 4.32 muestra gráficamente el resultado de este método para un parámetro de 2%.

TABLA IX
RESULTADOS DE UN EXPERIMENTO CON CLASES DISTRIBUIDAS EN CÍRCULOS CONCÉNTRICOS

Método	Error de prueba
K-Vecinos K=1	Ep=0
K-Vecinos K=5	Ep=0
K-Vecinos K=10	Ep=0
K-Vecinos K=17	Ep=0.12
Naive Bayes	Ep=0.14
Regresión Grado=1	Ep=0.48
Regresión Grado=2	Ep=0.1
Regresión Grado=3	Ep=0.1
Regresión Grado=4	Ep=0.03
R. Logística Grado=3	Ep=0.1
Arbol param=2%	Ep=0.08
Arbol param=8%	Ep=0.08
Arbol param=15%	Ep=0.13
Arbol param=18%	Ep=0.13
Arbol param=22%	Ep=0.22
Arbol param=26%	Ep=0.22

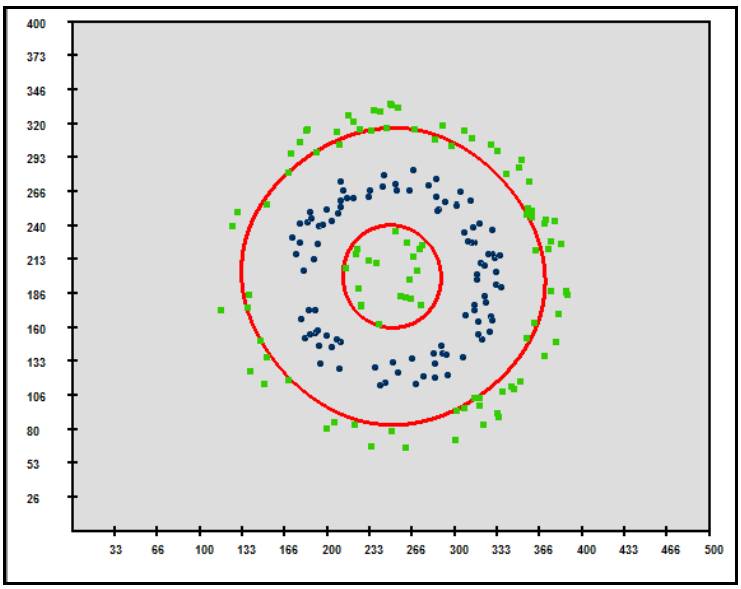

FIGURA 4.31 EXPRESIVIDAD DE LA REGRESION POLINOMIAL GRADO 4 APLICADA A LA DISTRIBUCIÓN DE CÍRCULOS CONCÉNTRICOS

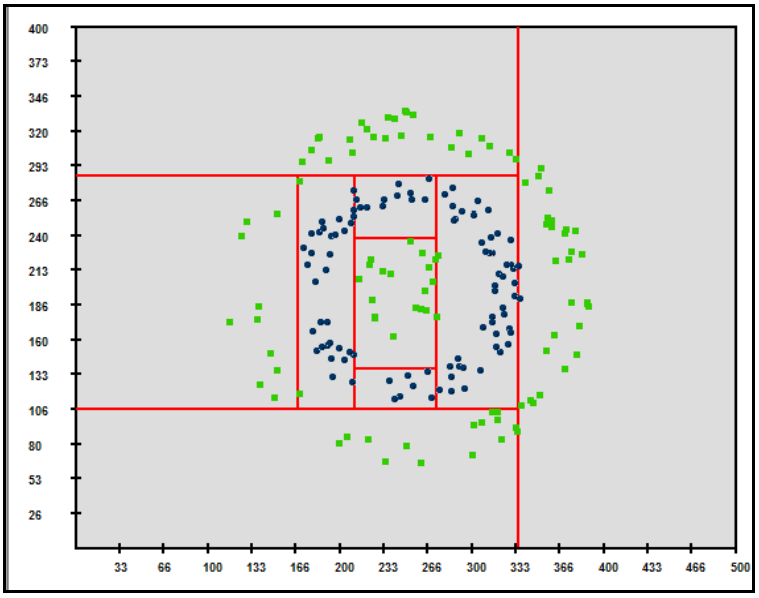

FIGURA 4.32 ÁRBOL DE DECISIÓN APLICADO A LA DISTRIBUCIÓN DE CÍRCULOS CONCÉNTRICOS