ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

Instituto de Ciencias Matemáticas

INGENIERÍA EN ESTADÍSTICA INFORMÁTICA

"Diseño e implementación de un aplicativo web para el aprendizaje de análisis discriminante"

TESIS DE GRADO

Previa a la obtención del título de:

INGENIERO EN ESTADÍSTICA INFORMÁTICA

Presentada por:

César E. Noboa Cisneros

GUAYAQUIL – ECUADOR

AÑO

2008

INTRODUCCIÓN

El modelo en el cual los resultados de una investigación constituyen el punto de partida para otra es un claro modelo de desarrollo, particularmente en el campo científico. El uso de Internet como medio de difusión del conocimiento está favoreciendo notablemente este desarrollo. Un muy sencillo ejemplo de esto es la publicación en Internet de *Sylvester* un programa (*conocido como librería*) creado por James Coglan útil para la manipulación de vectores y matrices. Un programador que obtenga esta *librería* podrá calcular determinantes, inversas y otras operaciones con matrices con relativa facilidad para sus propios proyectos.

A menudo la información que viaja a través de Internet rompe las barreras geográficas. Personas de diferentes partes del mundo pueden aprender acerca de un tópico en particular sin sujetarse a un horario. La BBC (*British Broadcasting Corporation*) de Londres, por ejemplo, permite por medio de Internet aprender inglés a personas de distinta raza y lengua. Tal parece

que los efectos de una página web publicada en Internet podrían ir mucho más allá de lo que jamás antes se hubiera pensado.

En términos sencillos, una página web es un documento que puede contener entre otras cosas texto, imágenes, videos al igual que vínculos por medio de los cuales el usuario puede "viajar" a otras páginas. Las páginas web se encuentran en un sistema conocido como WWW (World Wide Web) el cual fue desarrollado por el inglés Tim Berners-Lee y el belga Robert Cailliau. El sistema WWW es uno de los servicios que se ofrecen en la red Internet. La páginas web pueden estar llenas de dinamismo de tal forma que respondan a diversas acciones ejecutadas por el usuario. Páginas web de esta naturaleza pueden ser utilizadas para fines pedagógicos donde estudiantes alrededor del mundo tengan la posibilidad de reforzar algún conocimiento adquirido por medios tradicionales. Bajo esta tónica es presentado el aplicativo web que en la presente tesis se ha desarrollado.

CAPITULO I

1. ANALISIS DISCRIMINANTE

En el sentido más amplio uno de los objetivos del análisis discriminante es inferir a cual de un grupo de k clases predeterminadas pertenece una instancia conociendo previamente las clases a la que pertenecen un grupo de n instancias anteriores, donde cada instancia solamente pertenece a una clase. Cada instancia es determinada por un conjunto de p variables que pueden ser nominales o numéricas.

La tabla I muestra un subconjunto de 8 registros provenientes de una base de datos hipotética, en este caso podríamos decir que cada registro es una instancia. Si se supone que el riesgo de enfermar de osteoporosis[1] o si la persona ya está enferma de osteoporosis se determina según un examen que se le realiza al paciente conocido como densitometría[2], tenemos que un paciente cualquiera puede estar en uno de los cuatro casos listados.

TABLA I

EJEMPLO DE REGISTROS PARA UN ANÁLISIS DISCRIMINANTE

No.	No. ID	Edad	Sexo	Raza predominante	Practica deporte	Dieta Balanceada	Nivel de Osteoporosis
1	11023	52	M	Negra	N	N	Riesgo Medio
2	15890	35	F	Blanca	S	S	Riesgo Bajo
3	22560	18	M	Asiática	S	N	Riesgo Bajo
4	15000	65	F	Blanca	N	N	Enfermo
5	12200	42	F	Negra	N	N	Riesgo Medio
6	14890	29	M	Indígena	S	N	Riesgo Medio
7	11534	39	M	Blanca	N	N	Riesgo Alto
8	12667	49	F	Negra	N	N	??

[1] La osteoporosis es una enfermedad que consiste en la pérdida constante de la densidad mineral ósea lo cual puede llevar a fracturas.
[2] La densitometría es un examen que mide la densidad mineral de los huesos.

Cada instancia en la tabla I está compuesta por una clase (nivel de osteoporosis) y 6 variables que se supone determinan la clase a la que pertenece cada instancia (paciente), aunque la variable categórica No. ID simplemente es una identificación por lo tanto no ejerce ninguna influencia.

Dada la muestra de los 7 primeros pacientes de la tabla I sería útil conocer a que clase pertenecería la instancia # 8 de la cual se conocen sus variables pero no su nivel de osteoporosis.

Al igual que el caso de la tabla I existen una serie de situaciones en las que el análisis discriminante es procedente.

Si se considera el caso particular de n instancias determinadas por 2 variables numéricas X y Y que pertenecen a una de 2 clases diferentes, se tiene que el par ordenado (X_i,Y_i) pertenece a la clase C_1 o a la clase C_2. La Figura 1.1 representa esta situación gráficamente.

Pudo haber ocurrido que al intentar determinar la clase (el nivel de osteoporosis) a la que pertenece la instancia del registro No. 8 del

ejemplo anterior se cometa un error. El análisis discriminante no está exento de la mala clasificación tal como se aprecia en la Figura 1.1 donde la curva discriminante falla en algunos de los puntos. Aunque los errores de mala clasificación pueden ocurrir, al discriminar se intenta minimizar la probabilidad de una mala clasificación (o maximizar la probabilidad de clasificar correctamente).

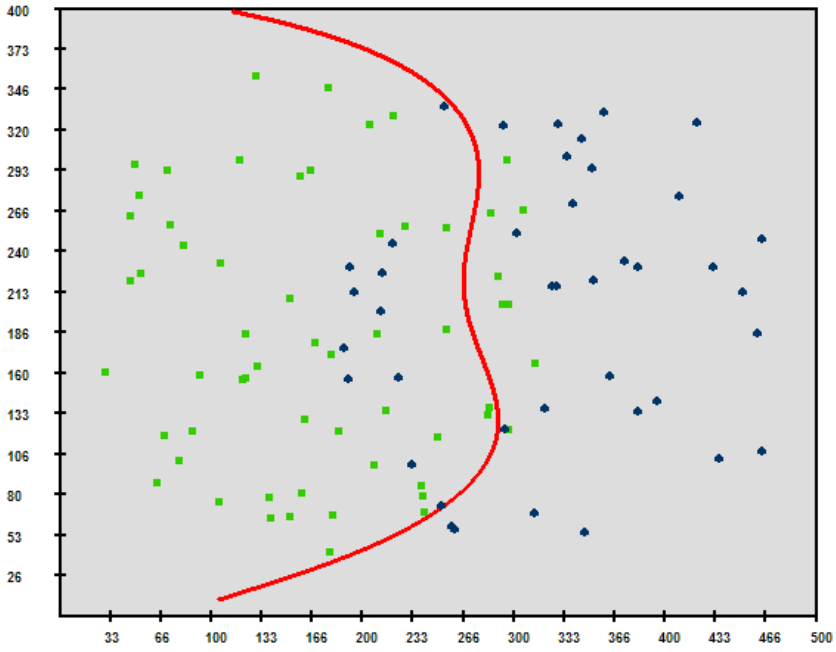


FIGURA 1.1 ILUSTRACIÓN DE UNA CURVA DISCRIMINANTE

A este respecto es útil el conocido teorema de Bayes:

$$P(A \setminus B) = \frac{P(B \setminus A) \cdot P(A)}{P(B)}$$

El cual da la probabilidad de un evento A dado que ha ocurrido otro evento B. En esta expresión P(A) se conoce como probabilidad a priori del evento A, pues es la probabilidad de que ocurra el evento A sin considerar que ya ha ocurrido el evento B, mientras que P(A\B) se conoce como probabilidad a posteriori.

Para el caso que se mencionó en la tabla I bien podríamos expresar la probabilidad P(A\B) del teorema de Bayes de la siguiente forma:

p=Prob(N= R. Alto \ B) =P(N=R. Alto\ X_1=49;X_2=F;X_3=Negra;X_4=N;X_5=N)

donde:

N: Nivel de osteoporosis

X_1: Edad

X_2: Sexo

X_3: Raza predominante

X_4: Practica deporte

X_5: Dieta balanceada

B: Evento en el que X_1=49; X_2=F; X_3=Negra; X_4=N y X_5=N a la vez.

Siguiendo el teorema de Bayes se tiene que:

$$p = \frac{P(X1 = 49; X2 = F; X3 = Negra; X4 = N; X5 = N \setminus N = Alto) \cdot P(N = Alto)}{P(X1 = 49; X2 = F; X3 = Negra; X4 = N; X5 = N)}$$

Donde $P(X_1,X_2,X_3,X_4,X_5)$ es la función de probabilidad conjunta.

Si se tuviesen las funciones para calcular las probabilidades de la expresión anterior se obtuviese un valor para Prob(N= Riesgo Alto \ B). De la misma forma se obtuviese las probabilidades Prob(N=R Medio \ B) Prob(N= R Bajo \ B) y Prob(N=Enfermo \ B). De estas 4 probabilidades se escogería la más alta es decir, aquel nivel de osteoporosis (o clase) que tiene la más alta probabilidad de ocurrir dado B.

Más adelante se verá el clasificador Naive Bayes que añade una suposición a este enfoque para producir resultados.

1.1 Minería de datos

Se podría decir que la minería de datos se encarga de la extracción de conocimiento a partir de datos. Los datos pueden provenir de grandes bases transaccionales de sistemas informáticos. El aprendizaje a partir de los datos no es una tarea sencilla e involucra algunos factores a considerar.

La minería de datos puede realizar algunas tareas con los datos tales como la discriminación, la categorización, la regresión, entre otras. De igual manera existen algunos métodos para realizar estas

tareas. En el presente trabajo se mencionan 5 métodos para realizar la misma tarea: la discriminación.

Puesto que no todos los métodos proporcionan los mismos resultados, existen métodos que son más apropiados para un cierto conjunto de datos antes que otros. Cada método tiene sus fortalezas y debilidades basadas en una serie de características que permiten al analista decidir cual o cuales de estos métodos utilizar. Algunas de estas características son las siguientes:

Precisión: En el caso del análisis discriminante es la capacidad que tiene el método de clasificar correctamente un nuevo elemento. Una forma natural de selección sería escoger el método que discrimine mejor para una determinada muestra. Para el análisis discriminante una medida de precisión de un método es el porcentaje de malas clasificaciones. Aquí, es necesario establecer la diferencia entre error de entrenamiento y error de prueba.

Error de entrenamiento**:** Un determinado método opera en base a una muestra de datos que se le proporciona, ya sea explícita o implícitamente basará sus respuestas en estos datos los cuales son conocidos como datos de entrenamiento. A partir de este conjunto

de entrenamiento el método intentará generalizar para datos futuros o no proporcionados. En el caso de discriminación un método no siempre logrará separar completamente los datos de entrenamiento, en la Figura 1.2 se muestra que el método ha clasificado incorrectamente 3 puntos azules. El error de entrenamiento será el porcentaje de malas clasificaciones en los datos de entrenamiento, en el caso de la figura será 8/40=20%=0.2

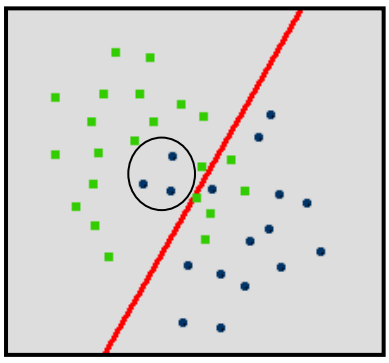


FIGURA 1.2 ERROR DE ENTRENAMIENTO

Aunque el error de entrenamiento es una medida del posible error del método no será adecuado utilizarlo porque por lo regular este error estará subestimado pues está ajustado a los datos.

Error de prueba: El procedimiento para evaluar el error de un método es tomar de la muestra total un conjunto de datos que el método no tome en cuenta, este conjunto conocido como conjunto de prueba podrá ser utilizado para evaluar el método una vez que

éste sea capaz de responder. Será más confiable evaluar el método con este conjunto de prueba puesto que el método no basa sus inferencias en este conjunto. En el caso de discriminación el error de prueba es el porcentaje de malas clasificaciones que realiza el método con los datos de prueba.

Estabilidad: Es la variabilidad que tiene el método a diferentes muestras de la misma población. El método es inestable cuando sus resultados varían significativamente al aplicarlo a una muestra diferente de la misma población.

Expresividad: Es la capacidad que tiene el método para capturar el patrón de los datos. En el caso de la discriminación, es la capacidad que tiene el método de separar los datos según la clase de cada miembro. Las figura 1.3 muestra dos grados diferentes de expresividad para la misma muestra. Es notorio en la figura que la curva discriminante de la izquierda no puede separar completamente los datos debido a su rigidez. Se puede decir entonces que el método aplicado a la derecha es más expresivo que el método de la izquierda.

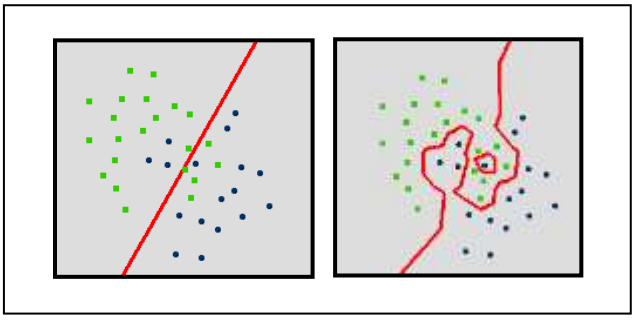


FIGURA 1.3 GRADOS DE EXPRESIVIDAD

Comprensibilidad: Aunque un método pueda realizar una tarea no necesariamente producirá un modelo que se pueda aplicar a futuras observaciones y en caso de que lo produzca éste no siempre resultará comprensible. En este punto es necesario mencionar que existe una división de los métodos en relación a la forma como proceden para dar resultados.

Métodos anticipativos: Estos métodos emplean largo tiempo en aprender o entrenarse pero poco tiempo en responder. Estos métodos emplean todo el conjunto de datos para generar algún tipo de modelo que sirva para decisiones futuras, en el momento que logran obtener el modelo (*aprender*), los datos ya no son más útiles y pueden ser desechados, debido a que el modelo será suficiente para responder a cualquier observación futura. Estos métodos son globales en el sentido de que los modelos que producen están basados en todo el conjunto de datos. Un método anticipativo, por ejemplo, es el método de regresión lineal que se verá más adelante.

Métodos retardados: Es posible que un método no produzca un modelo general que responda a cualquier observación futura, sino que en el momento que se le pregunte comience a procesar y responda. A diferencia de los anticipativos estos métodos utilizan las observaciones anteriores cada vez que se les pregunta, por lo tanto, el conjunto de datos no puede ser desechado. Estos métodos ocupan muy poco tiempo en entrenarse pero mucho tiempo en responder. Puesto que no producen un modelo general los métodos retardados son incomprensibles.

Se podría decir también que desde el punto de vista humano ciertos métodos son más comprensibles que otros, como se verá más adelante, el sistema de reglas que produce el árbol de decisión es muy comprensible.

Robustez al ruido o a outliers: Esta característica se refiere a la sensibilidad que tiene un método a los datos incorrectos, datos aberrantes o extremos (*outliers*). Se dice que un método es robusto a outliers, si los valores anómalas ejercen muy poca influencia sobre su funcionamiento.

Costo computacional: Esta característica se refiere a la eficiencia del aprendizaje de un método o el tiempo que tarda en responder a observaciones futuras. Aunque el tiempo que tardan los métodos anticipativos en generar un modelo puede ser alto cuando ya lo tienen el tiempo que tardan en responder a una observación futura es casi instantáneo. Sin embargo los métodos retardados casi no emplean ningún tiempo en entrenarse y su mayor tiempo lo emplean en responder a una observación futura. Dependiendo del número de datos el costo computacional de estos métodos suele ser mayor en relación al costo computacional de los métodos anticipativos.

1.2 Método de los K-Vecinos más cercanos.

El método de los k-vecinos más cercanos (knn de k-nearest neighbors) utiliza un sencillo criterio de clasificación, sin embargo, este método dependiendo del número de instancias puede demandar un considerable costo computacional.

Cuando se quiere inferir la clase a la que pertenece un nuevo ejemplo, como su nombre lo indica, este método busca los k

ejemplos más próximos a él (de los cuales, como ya se mencionó, ya se conocen sus clases) y le asigna al nuevo ejemplo la clase que más se repite en estos k ejemplos.

Para el caso particular en que cada ejemplo viene determinado por 2 variables éstos pueden ser graficados como puntos en un plano bidimensional (véase la Figura 1.4). Para un nuevo punto se pueden obtener sus k vecinos más cercanos utilizando la distancia euclídea entre 2 puntos y obtener la clase que más se repite en estos vecinos.

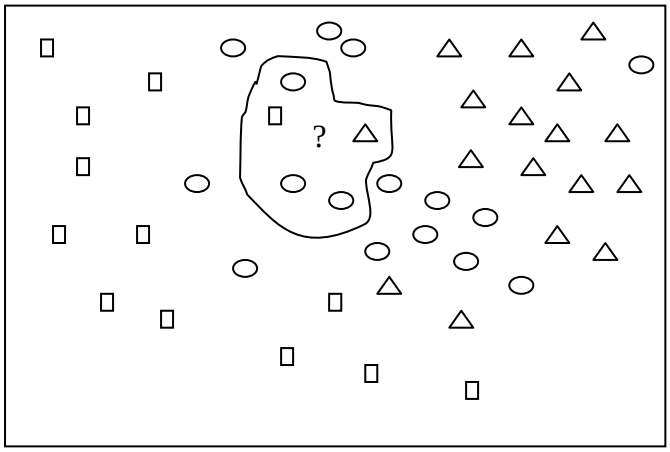


FIGURA 1.4 ILUSTRACIÓN DE LOS 5-VECINOS MÁS CERCANOS

Si se considera k=5 se observa en la figura 1.4 que el nuevo punto denotado por "?" cuya clase es desconocida tiene 1 vecino de

clase “triángulo”, 1 vecino de clase “cuadrado” y 3 vecinos de clase “elipse”. Por lo tanto se inferirá que este punto es de tipo “elipse”.

El método de los k vecinos es bastante flexible y es capaz de trazar curvas discriminantes bastante irregulares, sin embargo cuando k=1 este método puede trazar curvas que no clasifiquen tan correctamente nuevos ejemplos puesto que no toma en cuenta la densidad o la región donde se encuentra el ejemplo. En la figura 1.5 se puede observar este hecho. El nuevo punto “?” en la figura se lo clasifica como “elipse “ a pesar de que se encuentra en una región poblada por puntos de tipo “triángulo”.

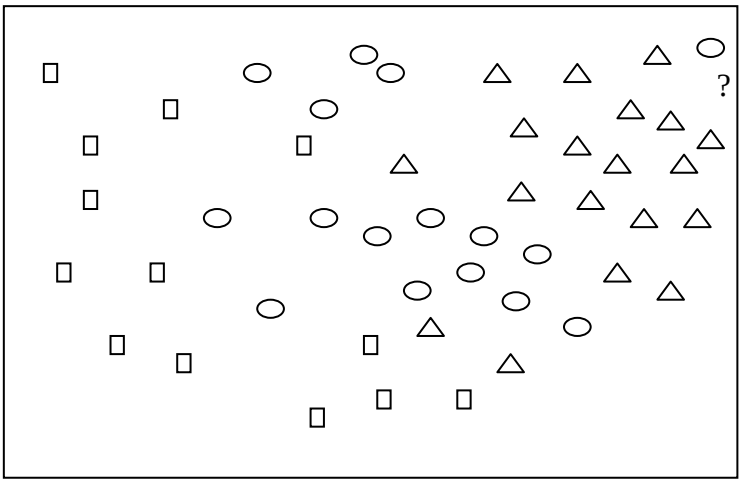


FIGURA 1.5 LIMITACIÓN DE K-VECINOS PARA K=1

1.3 Método de Naive Bayes Kernel

En una sección anterior se expuso el uso del teorema de Bayes en el análisis discriminante. En resumen se mencionó que si se tienen k clases { C_1,C_2,C_3,……,C_k } y n variables { X_1,X_2,…..,X_n } se pueden obtener las probabilidades condicionales de obtener cada clase dado que el vector X=(X_1,X_2,X_3,……,X_n) toma los valores {a_1,a_2,a_3,…..,a_n} de la siguiente forma:

$$P(C_1 \setminus a_1, a_2, a_3, ..., a_n) = \frac{P(a_1, a_2, a_3, ..., a_n \setminus C_1)P(C_1)}{P(a_1, a_2, a_3, ..., a_n)}$$

$$P(C_2 \setminus a_1, a_2, a_3, ..., a_n) = \frac{P(a_1, a_2, a_3, ..., a_n \setminus C_2)P(C_2)}{P(a_1, a_2, a_3, ..., a_n)}$$

$$P(C_3 \setminus a_1, a_2, a_3, ..., a_n) = \frac{P(a_1, a_2, a_3, ..., a_n \setminus C_3)P(C_3)}{P(a_1, a_2, a_3, ..., a_n)}$$

……

$$P(C_k \setminus a_1, a_2, a_3, ..., a_n) = \frac{P(a_1, a_2, a_3, ..., a_n \setminus C_k)P(C_k)}{P(a_1, a_2, a_3, ..., a_n)}$$

La clase C que produce la probabilidad condicional máxima es la que se elige como clase de pertenencia del elemento {a_1,a_2,a_3,…..,a_n}. Puesto que el denominador de las expresiones anteriores es el mismo se puede prescindir de éste para calcular la probabilidad máxima, es decir las siguientes 2 expresiones producen el mismo resultado:

$$Max\left\{\frac{P(a_1,a_2,a_3,....,a_n \setminus C_1)P(C_1)}{P(a_1,a_2,a_3,....,a_n)},.....,\frac{P(a_1,a_2,a_3,....,a_n \setminus C_k)P(C_k)}{P(a_1,a_2,a_3,....,a_n)}\right\}$$

$$Max\{P(a_1,a_2,a_3,....,a_n \setminus C_1)P(C_1),.....,P(a_1,a_2,a_3,....,a_n \setminus C_k)P(C_k)\}$$

Como ya se mencionó $P(a_1,a_2,a_3,.....,a_n \setminus C_i)$ es la probabilidad conjunta dado que pertenecen a C_i. El método de Naive Bayes supone que las variables $\{X_1,X_2,.....,X_n\}$ son condicionalmente independientes, por lo tanto:

$P(X_1,X_2,X_3,.....,X_n \setminus C_i)=P(X_1 \setminus C_i)P(X_2 \setminus C_i)P(X_3 \setminus C_i).....P(X_n \setminus C_i)$.

Con esta suposición la expresión anterior se simplifica a:

$$Max\left\{P(C_1)\prod_{i=1}^{n}P(a_i \setminus C_1),P(C_2)\prod_{i=1}^{n}P(a_i \setminus C_2),....,P(C_k)\prod_{i=1}^{n}P(a_i \setminus C_k)\right\}$$

Si se está tratando con valores numéricos no es común que se conozca las funciones de probabilidad de cada variable, por ello para estimar la función de densidad de las variables se hace uso de los estimadores núcleo K(x).

Una función de densidad f(x) se puede estimar haciendo uso de un estimador núcleo K(x) de la siguiente manera:

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

El cuadro 1.1 contiene algunas opciones para las funciones núcleo K(x) donde la función I(u) es como se indica en el cuadro.

El parámetro h de esta estimación requiere una muy especial atención. Si se trata de la función núcleo G(u) del cuadro 1.1 conocida como función núcleo gaussiano un posible valor de h podría ser:. $h_G = (1.06s)\cdot n^{-0.2}$

Donde s es la desviación estándar de la muestra.

Para la función núcleo Q(u) del cuadro 1.1 un valor de h podría ser:

$$h_Q = 2.62(1.06s)\cdot n^{-0.2}$$

$$K(u) = \frac{1}{2} I(|u| \le 1)$$

$$K(u) = (1 - |u|) I(|u| \le 1)$$

$$K(u) = \frac{3}{4}\left(1 - u^2\right) I(|u| \le 1)$$

$$Q(u) = K(u) = \frac{15}{16}\left(1 - u^2\right)^2 I(|u| \le 1)$$

$$G(u) = K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$$

$$I(|u| \le 1) = \begin{cases} 1, & si \;\; |u| \le 1 \\ 0, & caso \;\; contrario \end{cases}$$

CUADRO 1.1 ALGUNAS FUNCIONES NÚCLEO K(X) UTILIZADAS

Utilizando la función núcleo para estimar las funciones de densidad la expresión para la máxima probabilidad se podría reescribir así:

$$Max\left\{\hat{P}(C_1)\prod_{i=1}^{n}\hat{f}(a_i \setminus C_1), \hat{P}(C_2)\prod_{i=1}^{n}\hat{f}(a_i \setminus C_2), ..., \hat{P}(C_k)\prod_{i=1}^{n}\hat{f}(a_i \setminus C_k)\right\}$$

$$donde \quad \hat{P}(C_i) = \frac{\#\ de\ elementos\ \ de\ la\ \ clase\ \ C_i}{n}$$

1.4 Método de Regresión lineal

Este método pretende modelar el comportamiento de una variable respuesta en función de un conjunto de n variables explicativas $X_1, X_2, X_3, ..., X_n$. Si se considera el caso simple de una variable respuesta y una sola variable explicativa X, se podría pensar que un modelo para este caso sería $Y=f(X)+Є$ donde Є es el componente aleatorio del modelo. La función f(X) podría ser lineal $f(X)= \beta_0+\beta_1X$, podría ser cuadrática $f(X)= \beta_0+\beta_1X+ \beta_2X^2$ o en general polinómica de grado n con $f(X)= \beta_0+\beta_1X+ \beta_2X^2+……+ \beta_nX^n$.

Si se tienen más de 1 variable explicativa los módelos fácilmente aumentan de complejidad. En el caso de 2 variables explicativas X_1 y X_2 un modelo cuadrático sería $f(X)= \beta_0+\beta_1 X_1+ \beta_2 X_2+ \beta_3X_1X_2+ \beta_4(X_1)^2+ \beta_5(X_2)^2$ el cual como se aprecia tiene 6 términos y un polinomio de grado 4 tendrá 15 términos.

El objetivo de este método es estimar el vector $\beta=(\beta_0, \beta_1, \beta_2,\ldots, \beta_k)$ tal que la siguiente expresión sea mínima:

$$\sum_{i=1}^{n}\left(y_i - f(x_{i1}, x_{i2}, x_{i3},...., x_{im})\right)^2$$

En la expresión anterior se considera que existen n valores tanto para la variable respuesta como para cada variable explicativa. También se observa que existen m variables explicativas.

Considerando el caso particular en que solamente existe una variable explicativa o independiente X y se plantea un modelo cuadrático, con n valores para las variables X y Y se tiene un conjunto de n ecuaciones, donde b_i es el estimador de β_i que se desea encontrar

$$\begin{cases} y_1= b_0+b_1 x_1+ b_2 (x_1)^2 \end{cases}$$

$y_2 = b_0 + b_1 x_2 + b_2 (x_2)^2$

$y_3 = b_0 + b_1 x_3 + b_2 (x_3)^2$

.....

$y_n = b_0 + b_1 x_n + b_2 (x_n)^2$

La versión matricial del sistema anterior sería:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \cdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ 1 & x_3 & x_3^2 \\ \cdots & & \\ 1 & x_n & x_n^2 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix}$$

Lo cual podría ser reescrito como Y=XB considerando que Y,X y B son matrices. La solución del sistema matricial que considera el criterio de mínimos cuadrados que se mencionó anteriormente es $B=(X'X)^{-1}X'Y$ donde X' es la matriz transpuesta de X.

Anteriormente se expuso que en análisis discriminante se intenta clasificar un elemento determinado por n variables en una de k clases posibles. Se podría plantear este problema bajo el enfoque de regresión lineal suponiendo que las n variables que determinan un elemento son las variables explicativas y la variable de respuesta

Y es una medida que es útil para clasificar de donde proviene el elemento.

Si se considera el caso general en que se tienen k clases posibles y n variables que determinan cada elemento podría resolverse el problema planteando k modelos de regresión del tipo $Y_i = f(X_1, X_2, X_3,\ldots\ldots, X_n)$. Es decir un modelo de regresión por cada clase. Puesto que las clases son valores no cuantitativos, la variable respuesta Y_i del modelo de regresión $f(X_1, X_2, X_3,\ldots\ldots, X_n)$ se establecerá a 1 si es que el elemento $(X_1, X_2, X_3,\ldots\ldots, X_n)$ pertenece a la clase i y se establecerá a 0 en caso contrario. De esta manera se podrán encontrar las estimaciones de los β's para cada modelo de regresión.

Utilizando el enfoque anterior para clasificar un nuevo elemento con variables $(X_1, X_2, X_3,\ldots\ldots, X_n)$ se deberá evaluar los valores de este nuevo elemento en cada una de los k modelos de regresión. Se inferirá que este elemento pertenece a la clase i si su correspondiente variable respuesta Y_i es la máxima dentro de todas las variables respuesta.

Si se considera el caso particular en que un determinado elemento solamente puede pertenecer a 2 clases posibles, se tiene el caso en que la variable respuesta Y es binaria. En este caso se podría generar un solo modelo de regresión en el cual la variable Y se la establecería como 1 si el elemento o vector $(X_1, X_2, X_3, \ldots\ldots, X_n)$ pertenece a una clase y 0 si pertenece a la otra. Una política para clasificar un nuevo elemento, dado que se han encontrado las estimaciones de los coeficientes β's del modelo de regresión propuesto es evaluar los valores de las variables del nuevo elemento en el modelo de regresión, si la variable respuesta Y es mayor o igual a 0.5 se inferirá que el nuevo elemento pertenece a la clase que se codificó como 1 al plantear el problema, caso contrario se inferirá que pertenece a la otra clase.

1.5 Método de regresión logística

Para este método se considerará el caso en el que la variable respuesta es binaria.

Si se desea aplicar regresión lineal a un conjunto de datos en donde la variable respuesta Y es binaria, 1 ó 0, una pregunta que surgiría sería si E[Y]=Prob(Y=1). En el caso de regresión lineal es posible que los valores $\hat{Y}_i$ sean menores que 0 o mayores que 1 lo cual no puede ser interpretado como una probabilidad. La regresión logística es un método que "acomoda" la variable respuesta para que ésta "caiga" en el intervalo [0,1].

En regresión lineal se plantearon modelos del tipo:

$$Y_i=\beta_0+\beta_1X_{1i}+ \beta_2X_{2i}+ \beta_3X_{3i}+\ldots\ldots+ \beta_nX_{mi}$$

Donde i=1,2….,n y hay m variables explicativas.

Para regresión logística se planteará el siguiente modelo:

$$\ln\left(\frac{p_i}{1-p_i}\right)=\beta_0+\beta_1X_{1i}+\cdots\cdots+\beta_nX_{mi}$$

Donde p_i es igual a Prob(Y_i=1). El modelo de regresión logística anterior también puede ser escrito como:

$$p_i=\frac{1}{1+e^{-(\beta_0+\beta_1X_{1i}+\cdots\cdots+\beta_nX_{mi})}}$$

En esta última expresión se puede observar que a diferencia de la regresión lineal en la regresión logística sean cuales sean las estimaciones para $\beta_0,\beta_1,\beta_2,\beta_3,\ldots,\beta_m$ y sean cuales sean los valores

de X_1, X_2, X_3,……, X_m la respuesta p_i siempre estará en el intervalo [0,1].

Puesto que p_i es la Prob(Y_i=1), si ésta es mayor que 0,5 el elemento cuyas variables se evaluaron pertenecerá a una clase determinada, si p_i<0.5 pertenecerá a la otra. También se puede estar interesado en graficar la curva discriminante (cuando es posible) o al menos conocer la ecuación algebraica de esta curva. El límite de discriminación, es decir la frontera entre una clase y la otra ocurrirá cuando p_i=0.5. Para hallar la ecuación de la curva discriminante se reemplaza p_i=0.5 en el modelo:

$$p_i = \frac{1}{1+e^{-(\beta_0+\beta_1 X_{1i}+\cdots+\beta_n X_{mi})}}$$

Quedando:

$$0.5 = \frac{1}{1+e^{-(\beta_0+\beta_1 X_{1i}+\cdots+\beta_n X_{mi})}}$$

Esto ocurrirá cuando $\beta_0+\beta_1 X_{1i}+ \beta_2 X_{2i}+ \beta_3 X_{3i}+\ldots\ldots\ldots+ \beta_n X_{mi} = 0$. Siendo ésta la ecuación de la curva discriminante.

Para el caso particular en que existen sólo 2 variables explicativas, la curva discriminante podrá visualizarse en un plano bidimensional como se verá más adelante.

Para estimar los parámetros del modelo logístico se utiliza el criterio general de máximo-verosimilitud lo que desemboca en la siguiente ecuación matricial iterativa:

$B^{t+1}=(X'WX)^{-1}X'Wz$, donde $z=XB^t+W^{-1}(Y-P)$, donde B^t es el estimador del vector de coeficientes β en la iteración t.

La ecuación matricial para hallar B^{t+1} es bastante similar a la expresión $B=(X'X)^{-1}X'Y$ que es la solución para encontrar los estimadores de los coeficientes en el método de regresión lineal que expuso anteriormente.

Una de las diferencias es la presencia de la matriz W, conocida como matriz de ponderación que se define como:

$$W=\begin{bmatrix} p_1(1-p_1) & & & \\ & p_2(1-p_2) & & \\ & & \cdots & \\ & & & p_n(1-p_n) \end{bmatrix}$$

Una de las suposiciones de la regresión lineal $Y_i = f(X_i)+ \epsilon_i$ es que $Var[\epsilon_i]=\sigma^2$, es decir la varianza de los errores se la supone constante para cada observación. Sin embargo, esta suposición no siempre es realista puesto que es posible que esta varianza cambie conforme los valores de X_i cambian. En el modelo de regresión

logística se supone que $Var[\epsilon_i]=\sigma_i^2$, lo cual indica que la varianza depende de la observación i. La presencia de la matriz W se justifica debido a esta última suposición puesto que Y_i es una variable aleatoria binomial con media $E[Y_i]=1(pi)+0(1-pi)=p_i$ y varianza $Var[Y_i]=p_i(1-p_i)$, siendo estas varianzas los pesos de la matriz W.

Para encontrar $B^{t+1}=(X'WX)^{-1}X'Wz$, donde $z=XB^t+W^{-1}(Y-P)$ se aplica el siguiente método iterativo:

1) Se inicia con $B^0=0$ y $p_i=0.5$ para todos los i. Con estos valores iniciales se construye la matriz W y se encuentra la matriz respuesta z.
2) Se calcula B^1 con la matriz z anterior.
3) Se utiliza los β's de B^1 para calcular las nuevas probabilidades p_i utilizando el mismo modelo: $p_i = \frac{1}{1+e^{-(\beta_0+\beta_1 X_{1i}+\cdots+\beta_n X_{mi})}}$
4) Se calcula nuevamente la matriz respuesta z.
5) Se calcula B^2 y se continúa con el proceso iterativo hasta que los valores de B converjan.

Aunque la convergencia no se garantiza, es usual que ocurra.

1.6 Árbol de decisión.

El árbol de decisión es un método que realiza la tarea de clasificación mediante segmentaciones sucesivas del conjunto de datos.

La figura 1.6 muestra el esquema que en general sigue un árbol de decisión al realizar la tarea de discriminación para el caso de 2 clases.

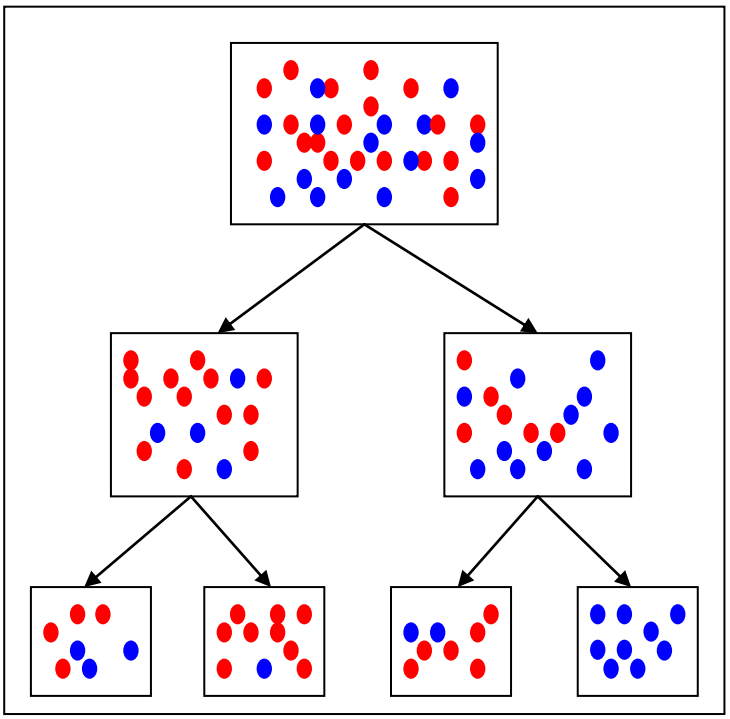


FIGURA 1.6 ESQUEMA DE LA DISCRIMINACIÓN UTILIZANDO UN ARBOL DE DECISIÓN

Del árbol de decisión del esquema anterior se derivan algunas observaciones:

- El nodo raíz contiene todos los elementos.
- Cada segmentación produce hijos que son más puros que el padre, es decir, que tienen menos diversidad. En el esquema anterior se nota que en el hijo izquierdo del nodo raíz el porcentaje de elementos azules es menor que en el caso de su padre, en este sentido es más puro. De igual manera en el hijo derecho el porcentaje de elementos rojos es menor que en el caso de su padre.
- Un nodo cualquiera deja de tener hijos cuando es totalmente puro, es decir tiene el 100% de sus elementos de una clase determinada. En este caso se dice que este nodo es una hoja del árbol. En el esquema anterior se observa que 3 de los 4 nodos finales pueden seguirse segmentando. En la práctica no siempre se exige que todas las hojas del árbol sean 100% puras, sino que se puede establecer una condición de parada, como por ejemplo, que el número de elementos mínimo que tenga un nodo para que pueda tener hijos (ser segmentado) sea un tanto por ciento del número total de elementos.

En el esquema del árbol de decisión presentado anteriormente se supone que cada elemento (punto rojo o azul) tienen una cierta cantidad de atributos o variables que pueden ser nominales o numéricos. Cada segmentación solamente se la hará en base a un atributo a la vez. Aunque el árbol del esquema anterior es binario, una partición en el caso nominal puede generar más de 2 hijos.

Si un atributo Xi es numérico continuo y puede tomar valores en el intervalo [a,b] entonces las 2 condiciones que segmentarán un nodo en base a ese atributo serán las siguientes: $Xi \leq k$ y $Xi > k$ donde $k=(v+w)/2$ donde v y w son valores consecutivos en todo el conjunto de valores que toma el atributo Xi.

De esta manera si se supone que el árbol del esquema anterior tiene elementos con 2 atributos numéricos continuos X y Y un resultado podría ser el siguiente:

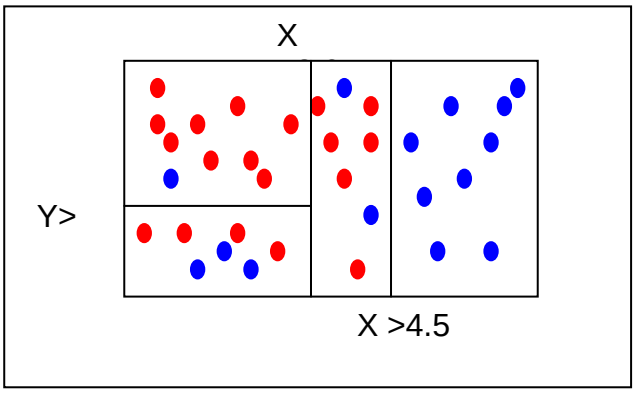


FIGURA 1.7 RESULTADO DE UNA DISCRIMINACIÓN UTILIZANDO UN ÁRBOL DE DECISIÓN

Considerando las 3 particiones de la figura 1.7 entonces el árbol de la figura 1.6 quedaría como se muestra en la figura 1.8.

Suponiendo que los puntos de la figura 1.7 son los datos de entrenamiento, entonces una de las reglas de generalización que producirá el árbol de decisión entrenado será: *si X>4.5 entonces el elemento es de la clase "azul"*. De esta manera si se quisiese conocer la clase de un nuevo elemento cuyo atributo X=4.8 por ejemplo, entonces utilizando la regla anterior se inferirá que este nuevo elemento es de la clase "azul".

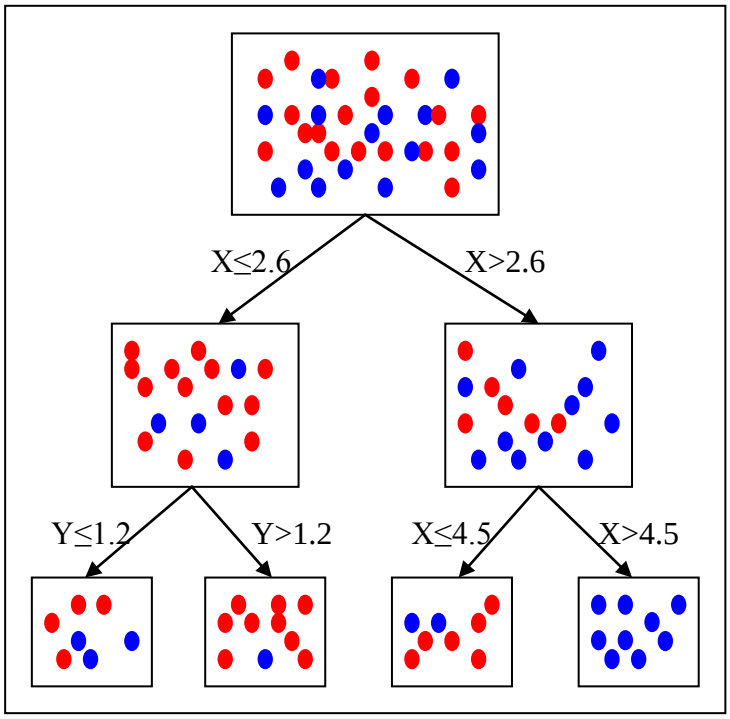


FIGURA 1.8 ESQUEMA DE UN ÁRBOL DE DECISIÓN CON ATRIBUTOS NUMÉRICOS

En el árbol anterior la primera partición se baso en el atributo X y no en el atributo Y. En la condición se escogió la constante 2.6 con la cual comparar el atributo X y no se escogió otra constante como 3.2 por ejemplo. La razón por la cual se lo hizo así es porque la partición que utiliza la condición X>2.6 es la óptima de acuerdo a un criterio determinado.

Existen distintos criterios para seleccionar la "mejor" partición, los cuales utilizan funciones de impureza del tipo $f(p_{1j}, p_{2j}, p_{3j},\ldots, p_{cj})$ donde p_{ij} es la proporción de elementos de la clase i en el nodo j. Entre algunos criterios con sus funciones de impureza tenemos los siguientes:

- Error esperado: $f(p_{1j}, p_{2j}, p_{3j},\ldots, p_{cj})=\min(p_{1j}, p_{2j}, p_{3j},\ldots, p_{cj})$
- GINI: $f(p_{1j}, p_{2j}, p_{3j},\ldots, p_{cj}) = 1- \sum(p_{ij})^2$
- Entropía: $f(p_{1j}, p_{2j}, p_{3j},\ldots, p_{cj}) = \sum p_{ij} \log(p_{ij})$
- DKM: $f(p_{1j}, p_{2j}, p_{3j},\ldots, p_{cj}) = 2 (\prod p_{ij})^{1/2}$

Cuando se trata de atributos numéricos por cada partición se producen 2 nuevos nodos del árbol, según el criterio que se seleccione se podrá calcular la función de impureza para cada nodo. De esta forma para cada posible partición se podrá obtener el promedio ponderado de la impureza de sus 2 hijos, de la siguiente manera:

$$\text{Impureza(partición)} = p_i \, f(p_{1i}, p_{2i}, p_{3i}, \ldots, p_{ci}) + p_d \, f(p_{1d}, p_{2d}, p_{3d}, \ldots, p_{cd})$$

Donde p_i es la proporción de elementos del nodo padre que pertenecen al hijo izquierdo y p_d es la proporción de elementos que pertenecen al hijo derecho. Entonces, de todas las posibles particiones se escogerá aquella que produzca la menor impureza ponderada.

El tratamiento anterior podrá extenderse cuando no se trate de árboles binarios simplemente obteniendo el promedio ponderado de las impurezas de los hijos que se produzcan con cada partición.

CAPÍTULO II

2. JAVASCRIPT (JSCRIPT)

2.1 Introducción

JavaScript es un lenguaje de programación creado para programar páginas Web. Aunque quizás su nombre pudiera sugerir lo contrario, Java y Javascript son lenguajes de programación diferentes y fueron pensados para diferentes propósitos.

Por una parte, Java es un potente aunque más complejo lenguaje de programación orientado a objetos creado por la empresa Sun Microsystems en la década de 1990. Este lenguaje es de propósito general lo cual significa que con éste se pueden crear aplicaciones de distinta naturaleza que no tengan ninguna relación con Internet.

Por otro lado, JavaScript es un modesto aunque sencillo lenguaje de programación destinado únicamente para la manipulación de páginas web. En sus inicios la compañía Netscape desarrolló este lenguaje bajo el nombre de LiveScript, sin embargo, poco después Netscape formó una alianza con Sun Microsystems (creador de Java) para desarrollar JavaScript. De esta alianza surgió un lenguaje más sencillo de utilizar que Java y útil para dotar de dinamismo a las aplicaciones web. De esta manera, el explorador Netscape 2.0 fue el primero capaz de interpretar código JavaScript.

La compañía Microsoft ha creado su propia versión de JavaScript llamada JScript la cual es empleada para programar y ejecutar páginas Web en uno de los exploradores de mayor uso: Internet Explorer.

JavaScript no es el único lenguaje en el que se puede programar una página web. VBScript (Visual Basic Script) es otro programa que puede ser utilizado para este propósito, sin embargo, aquí se tratará acerca de algunas características de JavaScript.

El conocido lenguaje HTML (Hyper Text Markup Language) que se traduce como Lenguaje de marcas hipertextuales es un lenguaje de presentación utilizado para estructurar el contenido de páginas web. Utilizando HTML es posible crear botones, colocar imágenes, etiquetas, cuadros de texto, crear hipervínculos, entre otras cosas. Sin embargo, en ocasiones el sólo conocimiento de HTML no es suficiente para darle a una página web la interactividad y flexibilidad que se requiere. En cambio el conocimiento de HTML unido al de JavaScript es una útil combinación que puede ayudar a los programadores a desarrollar interesantes páginas web.

JavaScript maneja las mismas nociones de otros lenguajes que manejan programación estructurada como Pascal y C++, entre otras cosas maneja estructuras de repetición, estructuras de decisión, funciones, variables, recursividad, etc. Los programadores de C++ encontrarán que existe similitud en la sintaxis de su programa con la

de JavaScript. Sin embargo, a diferencia de Pascal y C++ en JavaScript no es posible definir directamente una matriz de nxm ni definir el tipo de una variable de forma explícita.

Una de las cosas interesantes de los lenguajes HTML y JavaScript es que sus códigos pueden ser escritos sin ninguna dificultad en un archivo de texto. Es decir, un programador puede escribir su "sofisticado" código en el simple Block de Notas de Microsoft Windows. Lo único que necesita hacer al terminar de escribir su código es cambiar la extensión del archivo de *.txt a *.html y ya tendrá una página Web.

Existen programas especializados para escribir códigos de páginas Web como Front Page de Microsoft, Microsoft Script Editor o DreamWeaver, pero éstos no son imprescindibles para programar aunque sí brindan una serie de facilidades en relación al sencillo Block de Notas.

Por otro lado, una de las desventajas de la programación de código script es que la correcta ejecución de un programa depende del explorador del cliente. Lamentablemente los exploradores no tienen

un estándar por ello es posible que un código funcione perfectamente en Mozilla Firefox pero presente problemas al querer ser ejecutado en Internet Explorer, por ejemplo. Las instrucciones o funciones que un explorador reconoce no siempre serán reconocidas en otro. Es por esto que es conveniente que los programadores prevengan esta situación diseñando sus códigos de tal manera que dependiendo del tipo de navegador del cliente se ejecute una versión de sus páginas.

2.2 Referencia rápida

Para adherir lenguaje ya sea JavaScript o VBScript en un código html se lo realiza utilizando el elemento <script language="lenguaje">. Dentro de este elemento puede escribirse todas las instrucciones que se desee como lo muestra el simple código siguiente:

- <HTML>
- <HEAD></HEAD>
- <BODY>
- <script language="javascript">
- <!--

- alert("JavaScript");
- //-->
- </script>
- </BODY>
- </HTML>

Al igual que otros elementos de HTML el elemento <script> debe cerrarse con la instrucción </script>. De esta manera se puede alternar dentro de un mismo programa código HTML y código JavaScript, con la facilidad de poder escribir código JavaScript en cualquier parte del programa utilizando las instrucciones <script></script>. En el código anterior se utiliza la instrucción alert(“JavaScript”), ésta permite mostrar mensajes por pantalla y es bastante útil en la depuración de programas. Si, por ejemplo, un programador no estuviera seguro del valor que toma la variable p en una parte del programa, puede incluir la instrucción alert(p).

Las variables de JavaScript son sensibles a las mayúsculas, es decir la variable p será diferente de la variable P. En la sintaxis, Javascript tiene similitudes con C++. Aquí se verán algunas de las

instrucciones y características que han sido utilizadas para el desarrollo de la página web que se mostrará en el capítulo III.

Estructuras de selección y repetición

La estructura de selección que se utiliza es la *if . . . [else]*, ésta permite ejecutar una instrucción o un grupo de instrucciones si una determinada condición es verdadera y caso contrario ejecutar otra u otras instrucciones. La sintaxis para esta estructura condicional es la siguiente:

- if (condicion)
- { instrucciones;}
- [else
- { instrucciones;}]

Un programador de C++ reconocerá que esta instrucción es exactamente la misma que en su lenguaje. Los corchetes [ ] presentes en la sintaxis anterior indican que la instrucción *else* es opcional.

Las estructuras de repetición utilizadas son la condicional *do...while* y la instrucción *for*. La primera estructura permite repetir una instrucción o un bloque de instrucciones mientras una condición sea

verdadera. La segunda permite repetir una instrucción o un bloque de instrucciones mientras la variable relacionada a la estructura cumpla una determinada condición, esta estructura es utilizada para tareas comunes como la lectura de un arreglo. La sintaxis de la estructura *do . . . while* es la siguiente:

- do
- {
- instrucción 1;
- instrucción 2;
- ……
- instrucción n;
- }
- while (condición);

La sintaxis de la instrucción *for* es la siguiente:

- for (i=Vinicial; condición; incremento/decremento)
- {
- Instrucción 1;
- Instrucción 2;
- ……
- Instrucción n;
- }

En la sintaxis anterior i es la variable relacionada al bucle, ésta es utilizada para evaluar una condición que determina si el ciclo debe detenerse e incremento o decremento hace que la variable i aumente o disminuya según sea el caso. Este incremento o decremento no tiene que ser de 1 en 1. El siguiente ejemplo muestra 4 mensajes con los números 10,7,4 y 1 respectivamente:

- <HTML>
- <BODY>
- <script language="javascript">
- <!--
- for (i=10;i>0;i=i-3)
- alert(i);
- //-->
- </script>
- </BODY>
- </HTML>

La instrucción *for* del código anterior hace que la variable i vaya reduciendo su valor de 3 en 3 empezando en 10 y terminando en 1.

Puesto que JavaScript es sensible a las mayúsculas es necesario ser muy cuidadoso, si por ejemplo se escribe la siguiente instrucción:

- If (a>5) alert(“Es mayor”);

Internet Explorer mostrará error puesto que la instrucción es *if* y no *If*.

Declaración de variables:

En cuanto a la declaración de variables JavaScript es muy cómodo para el programador puesto que una variable no necesita declararse previamente antes de ser utilizada. Así mismo en la declaración de una variable no se especifica el tipo. Una variable puede ser declarada y recibir un valor al mismo tiempo o solamente ser declarada haciendo uso de la palabra reservada *var*, como en los siguientes ejemplos:

- var x;
- var numero=0;
- var tasa=0.25, mes="Enero", dia;

Como se muestra en las líneas anteriores varias variables pueden ser declaradas al mismo tiempo. También se puede escribir una asignación múltiple en una misma línea de código de la siguiente manera: *tipoA_xl=tipoA_lx=tipoB_xl=tipoB_lx=40;*. De esta manera la variable *tipoA_xl* tendrá el valor de 40 al igual que las demás y no será necesario haber declarado previamente estas variables.

Declaración de arreglos:

La declaración de arreglos al igual que la de variables no especifica tipo de dato y es muy cómoda porque el tamaño del arreglo se adapta a la cantidad de valores que se le haya ingresado.

Una forma sencilla de declarar los arreglos es la siguiente:

- var nombre_de_arreglo = new Array();

En la linea anterior se puede prescindir de la palabra reservada var.

Al asignarle a una variable *New Array()* lo que en realidad se está creando es un objeto de tipo *Array* con sus propios métodos y propiedades. Para referirnos al elemento i del arreglo “colores”, por ejemplo, se especifica colores[i]. Nuevamente, al igual que en C++ los arreglos en JavaScript empiezan con el subíndice 0, de tal manera que el primer elemento del arreglo colores es *colores[0]* y no *colores[1]*.

Una propiedad bastante útil de este objeto es *length* la cual devuelve el tamaño o cuantos elementos tiene un arreglo. Hay que

tomar en cuenta que si el arreglo *colores* tiene 5 elementos su propiedad *length* será 5 pero su último elemento será *colores[4]*. El código que sigue a continuación ilustra lo que hasta aquí se ha mencionado en relación a los arreglos:

```
<HTML>
<BODY>
<script language="javascript">
<!--
colores = new Array("orange","red","cyan","brown","purple");
cadena=colores[0];
for (i=1;i<colores.length;i++)
cadena=cadena+"-"+colores[i];
alert(cadena);
//-->
</script>
</BODY>
</HTML>
```

Las líneas anteriores que es posible declarar un arreglo con el valor de cada uno de sus elementos a la vez:

colores = new Array("orange","red","cyan","brown",”purple”);

De esta manera la propiedad *length* del nuevo objeto llamado colores será igual a 5. La variable cadena es inicializada con el

valor string *"orange"* y la estructura *for* permite que se le vayan añadiendo los demás colores. La figura 2.1 muestra el resultado de este código anterior.

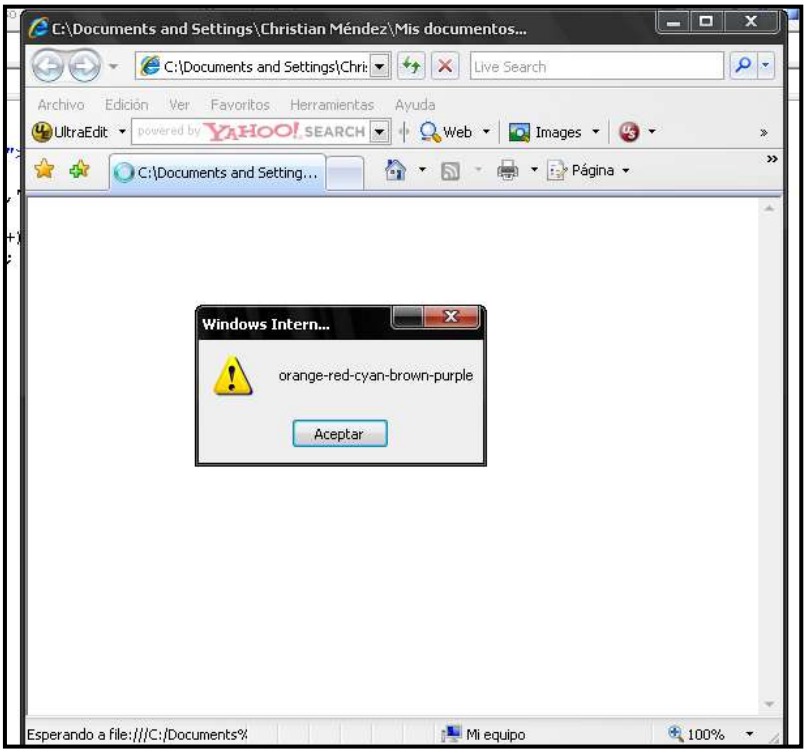


FIGURA 2.1 LECTURA DE UN OBJETO ARRAY()

El método write() del objeto document:

Una característica de JavaScript es que utilizando el método *write()* del objeto intrínseco *document* es posible escribir código HTML desde el propio código JavaScript. Por ejemplo si se deseara mostrar un botón se lo podría hacer comúnmente como sigue:

- <input type='button' value='Calcular' name='btnTexto' onclick='Mostrar()' ID=Button1>

Sin embargo también se podría crear el mismo botón desde el código JavaScript utilizando el método *write* de la siguiente forma:

- document.write("<input type='button' value='Calcular' name='btnTexto' onclick='Mostrar()' ID=Button1>")

Como se observa en el ejemplo anterior, utilizar el método *write* es como escribir directamente código HTML. Esto resulta muy útil puesto que se puede flexibilizar el código HTML. Por ejemplo, se puede crear un número variable de botones dependiendo del valor que toma una determinada variable.

La figura 2.2 muestra el código que flexibiliza la creación de botones que se mencionó anteriormente. Se observa que la cantidad de botones que se muestran depende de la variable *numero_de_botones*. La flexibilización va más allá, tanto la propiedad *value* como la propiedad *ID* de cada botón depende de la variable i.

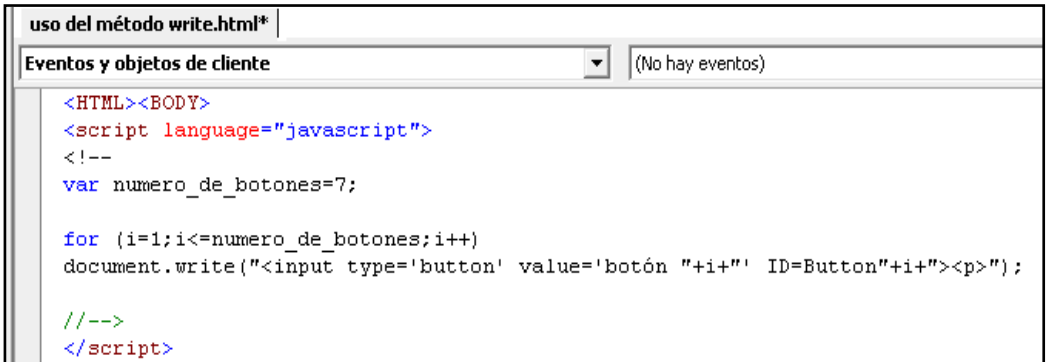


FIGURA 2.2 CÓDIGO PARA LA CREACIÓN FLEXIBLE DE BOTONES A TRAVES DEL METODO WRITE().

En el código anterior basta con cambiar el valor de la variable *numero_de_botones* a 12 para que se creen en pantalla esa cantidad de botones. Se nota entonces, que Javascript permite al programador ir un poco más allá en el desarrollo de páginas web. La figura 2.3 muestra la ejecución del código de la figura 2.2. El uso del elemento <p> al final de la instrucción document.write() permite que la creación de botones sea en columna y saltando un espacio.

Declaración de matrices:

Lamentablemente en JavaScript no es posible declarar directamente arreglos multidimensionales. Para obtener arreglos multidimensionales es necesario declarar un arreglo de arreglos, es

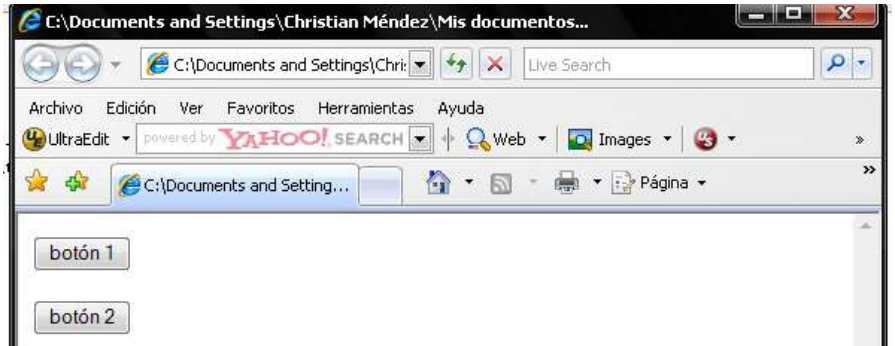


FIGURA 2.3 CREACIÓN FLEXIBLE DE BOTONES CON EL METODO WRITE().

decir un objeto de tipo Array donde cada elemento a su vez es un objeto Array. Esto se lo puede hacer de la siguiente forma:

a=new Array(new Array(),new Array(),new Array(),new Array());

Un ejemplo de la creación de una matriz de 3x2 por ejemplo, sería el siguiente:

a=new Array(new Array(2,4),new Array(3,6),new Array(41,10));

Para leer cualquier elemento una matriz se podrá utilizar la notación a[i][j] así, el elemento a[0][1] en el ejemplo anterior será igual a 4.

La figura 2.4 muestra un ejemplo más elaborado que ilustra el empleo de un arreglo de arreglos para simular matrices. En este ejemplo se crea una matriz de 4x20 con filas de 1's, 2's, 3's y 4's respectivamente. En el ciclo *for* anidado que está al final del código se nota que la variable i llega hasta 1 menos que *a.length* mientras que j llega hasta 1 menos *a[i].length*. La propiedad *length* puede ser utilizada puesto que tanto *a* como *a[i]* son objetos *Array*.

También aparece en el código de la figura 2.4 la facilidad de realizar la asignación de un arreglo completo a una variable tipo Array. Por ejemplo, el arreglo *uno* es un conjunto de 20 unos. Este se lo asigna directamente al arreglo a[0] de la siguiente manera: a[0]=uno.

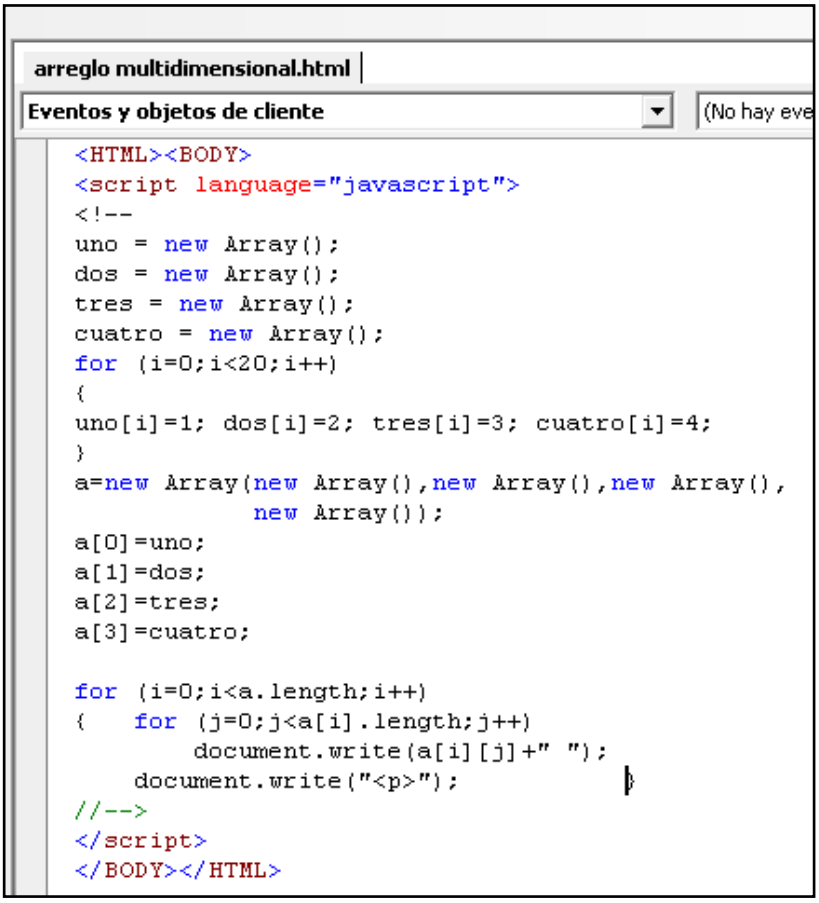


FIGURA 2.4 CÓDIGO JSCRIPT PARA LA CREACIÓN DE UN ARREGLO MULTIDIMENSIONAL

El resultado visual del código anterior es un arreglo rectangular de 4 x 20 tal como lo muestra la figura 2.5.

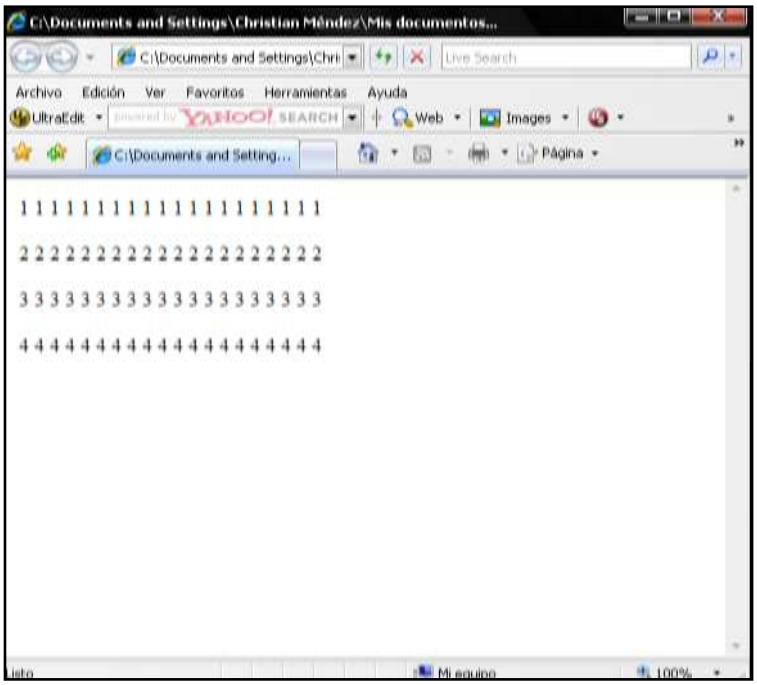


FIGURA 2.5 PRESENTACIÓN DE UNA MATRIZ DE 4 X 20

Uso de funciones:

Las funciones pueden ser declaradas en cualquier parte de un código JavaScript con la condición de que su invocación sea posterior a su declaración. También se puede invocar a estas funciones desde código HTML, al ocurrir un evento provocado por el usuario.

Los elementos de HTML contienen la posibilidad de actuar según un evento ocurra. Por ejemplo, el elemento <IMG> contiene la

posibilidad de ejecutar una acción al pasar el ratón sobre su imagen. Es común que una página web actúe de alguna manera cuando se da click a uno de sus botones. Un botón con propiedad *name* igual a *btnCalcular* sería el siguiente:

- <input type='button' value='Calcular' name='btnCalcular' onclick='Metodo()'>

En las líneas anteriores se muestra que el botón “btnCalcular” ejecutará la función “Metodo()” cuando se de click sobre él (evento onclick).

El siguiente código muestra el texto del botón btnTexto (su propiedad *value*) cuando se da click sobre él:

```
<HTML><BODY>
<input type='button' value='Calcular' name='btnTexto'
onclick='Mostrar()' ID=Button1>
<script language="javascript">
 <!--
function Mostrar()
{  alert(Button1.value);}
//-->
</script> </BODY>
</HTML>
```

La figura 2.6 muestra el resultado dar click sobre el botón btnTexto.

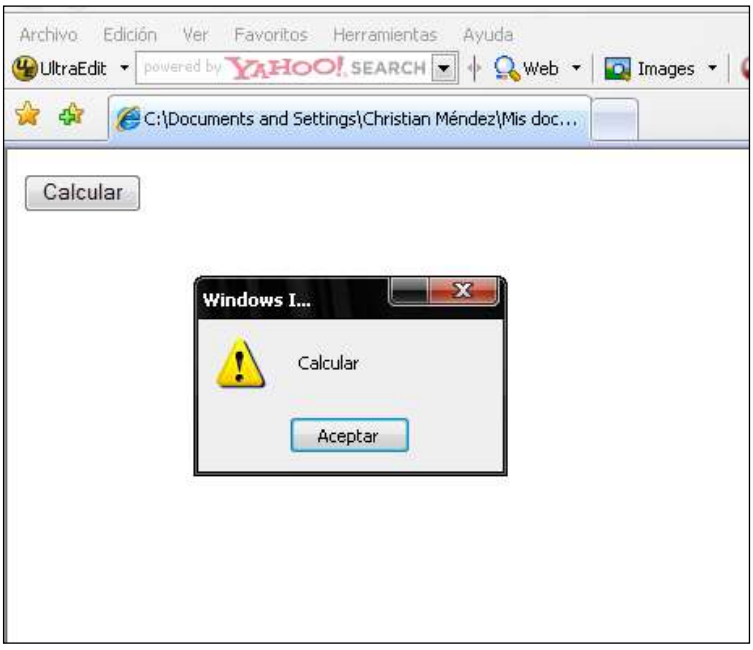


FIGURA 2.6 PRESENTACIÓN DE LA PROPIEDAD *VALUE* DE UN BOTÓN.

Uso de librerías:

En ocasiones los programadores hacen uso de clases junto con sus métodos y propiedades que ellos u otros han hecho anteriormente. De esta manera ellos pueden simplificar su trabajo. Las librerías son código JavaScript con extensión *.js. Las librerías *.js pueden ser llamadas utilizando la siguiente instrucción:

<script type="text/javascript" src="nombre_de_libreria.js"></script>

La página que se mostrará en el capítulo III utiliza 3 librerías: una para el manejo de matrices, las otras para la graficación de diferentes figuras y para el movimiento de imágenes respectivamente.

El objeto Math:

Puesto que la página que se ha diseñado en el presente trabajo aplica métodos de discriminación, el código para esta página requiere efectuar cálculos matemáticos, a este respecto resulta útil el objeto Math. Los métodos que son parte del objeto Math de JScript son los siguientes: *abs, acos, asin, atan, atan2, ceil, cos, exp, floor, log, max, min, pow, random, round, sin, sqrt, tan*. Algunos de estos métodos son utilizados en el desarrollo del programa.

El método random() es útil en este trabajo para generar conjuntos de datos parcialmente aleatorios. Este método devuelve un número pseudoaleatorio entre 0 y 1. Este método puede ser utilizado para generar un número aleatorio entre dos números cualesquiera a y b. La siguiente función llamada "aleatorio" genera un número pseudoaleatorio entre dos parámetros:

```
function aleatorio(a,b)
 {   return  Math.random()*(b-a)  + a;  }
```

El código que se muestra a continuación utiliza la función aleatorio para presentar 10 números pseudoaleatorios entre 4 y 9.

```
<HTML>
<BODY>
<script language="javascript">
<!--
function aleatorio_entero(a,b)
{    return Math.round( Math.random()*(b-a) + a); }
function aleatorio(a,b)
{    return Math.random()*(b-a) + a; }
for (i=1;i<=10;i++)
document.write("<p>"+aleatorio(4,9))
//-->
</script>
</BODY></HTML>
```

El resultado del código anterior se muestra en la figura 2.7. Se observa en esta figura que los números que se presentan no son enteros, para que éstos sean enteros se puede utilizar otro método del objeto Math llamado round() que redondea una cifra haciéndola entera. En el código anterior también se incluye una función llamada aleatorio_entero(a,b) que utiliza el método round() para

producir números enteros y ésta se la puede incluir en la instrucción document.write() si se desea números enteros.

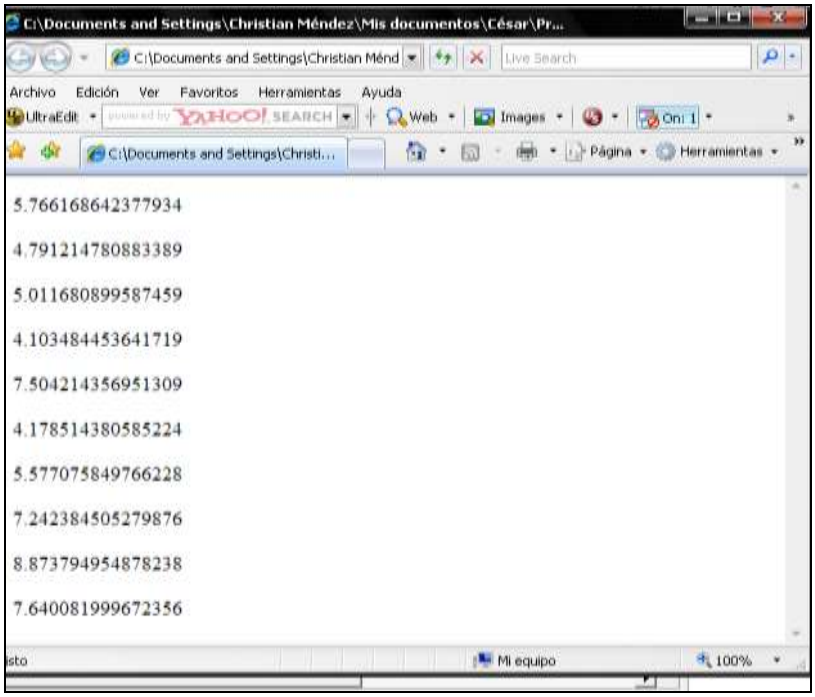


FIGURA 2.7 PRESENTACIÓN DE NÚMEROS UTILIZANDO EL MÉTODO RANDOM()

2.3 Librería JavaScript Vector Graphics: wz_jsgraphics.js

Esta es una librería desarrollada por Walter Zorn que contiene métodos para graficar una serie de figuras geométricas tales como: rectángulos, elipses, líneas, polígonos, arcos. Aunque los detalles para el uso de esta librería podrían encontrarse en el sitio web

www.walterzorn.com aquí se exponen algunos aspectos acerca de cómo usar esta librería en lo relacionado a la página web para el aprendizaje de análisis discriminante.

El primer paso es añadir la librería para que pueda ser utilizada. Esto se lo hace con la siguiente instrucción la cual ya se mencionó anteriormente:

<script type="text/javascript" src="wz_jsgraphics.js"></script>

Luego de esto es necesario definir un objeto de la clase jsGraphics() como si se definiera cualquier otro objeto:

var a = new jsGraphics();

Se puede definir un objeto jsGraphics() de dos maneras:

✓ Si el constructor no incluye argumento, el área de graficación será todo el documento: var a = new jsGraphics();

✓ Si el constructor incluye argumento, éste debe ser una capa previamente definida la cual será el área de graficación: var a=new jsGraphics(“miCapa”);

Será útil entonces definir una capa para que ésta represente el plano bidimensional donde se graficarán los puntos del conjunto de datos a discriminar.

Uno de los métodos que ofrece esta librería gráfica es fillEllipse(X,Y,ancho,alto). Esta instrucción sirve para graficar una elipse pintada por dentro, la cual se encuentra inscrita en el rectángulo de ancho y alto especificados y cuyo vértice superior izquierdo está en la columna X y la fila Y. La figura 2.8 muestra el esquema de graficación de una elipse. El rectángulo de la figura se muestra en líneas punteadas porque no se grafica sino que sólo sirve de referencia para graficar la elipse.

El código para graficar una elipse en la capa1 sería el siguiente:

```
var  plano=new jsGraphics(“capa1”);
plano.fillEllipse(0,0,20,20);
plano.paint();
```

De esta manera pudieran graficarse puntos (elipses muy pequeñas) en el plano bidimensional utilizando las coordenadas x,y de esta manera: plano.fillEllipse(x,y,3,3); Puesto que son elipses muy pequeñas (más bien círculos), en la instrucción anterior se nota que el ancho y alto de éstas es apenas 3.

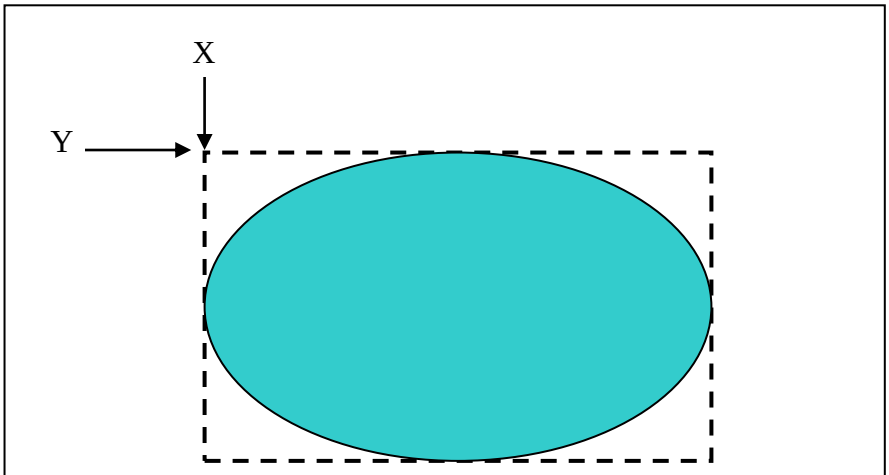


FIGURA 2.8 ESQUEMA DE GRAFICACIÓN DE UNA ELIPSE USANDO LA LIBRERÍA WZ_JSGRAPHICS.JS

La librería gráfica también permite ubicar una imagen cualquiera en la posición (x,y) como si se tratase de una elipse. Como se verá más adelante esto es muy útil para que los puntos graficados puedan ser desplazados por el usuario. La instrucción para realizar esto es la siguiente:

drawImage("src", X, Y, ancho, alto); donde src es el nombre de la imagen, por ejemplo “balloon.jpg”; y (X,Y) la posición donde se ubica el extremo superior izquierdo de la imagen.

Utilizando el método drawImage para una capa determinada y seleccionando las imágenes adecuadas se logran visualizar en el plano imágenes como si éstas fueran puntos, sin necesidad de graficar elipses. En la figura 2.9 se muestra la imagen  repetidas veces como puntos distribuidos en el plano.

El código para mostrar un punto parecido a los de la figura 2.9 es:

- var plano = new jsGraphics("myCanvas");
- plano.drawImage(“BD14868_.GIF”,30,50,10,10);
- plano.paint();

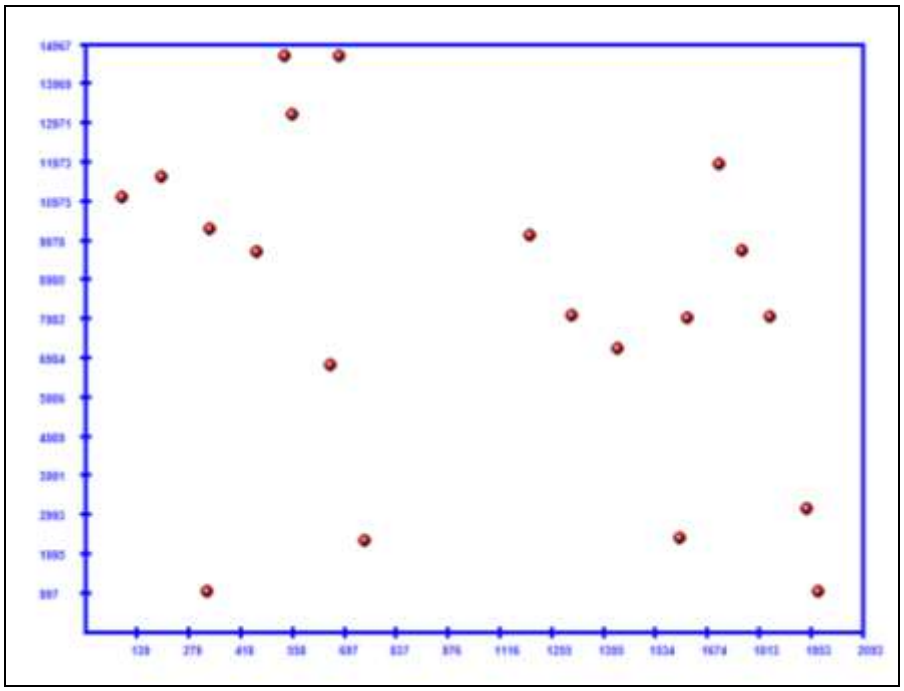


FIGURA 2.9 VISUALIZACIÓN DE IMÁGENES COMO PUNTOS EN EL PLANO UTILIZANDO EL MÉTODO DRAWIMAGE() DE LA LIBRERÍA WZ_JSGRAPHICS.JS

2.4 Librería JavaScript Drag & Drop: wz_dragdrop.js

Esta librería también ha sido desarrollada por Walter Zorn y se la utiliza para una de las partes esenciales de este proyecto. Aquí

sólo se presentará como hacer para que una imagen cualquiera se haga movible. Mayores detalles de cómo utilizar esta librería podrían encontrarse en el sitio www.walterzorn.com.

Con el uso la librería wz_dragdrop.js es posible desarrollar páginas web en donde el usuario puede desplazar tanto imágenes como capas (layers).

El procedimiento para hacer que una imagen se pueda desplazar libremente por el usuario se resume de la siguiente manera:

1) Se incluye la librería apropiada

2) Se crea las imágenes que se desee utilizando HTML común, por ejemplo:

<IMG src=”balloon.gif” name=”balon”>

< IMG src=”face.gif” name=”cara”>

3) En las últimas líneas antes de </BODY> se debe incluir la instrucción: SET_DHTML con la propiedad “name” de las imágenes que se desea sean movibles, para hacer movibles las imágenes creadas anteriormente la instrucción sería: SET_DHTML("balon", "cara");

Si solamente se deseara que la imagen “balon” sea movible, se deberá incluir solamente el parámetro “balon” en la instrucción SET_DHTML.

Cuando se explicó la librería wz_graphics.js se mencionó que para simular la graficación de puntos se podría utilizar el método fillEllipse o el método drawImage. La ventaja de utilizar drawImage es que los puntos pueden hacerse movibles, pues éstos en realidad son pequeñas imágenes. De esta forma podríamos tener n puntos en el plano xy que pudieran ser desplazados por el usuario.

CAPITULO III

3. DISEÑO DEL APLICATIVO WEB

La página web que se ha desarrollado para el aprendizaje de análisis discriminante permite probar 5 métodos que se expusieron en el capítulo I.

Aunque el análisis discriminante puede trabajar con elementos determinados por n variables ya sean categóricas o numéricas, esta herramienta se restringe a trabajar con elementos determinados por 2

variables numéricas. Esto se debe a que ésta es una herramienta gráfica que permite visualizar cada elemento según su clase en 2 dimensiones. Puesto que cada elemento tiene 2 variables o atributos pueden ser representados como puntos (pares ordenados) en un plano bidimensional.

Aunque en un mismo plano bidimensional pueden ser visualizados elementos de más de 2 clases, esta página sólo trabaja con datos provenientes de 2 clases posibles.

Al iniciar este aplicativo web se presenta un plano bidimensional con un conjunto de puntos de 2 clases junto con una sección donde el usuario puede manipular las distintas opciones que se ofrecen. La figura 3.1 presenta esta pantalla inicial.

La primera opción que ofrece la página es el método que se desea aplicar, la figura 3.2 presenta este detalle, donde la opción “Regresión” se refiere al método de los mínimos cuadrados ordinarios. A continuación se expondrá la manera de utilizar las otras opciones presentes en la página.

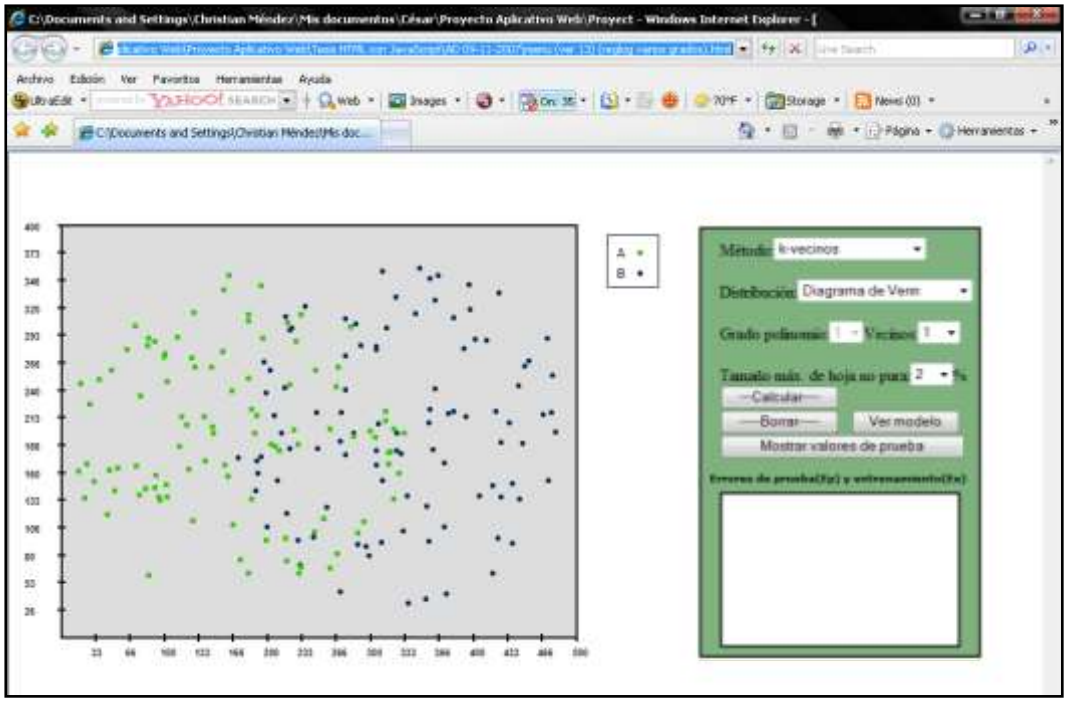


FIGURA 3.1 PANTALLA INICIAL DEL APLICATIVO WEB

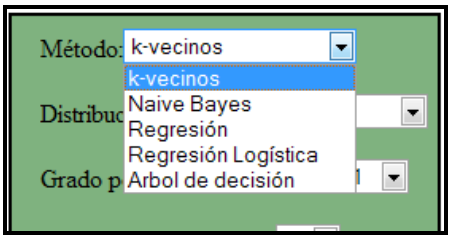


FIGURA 3.2 SELECCIÓN DEL MÉTODO

3.1 Distribución del conjunto de datos

En el plano bidimensional se presentan un conjunto de puntos distribuidos de una manera específica. Existen 3 distribuciones que el usuario puede seleccionar (figura 3.3):

- Diagrama de Venn
- Cuadrados traslapados
- Exponencial

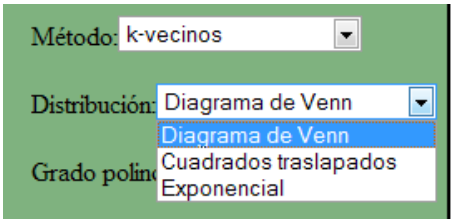


FIGURA 3.3 OPCIONES DE DISTRIBUCIÓN DE PUNTOS

Cuando se inicia la página la distribución por defecto es la diagrama de Venn como su nombre lo indica esta opción distribuye los puntos como un diagrama de Venn. La figura 3.4 muestra el esquema de esta distribución.

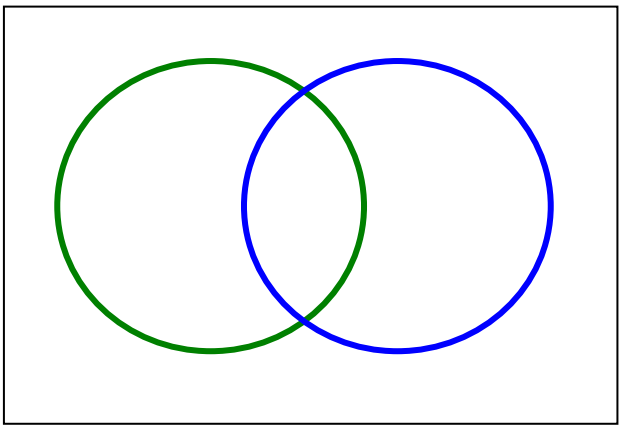


FIGURA 3.4 DISTRIBUCIÓN DE DIAGRAMA DE VENN

En cada círculo del esquema anterior se generarán puntos de una clase determinada, es decir, en el círculo verde caerán puntos o elementos (x,y) de la clase A y en el otro círculo los puntos de la clase B. Sin embargo dentro de cada círculo los puntos estarán distribuidos al azar. Como se observa en la figura 3.4 esta distribución de puntos sí tiene solapamiento, la intersección de los círculos es el área en donde hay puntos de ambas clases. La figura 3.1 muestra como luce este tipo de distribución. También en la figura 3.5 se ha sobrepuesto los círculos del diagrama a un conjunto de puntos generados con esta distribución.

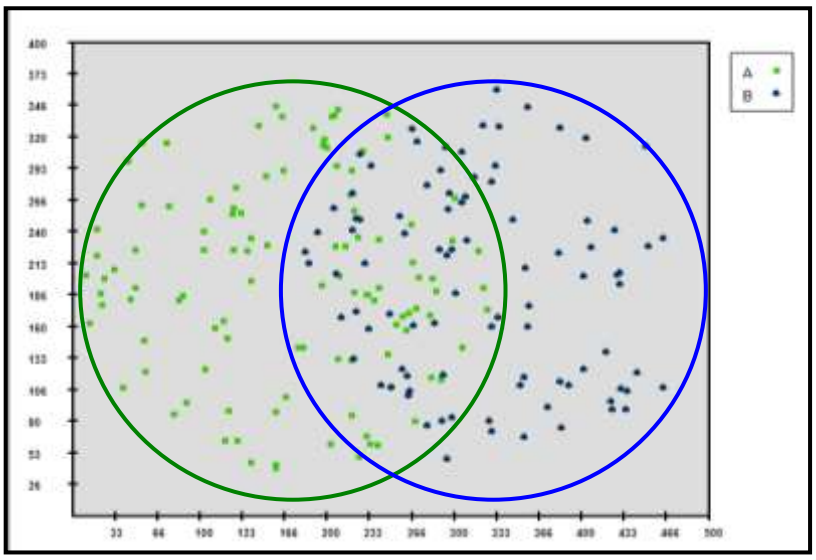


FIGURA 3.5 CONJUNTO DE PUNTOS DISTRIBUIDOS COMO DIAGRAMA DE VENN

Otra de las distribuciones llamada “cuadrados traslapados” genera cada tipo de puntos en un cuadrado (o rectángulo) correspondiente, estos rectángulos se superponen en una de sus esquinas. Un conjunto de puntos generados según la distribución “cuadrados traslapados” se muestra en la figura 3.6. Los cuadrados que están sobrepuestos en la figura permiten visualizar el esquema de la distribución.

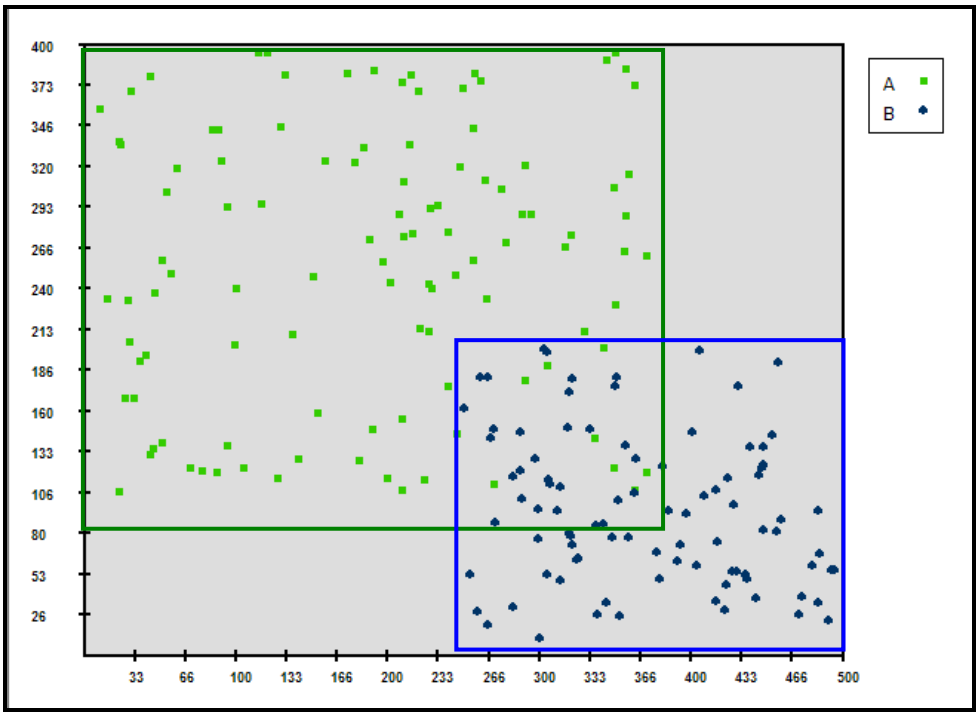


FIGURA 3.6 DISTRIBUCION DE CUADRADOS TRASLAPADOS

La última distribución que se ofrece se denomina “exponencial” porque tanto las variables X como Y de cada clase tienen función de densidad exponencial univariada $f(x)=(1/k)e^{-(x-L)/k}$ para $x \geq L$. La figura 3.7 muestra esta distribución aleatoria.

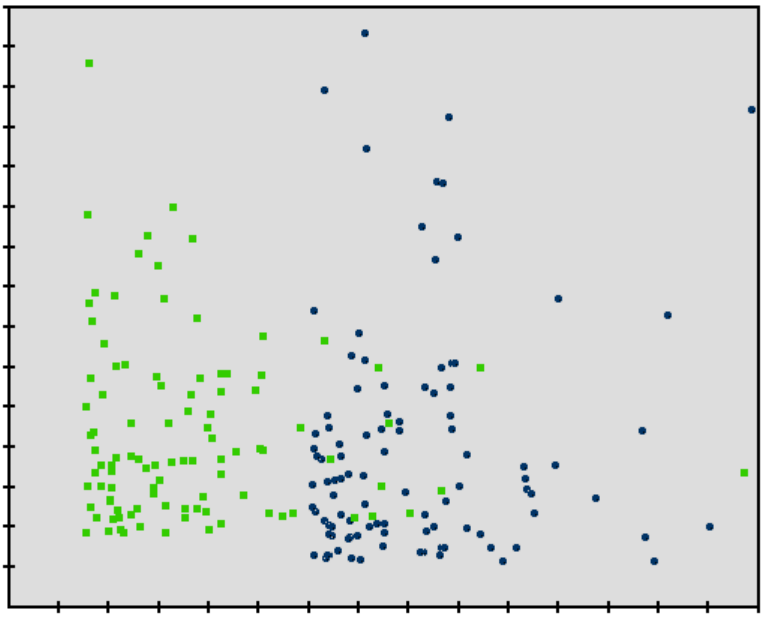


FIGURA 3.7 DISTRIBUCIÓN DE PUNTOS TIPO “EXPONENCIAL”

En esta página web el usuario tiene la posibilidad de desplazar cada punto generado para formar su propio conjunto de puntos y

experimentar cualquier método. La figura 3.8 muestra uno de estos conjuntos personalizados.

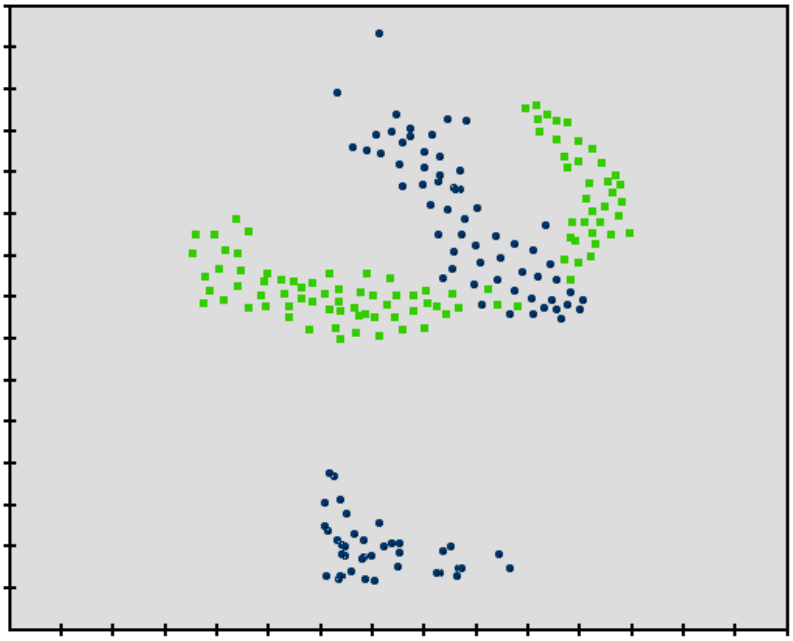


FIGURA 3.8 CONJUNTO DE DATOS PERSONALIZADO

3.2 Selección de parámetros

Cada método a excepción de Bayes Naive ofrece un parámetro el cual el usuario puede modificar antes de graficar.

El método de los kvecinos permite escoger la cantidad de vecinos, la figura 3.9 muestra esto. A su vez los métodos de regresión lineal y logística permiten escoger el grado del polinomio que utilizarán en su modelo. Por último, para el método del árbol de decisión se permite escoger el porcentaje de elementos al cual no puede llegar un nodo para que éste pueda ser particionado, a esta opción se le ha llamado "Tamaño máximo de hoja no pura", puesto que si una hoja no pura tuviese mayor número de elementos que el porcentaje fijado tuviese que ser particionada con lo cual dejaría de ser hoja. La figura 3.10 muestra la selección del parámetro de un árbol de decisión.

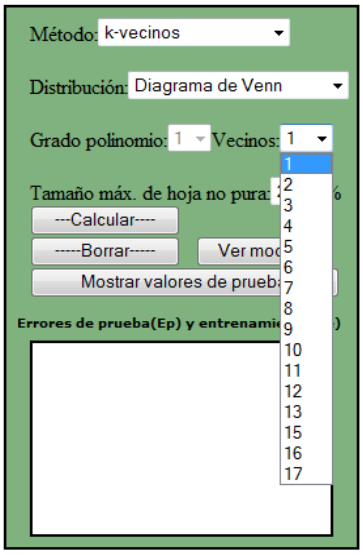


FIGURA 3.9 SELECCIÓN DEL NÚMERO DE VECINOS

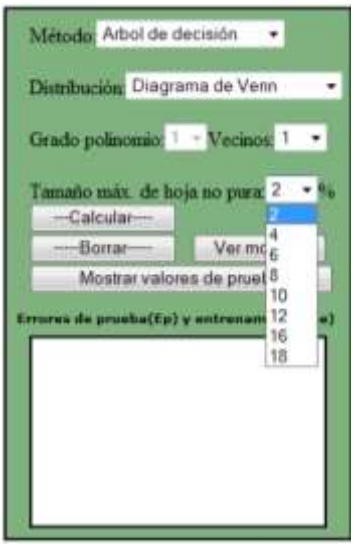


FIGURA 3.10 SELECCIÓN DEL PARÁMETRO DE PARADA DE UN ÁRBOL DE DECISIÓN

3.3 Botones de opción

El programa ofrece las opciones que se muestran en la figura 3.11. Después que el usuario ha seleccionado la distribución de puntos, el método a aplicar y el valor de su parámetro correspondiente

puede presionar “Calcular” para que el método se ejecute y comience a graficarse la curva discriminante en el plano.

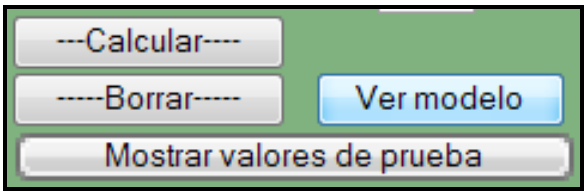


FIGURA 3.11 BOTONES DE OPCIÓN

Cada vez que el usuario presiona el botón “Calcular” se ejecuta el método que esté seleccionado en ese momento. Tal como lo muestra la figura 3.12 este aplicativo permite graficar varias curvas para discriminar el mismo conjunto de datos en el mismo plano diferenciándose éstas por su color.

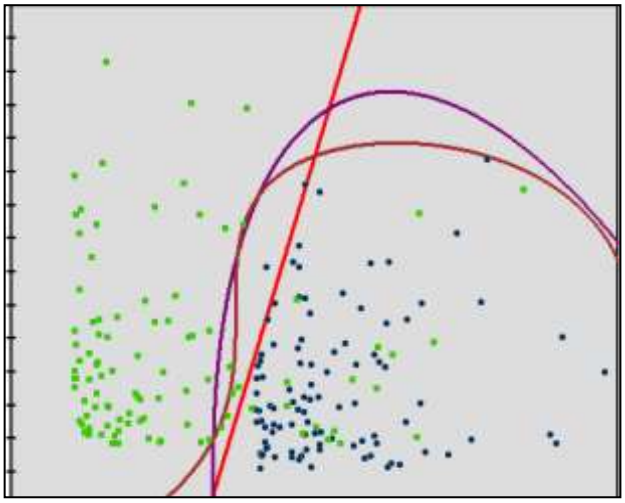


FIGURA 3.12 GRAFICACIÓN DE VARIAS CURVAS DISCRIMINANTES EN EL MISMO PLANO

Si el usuario ya no deseara ver los gráficos de los métodos anteriores bastará con presionar el botón “Borrar” que no elimina los puntos sino los trazos de cada método. Luego de esto podrá si desea seguir graficando curvas discriminantes con el mismo conjunto de datos.

Si el usuario desease probar otro conjunto de datos podrá hacerlo escogiendo en la lista de opciones de distribución de la cual se mencionó anteriormente. Si existiesen curvas discriminantes graficadas en el plano al escoger una nueva distribución éstas automáticamente se borrarán y se generará un nuevo conjunto de datos según la distribución seleccionada.

Los puntos que se visualizan en el plano son el conjunto de entrenamiento, en base a éstos se realiza la tarea de discriminación. El usuario podrá ver el conjunto de prueba en el cual se basa el error de prueba presionando el botón “Mostrar valores de prueba”. Las figuras 3.13 y 3.14 presentan la misma situación sin y con datos de prueba visualizados respectivamente.

Se nota en la figura 3.14 que los datos de prueba podrán ocultarse presionando el mismo botón.

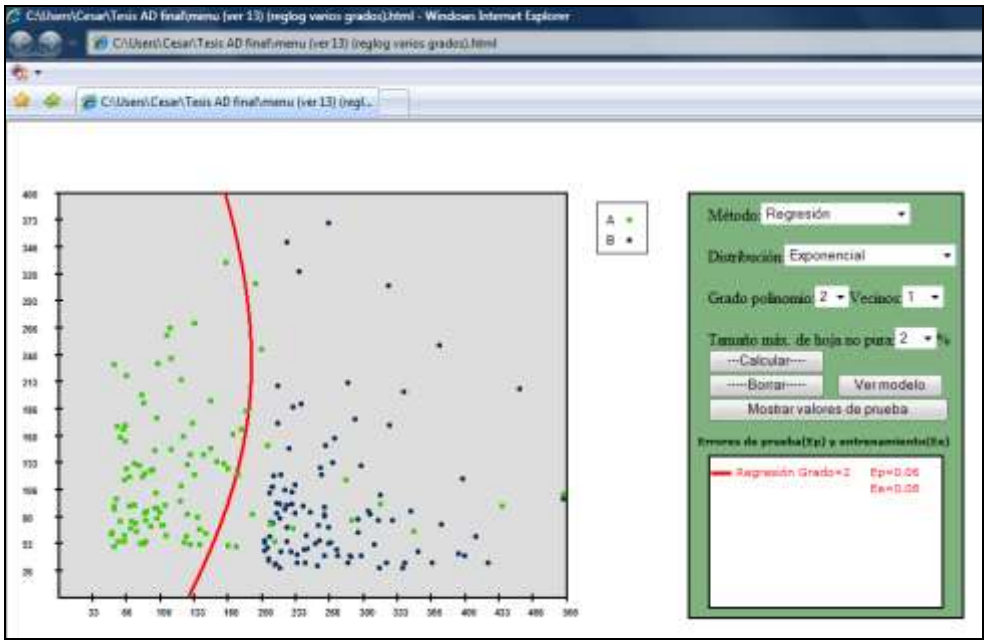


FIGURA 3.13 DISCRIMINACION SIN VISUALIZAR DATOS DE PRUEBA

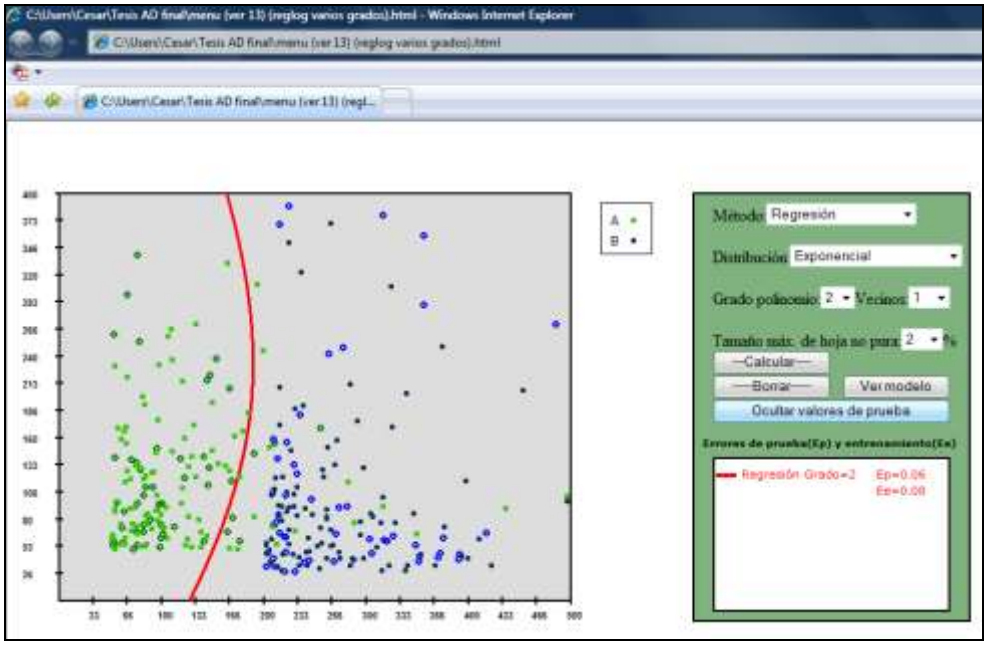


FIGURA 3.14 DISCRIMINACION VISUALIZANDO DATOS DE PRUEBA

Finalmente, el botón “Ver modelo” permitirá ver el modelo de aquellos métodos que son comprensibles, estos son: regresión lineal, regresión logística y árbol de decisión. La figura 3.15 muestra el modelo de la regresión de segundo grado aplicada en la figura 3.13.

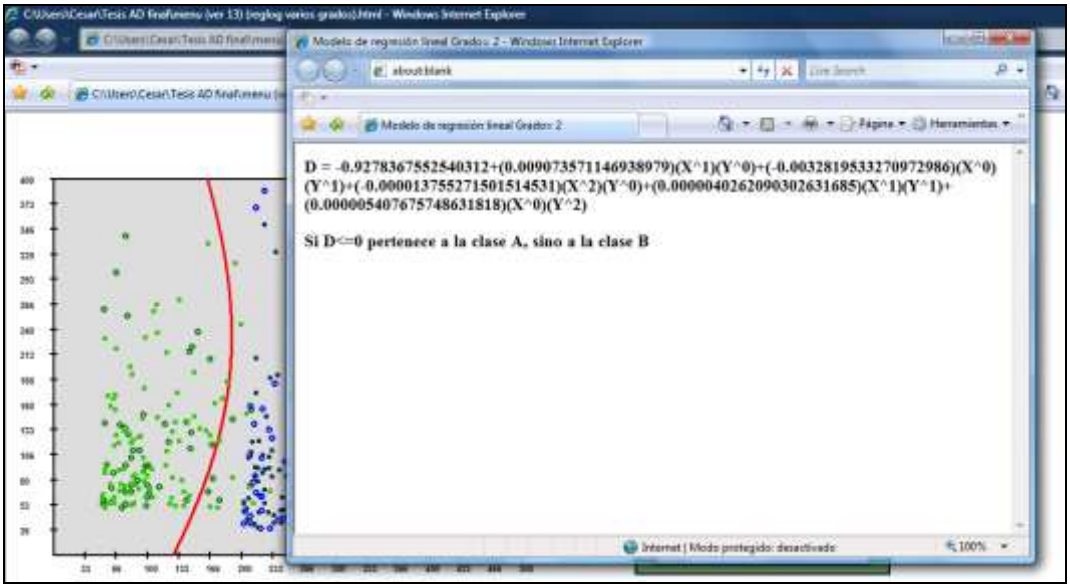


FIGURA 3.15 VISUALIZACIÓN DEL MODELO DE UN MÉTODO

3.4 Tablero de resultados

En el tablero de resultados es donde se muestran las respuestas analíticas de cada método, específicamente el error de prueba (Ep) y el error de entrenamiento (Ee). Este tablero puede contener más de 1 método con sus respuestas aplicado al mismo conjunto de datos, estos resultados permanecerán registrados en el tablero hasta que el usuario borre las curvas discriminantes del plano. La figura 3.16 muestra las gráficas de diferentes curvas con el mismo conjunto de datos y el registro de sus errores de prueba y

entrenamiento en el tablero de resultados (a la derecha de la figura).

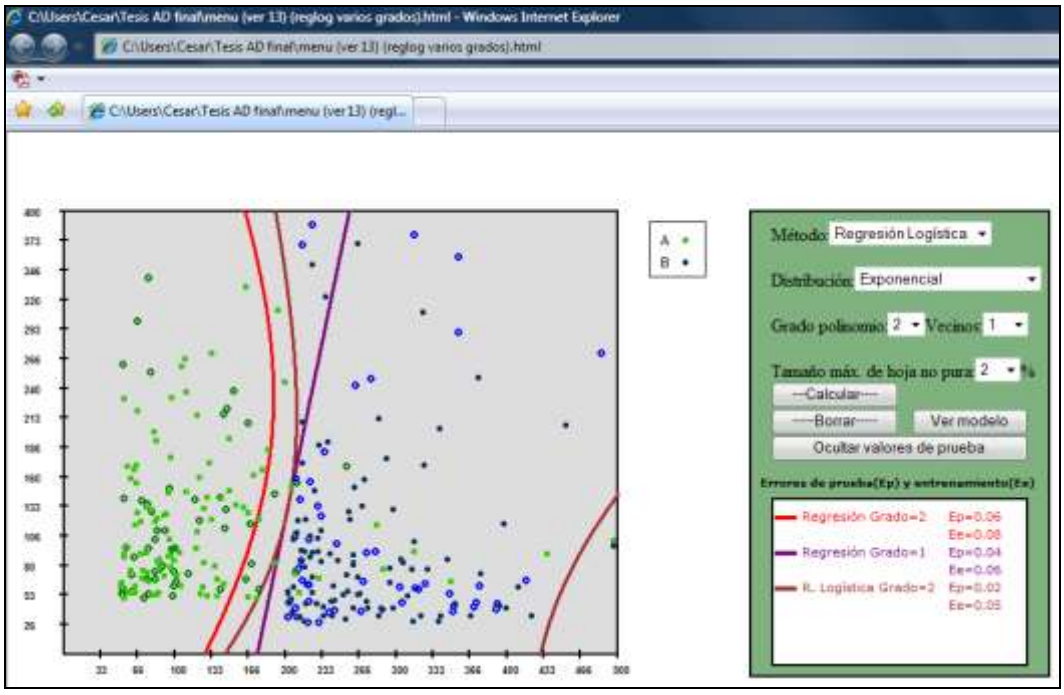


FIGURA 3.16 REGISTRO DE RESPUESTAS EN EL TABLERO DE RESULTADOS

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.

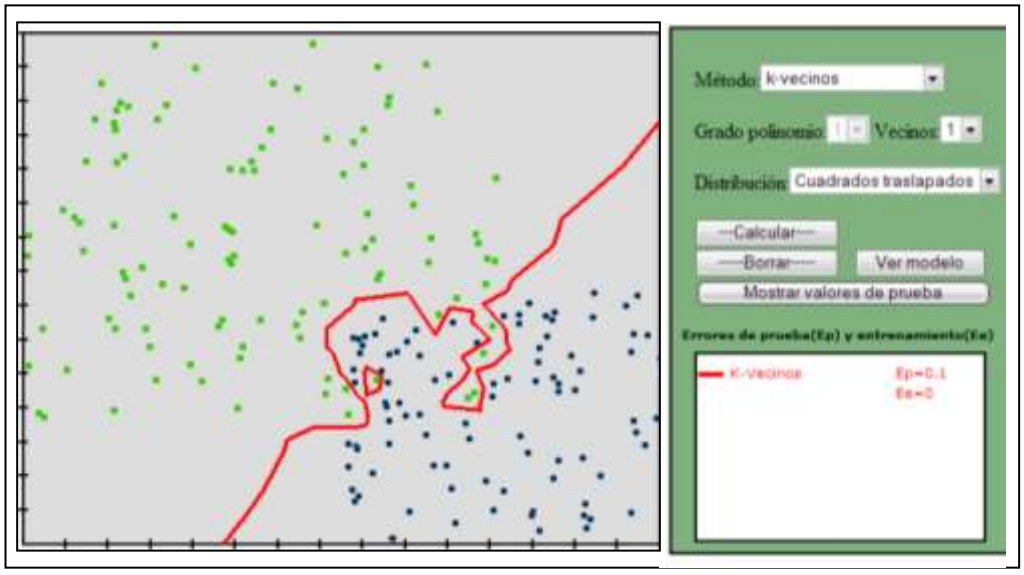


FIGURA 4.1 MUESTRA DE LA EXPRESIVIDAD

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.

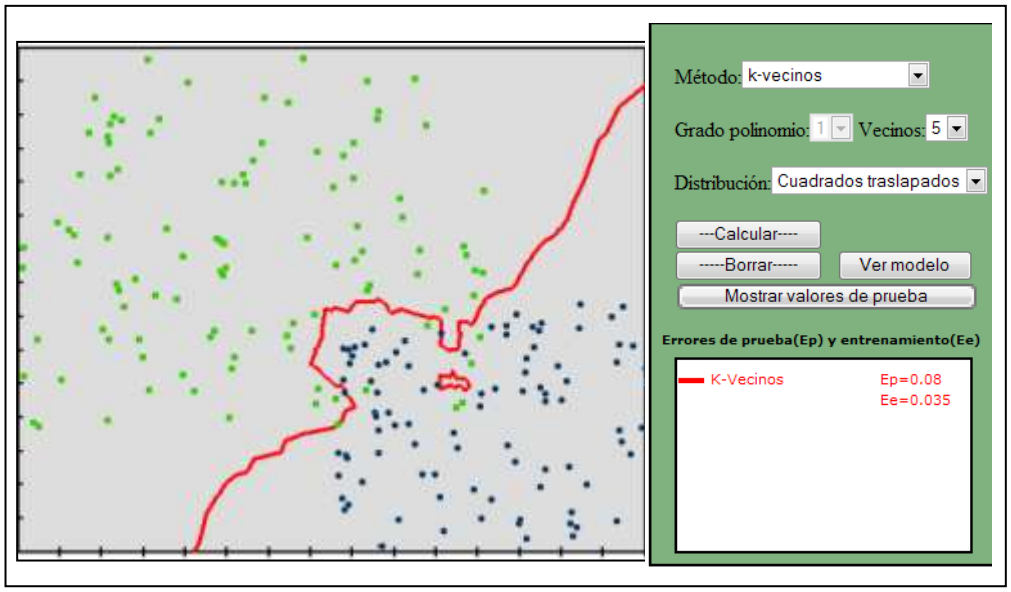


FIGURA 4.2 MUESTRA DE LA EXPRESIVIDAD DEL MÉTODO 5-VECINOS MÁS CERCANOS

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.

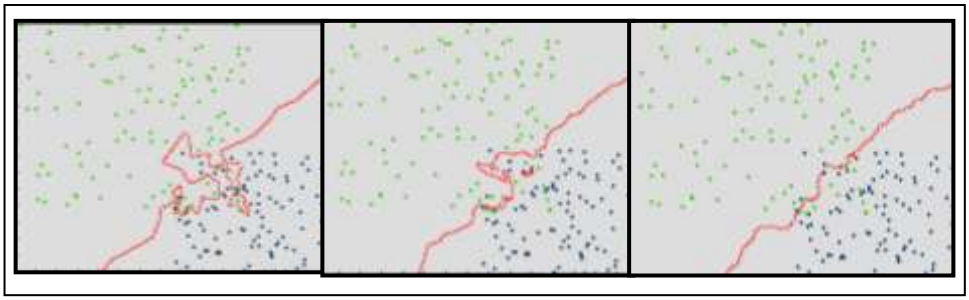


FIGURA 4.3 CAMBIOS EN LA EXPRESIVIDAD DE LOS K-VECINOS PARA K=1, 7 Y 15

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.

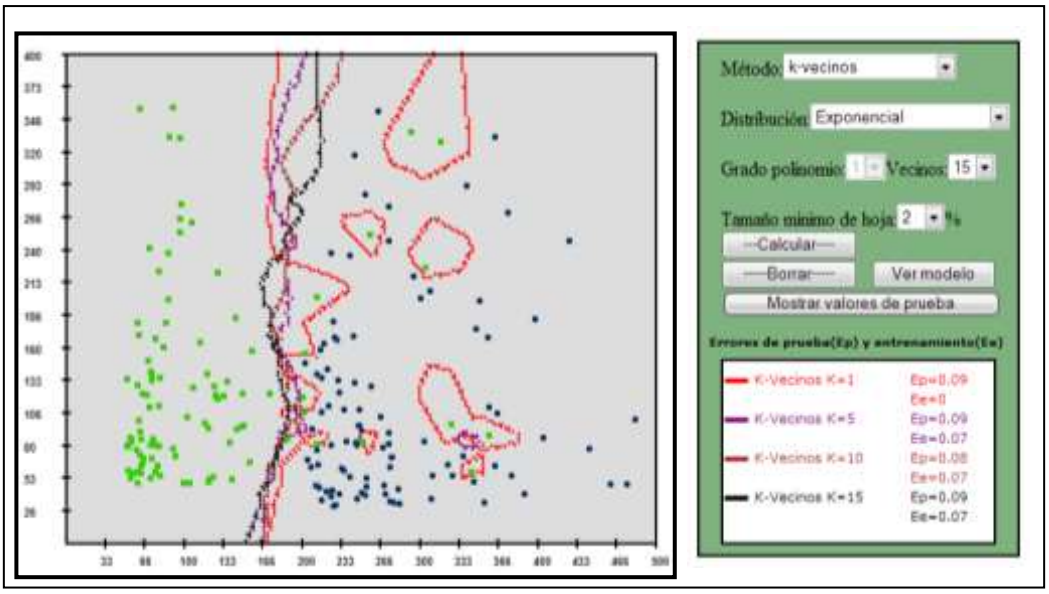


FIGURA 4.4 MÉTODO DE LOS K-VECINOS PARA K=1,5,10 Y 15 PARA CLASES DISTRIBUIDAS EXPONENCIALMENTE.

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.

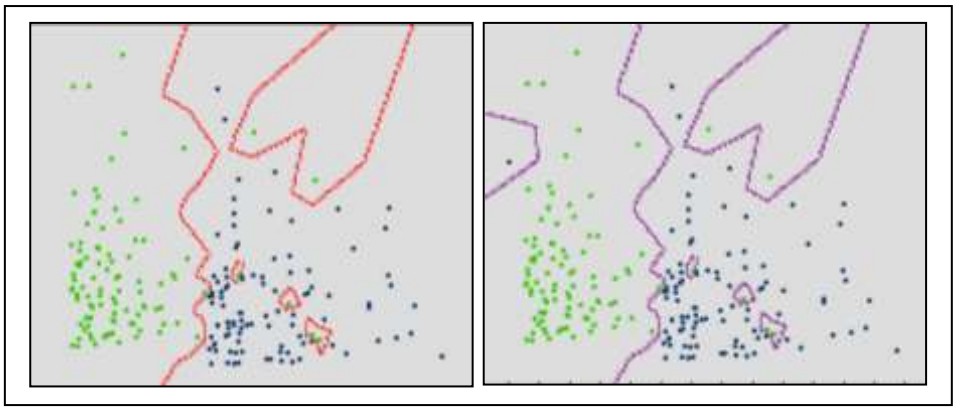


FIGURA 4.5 EJEMPLO DE LA SENSIBILIDAD DE 1-VECINO MÁS CERCANO A UN DATO ANÓMALA

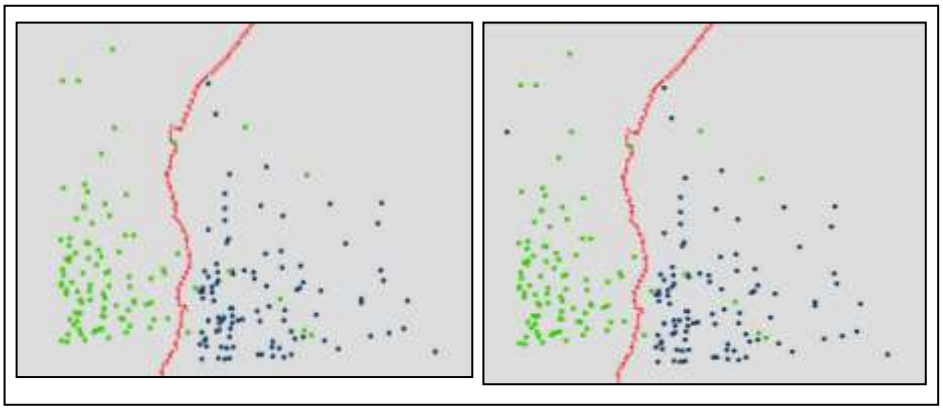


FIGURA 4.6 EJEMPLO DE LA SENSIBILIDAD DE 5-VECINOS MÁS CERCANOS A UN DATO ANÓMALA

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.

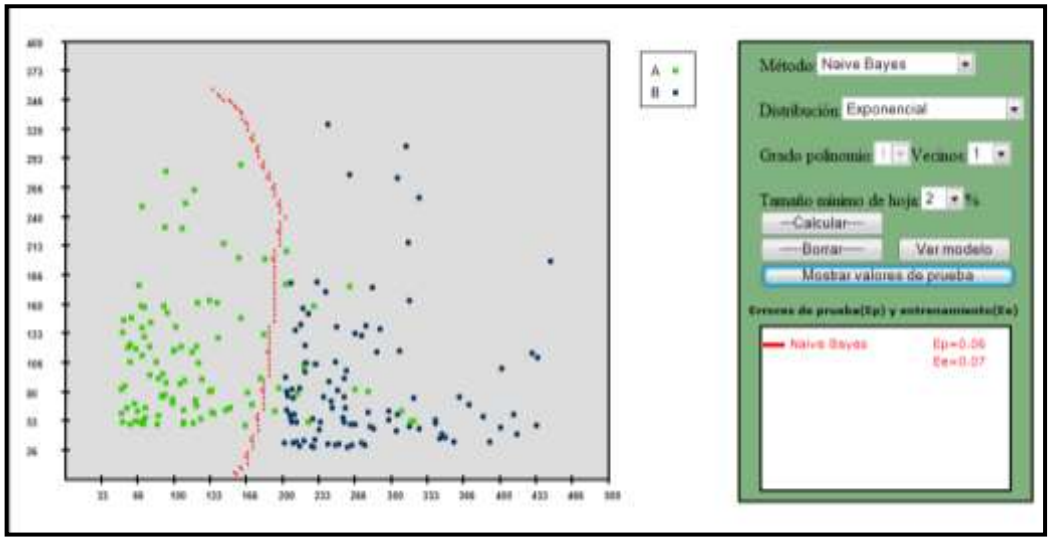


FIGURA 4.7 NAIVE BAYES APLICADO A MUESTRAS DE POBLACIONES EXPONENCIALES.

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.

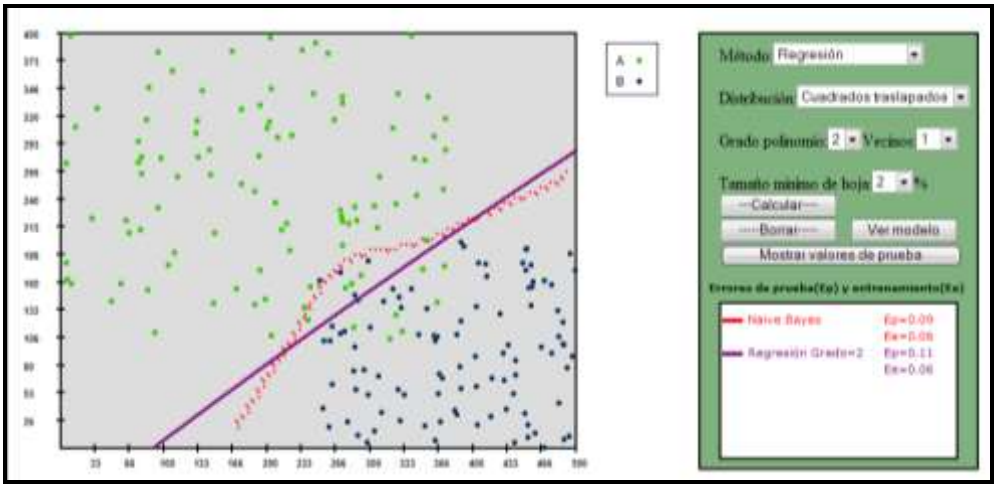


FIGURA 4.8 COMPARACION ENTRE EL MÉTODO DE NAIVE BAYES Y REGRESIÓN PARA DISTRIBUCION DE CUADRADOS TRASLAPADOS

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.

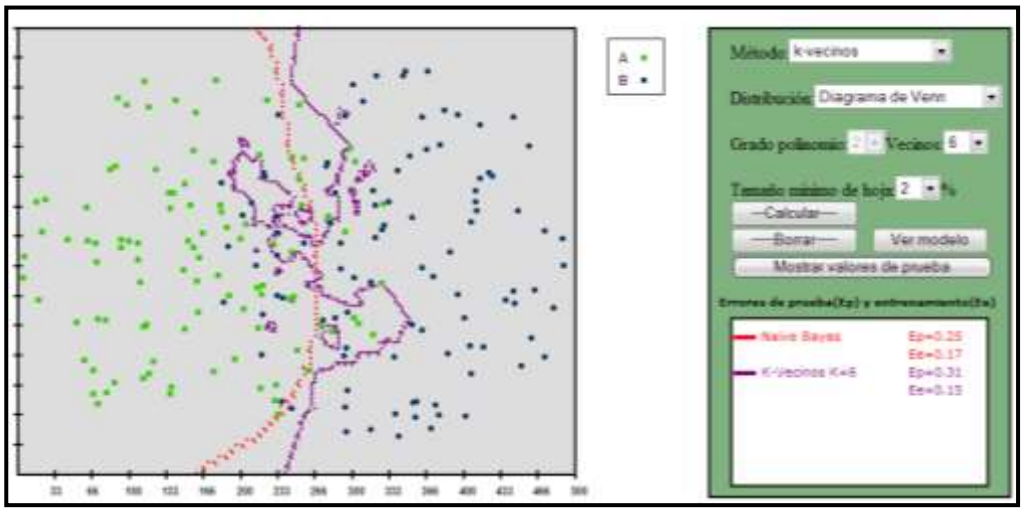


FIGURA 4.9 COMPARACIÓN ENTRE 6-VECINOS Y NAIVE BAYES

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.

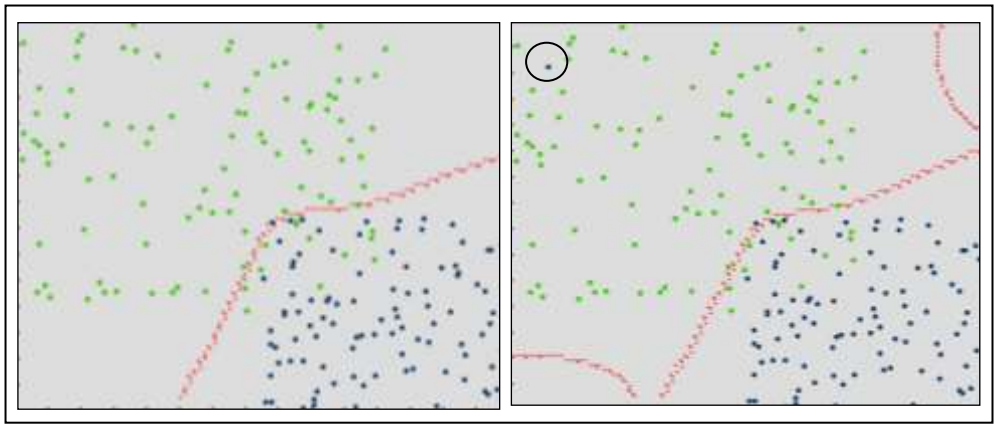


FIGURA 4.10 SENSIBILIDAD DE NAIVE BAYES KERNEL ANTE UN DATO ANÓMALA, DISTRIBUCIÓN DE CUADRADOS TRASLAPADOS

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.

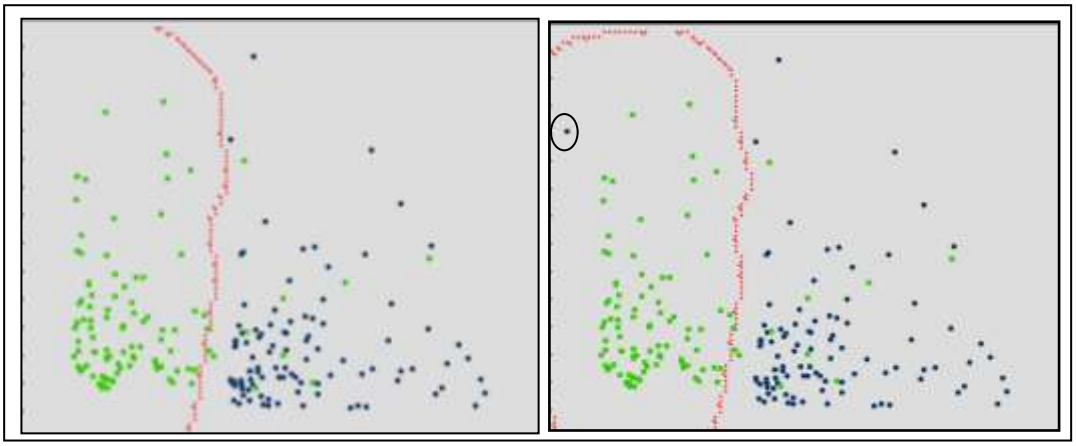


FIGURA 4.11 SENSIBILIDAD DE NAIVE BAYES KERNEL ANTE UN DATO ANÓMALA, DISTRIBUCIÓN EXPONENCIAL

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.

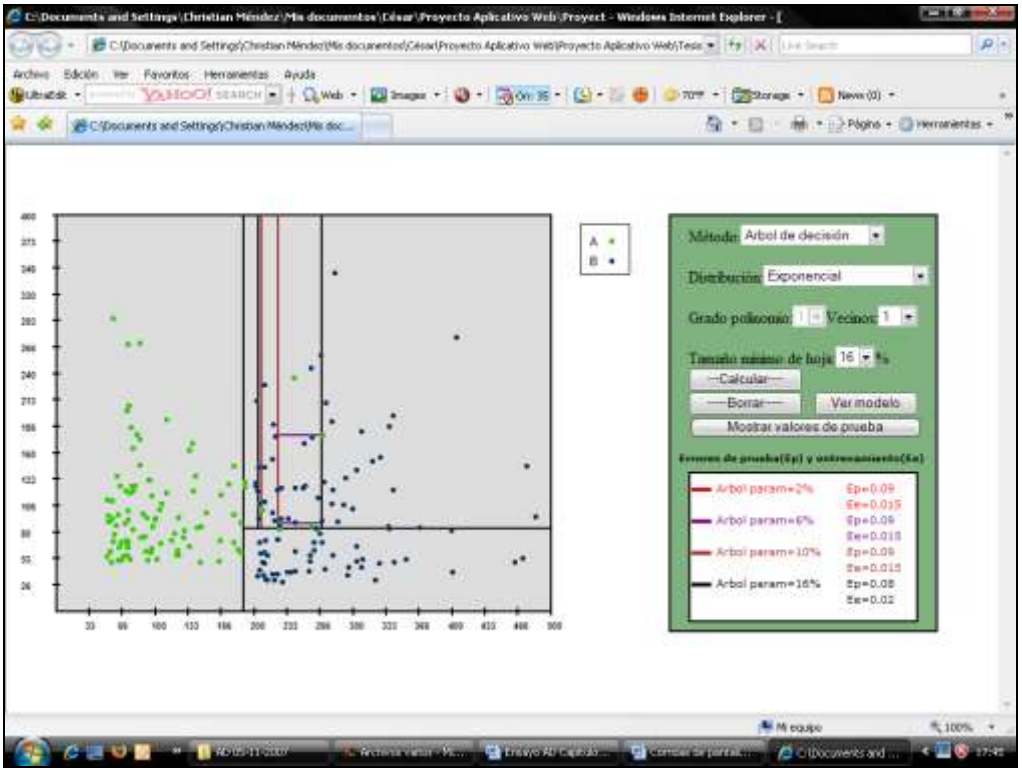


FIGURA 4.12 APLICACIÓN DEL ÁRBOL DE DECISIÓN VARIANDO S PARA LA MISMA MUESTRA

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.

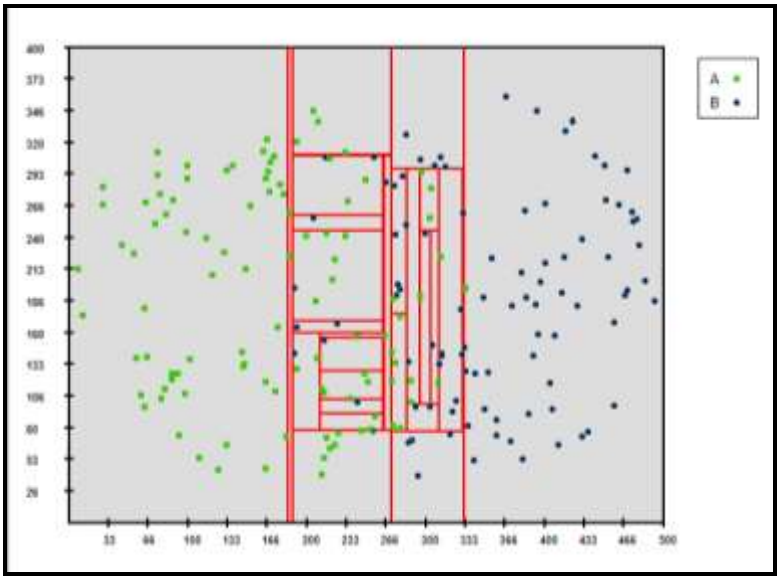


FIGURA 4.13 APLICACIÓN DE ÁRBOL DE DECISIÓN A LA DISTRIBUCIÓN DE DIAGRAMA DE VENN

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.

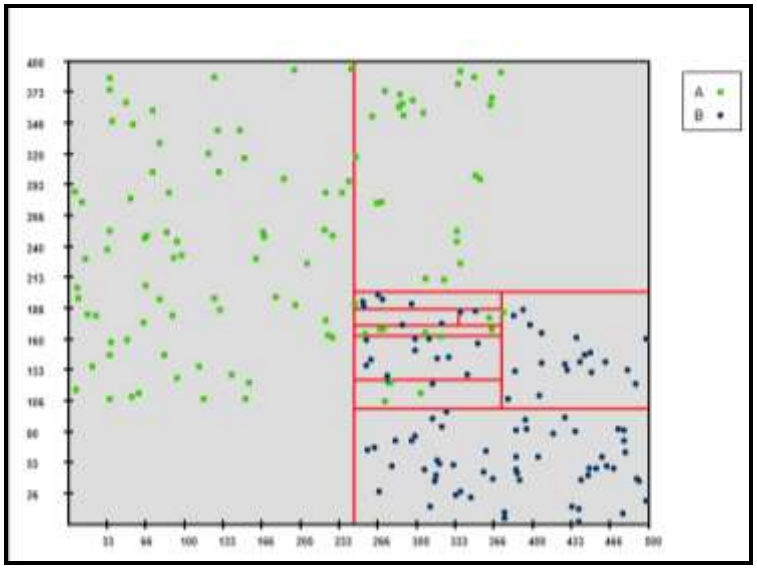
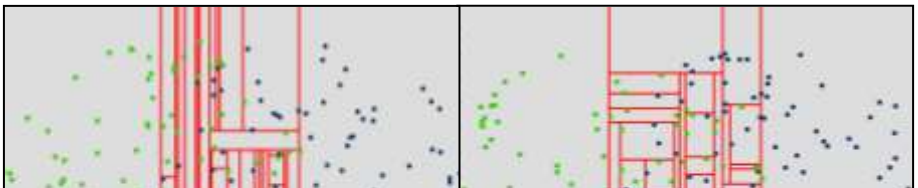


FIGURA 4.14 APLICACIÓN DE ÁRBOL DE DECISIÓN A LA DISTRIBUCIÓN DE CUADRADOS TRASLAPADOS

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%.

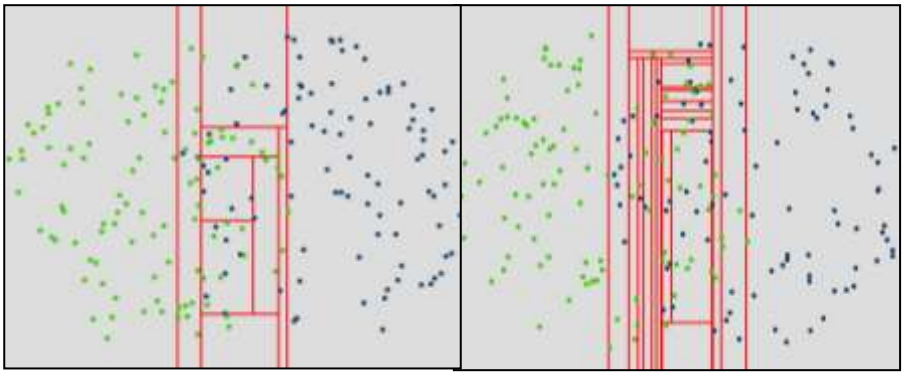


FIGURA 4.16 VARIACIÓN DEL ÁRBOL DE DECISIÓN A DIFERENTES MUESTRAS CON S=10%

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.

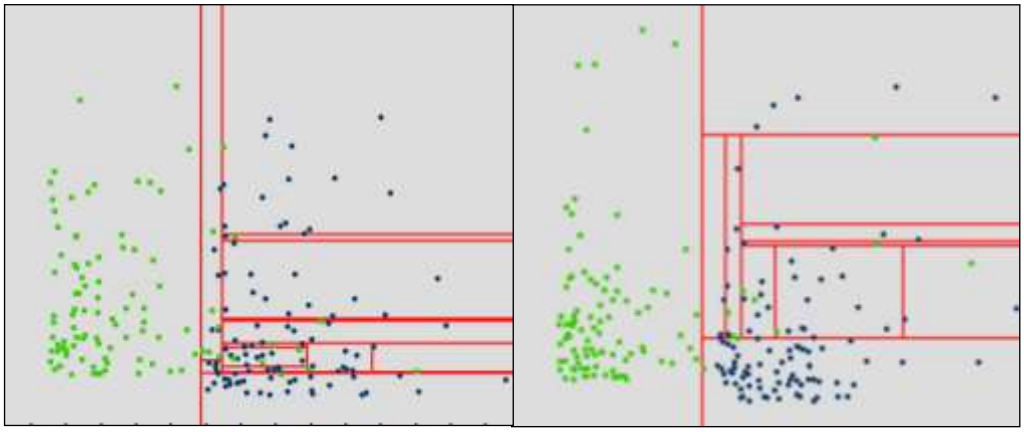


FIGURA 4.17 ÁRBOL DE DECISIÓN PARA 2 MUESTRAS DE POBLACIONES EXPONENCIALES CON S=6%

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.

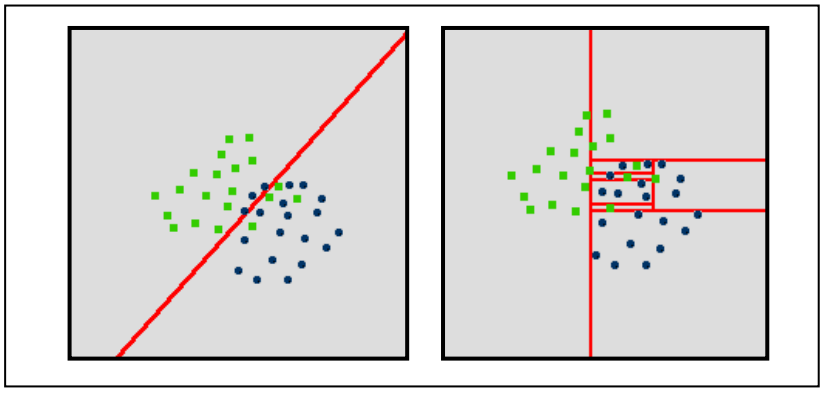


FIGURA 4.18 REGRESIÓN LINEAL DE GRADO 1 Y ÁRBOL DE DECISIÓN: DIFERENCIAS AL CAPTAR EL PATRÓN DE DATOS

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).

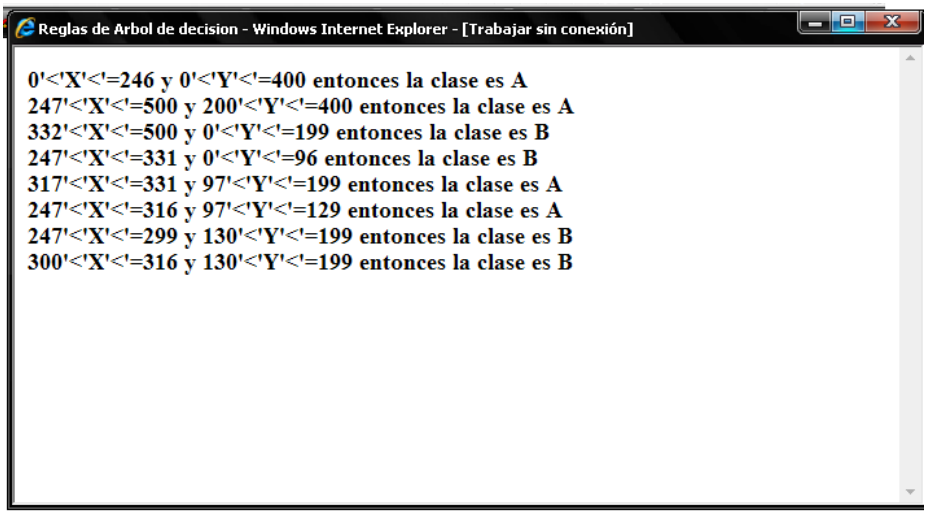


FIGURA 4.19 MODELO DE UN ÁRBOL DE DECISIÓN

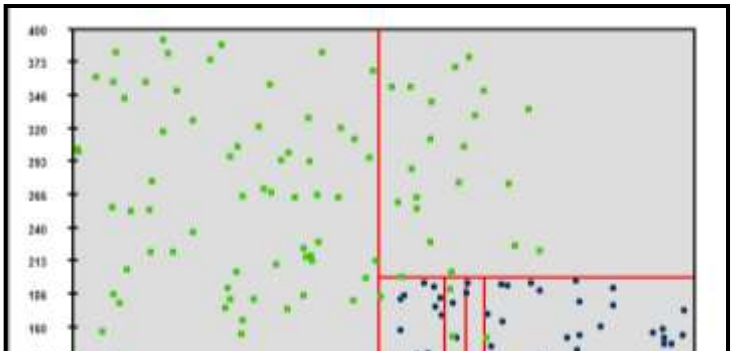


FIGURA 4.20 REPRESENTACIÓN GRÁFICA DEL SISTEMA DE REGLAS DE LA FIGURA 4.19

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

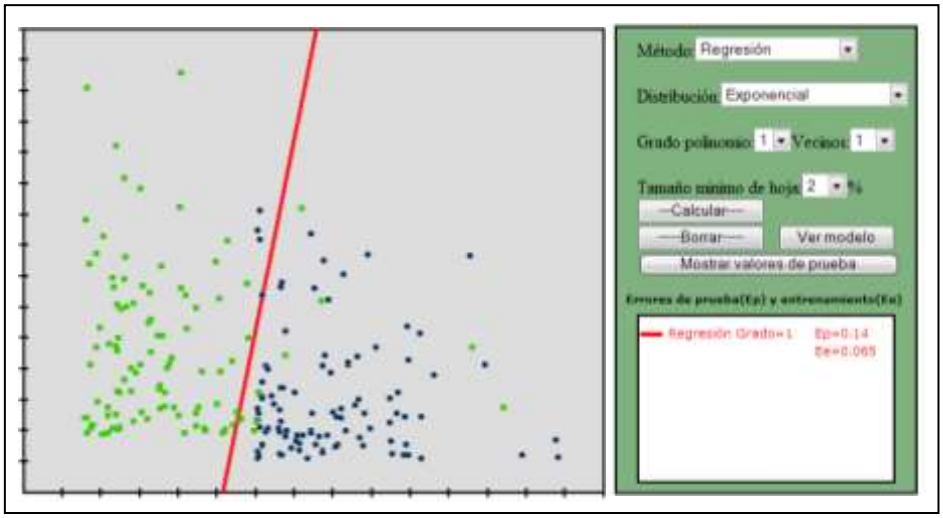


FIGURA 4.21 REGRESION GRADO 1 APLICADO A UN CONJUNTO DE DATOS EXPONENCIAL

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.

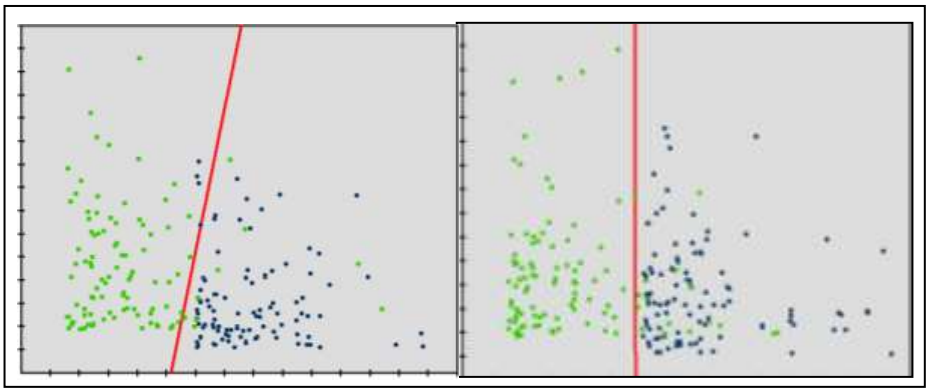


FIGURA 4.22 SENSIBILIDAD DE LA REGRESIÓN LINEAL AL CAMBIAR LA MUESTRA DE UNA POBLACIÓN CON CLASES EXPONENCIALES

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

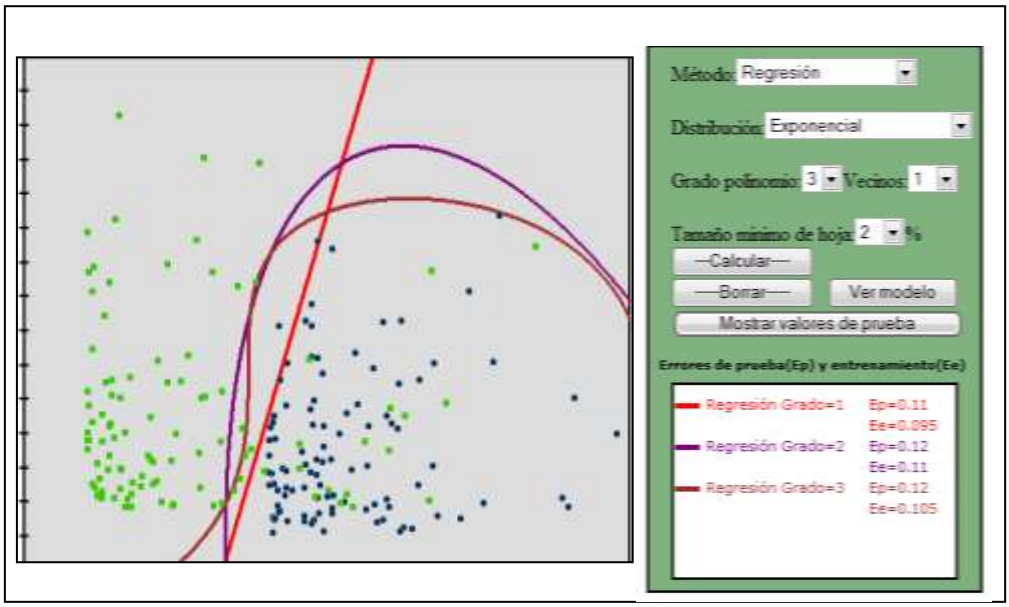


FIGURA 4.23 REGRESION GRADO 1,2 Y 3 PARA DISCRIMINAR CLASES CON DISTRIBUCIÓN EXPONENCIAL

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.

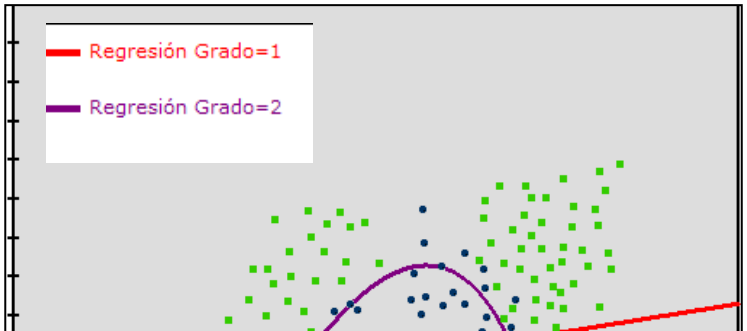


FIGURA 4.24 DIFERENCIAS EN LA EXPRESIVIDAD DE LA REGRESION DE GRADO 1 Y 2

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.

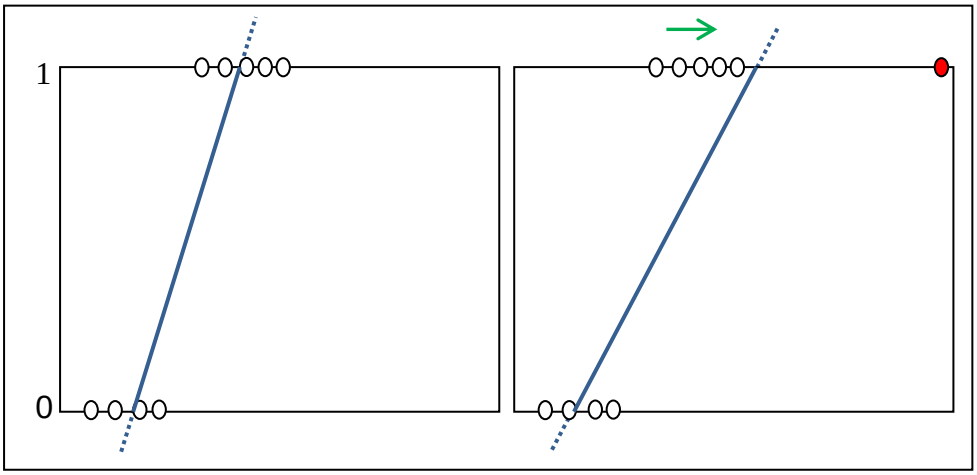


FIGURA 4.25 ESBOZO DE LA SENSIBILIDAD DE LA REGRESIÓN LINEAL ANTE OUTLIERS

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.

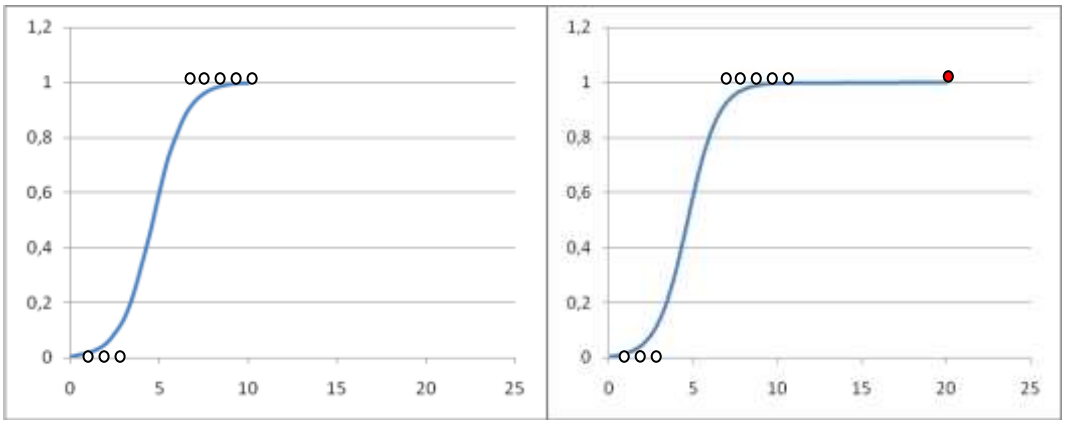


FIGURA 4.26 ESBOZO DE LA ROBUSTEZ DE LA REGRESIÓN LOGÍSTICA ANTE OUTLIERS

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.

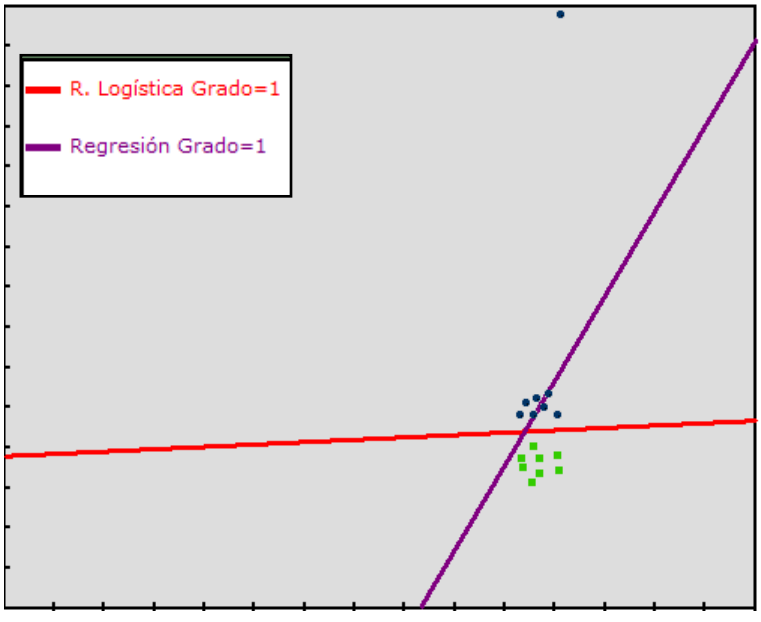


FIGURA 4.27 EJEMPLO DE UNA DIFERENCIA ENTRE REGRESIÓN LINEAL Y REGRESIÓN LOGÍSTICA

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.

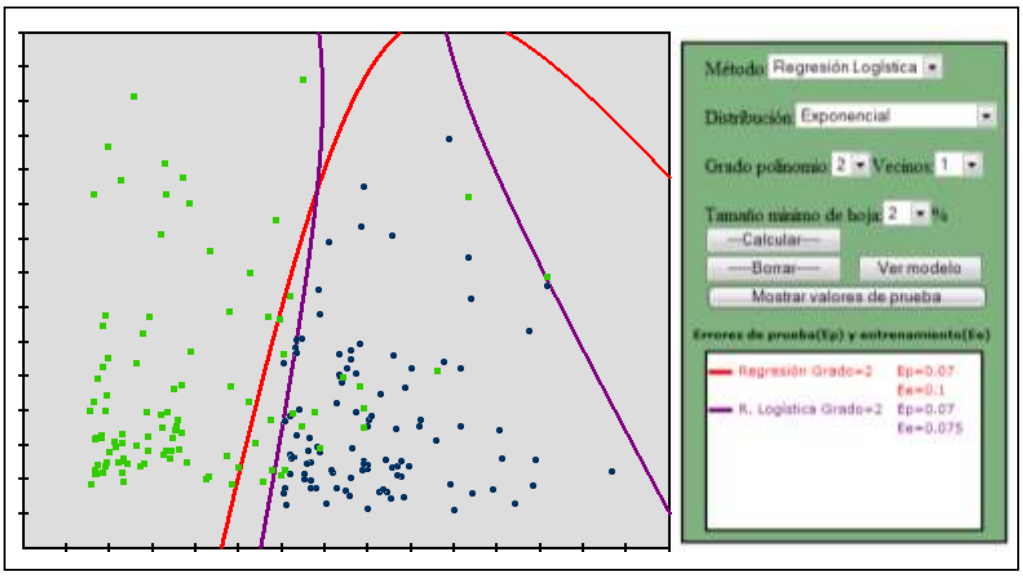


FIGURA 4.28 MÉTODOS DE REGRESIÓN GRADO=2 APLICADOS A UN CONJUNTO CON CLASES DISTRIBUIDAS EXPONENCIALMENTE.

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.

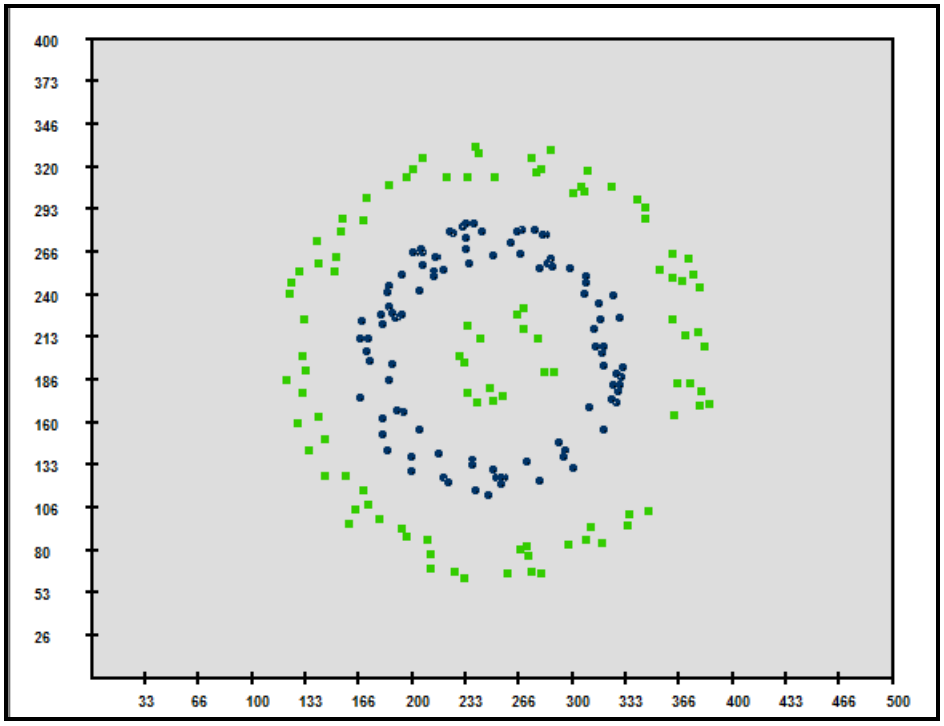


FIGURA 4.29 DISTRIBUCIÓN DE CÍRCULOS CONCÉNTRICOS

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.

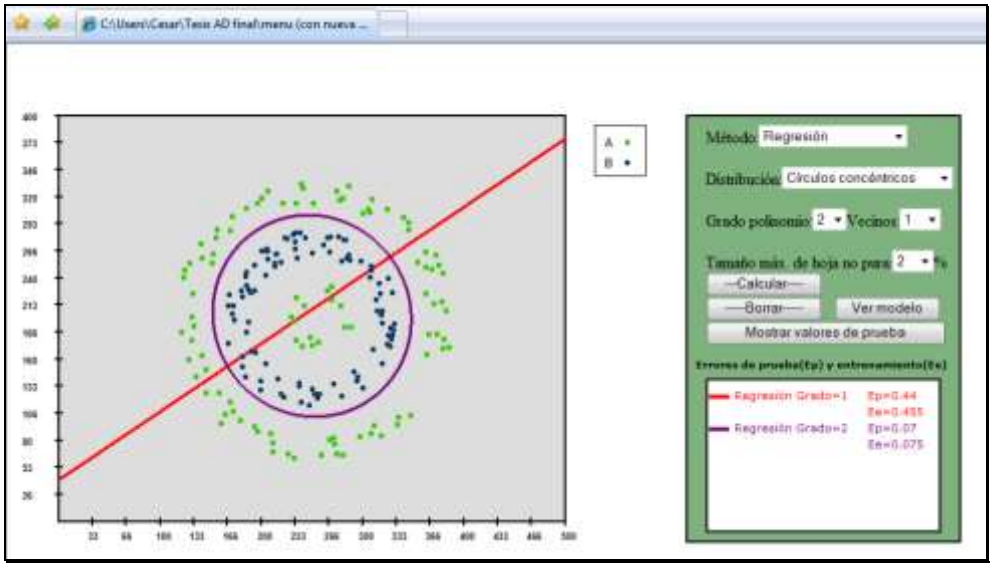


FIGURA 4.30 REGRESIÓN DE GRADO 1 Y 2 APLICADOS A LA DISTRIBUCIÓN DE CÍRCULOS CONCÉNTRICOS

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

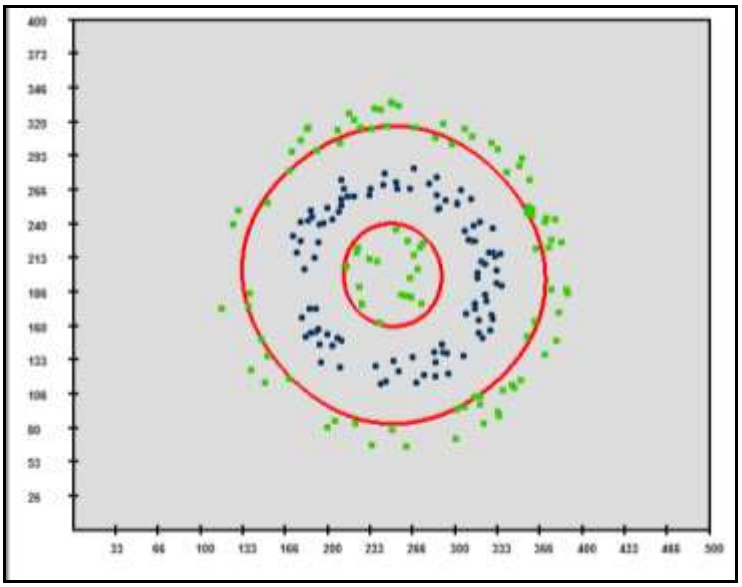


FIGURA 4.31 EXPRESIVIDAD DE LA REGRESION POLINOMIAL GRADO 4 APLICADA A LA DISTRIBUCIÓN DE CÍRCULOS CONCÉNTRICOS

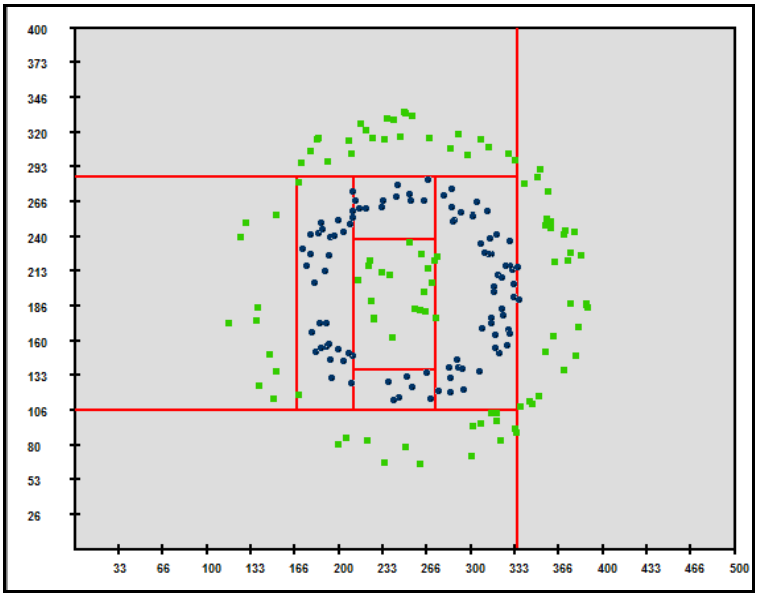


FIGURA 4.32 ÁRBOL DE DECISIÓN APLICADO A LA DISTRIBUCIÓN DE CÍRCULOS CONCÉNTRICOS

CONCLUSIONES Y RECOMENDACIONES

1. Esta página web brinda al usuario la posibilidad de desplazar uno o más puntos de la muestra para observar la forma de la curva discriminante antes y después de dichos movimientos. Sería interesante que el usuario convierta unos pocos puntos del conjunto de datos en datos extremos (*outliers*) y trate luego de aplicar su conocimiento para explicar la reacción o insensibilidad de un determinado método ante dichos cambios extremos. Disponiendo el conjunto de datos convenientemente el usuario podrá observar que la regresión lineal de grado 1 es sensible ante un dato extremo mientras que la regresión logística de grado 1 no siempre será influenciada por ese dato.

2. Puesto que esta herramienta ofrece al usuario la posibilidad de construir su propia distribución de datos, se recomienda que construya conjuntos de datos tan poco comunes o novedosos como la distribución de círculos concéntricos (*Figura 4.29*) u otros conjuntos donde la tarea de discriminación o de captación de patrones no sea una labor sencilla. Esto,

con el objetivo de palpar la potencia o debilidad de cada método según el caso propuesto. Específicamente, en la discriminación de un conjunto de datos con distribución de círculos concéntricos el método de los kvecinos más cercanos captó perfectamente el patrón de datos utilizando valores de k=1,5 y 10.

3. De los 5 métodos que se han expuesto en el presente trabajo se podría decir que no existe uno que sea el mejor para resolver la tarea de discriminación. Sin embargo, sí existen unos métodos que son más efectivos que otros dependiendo del conjunto al cual se apliquen.

4. Cuando se trata de distribuciones irregulares de datos que requieren una gran expresividad uno de los métodos que debería considerarse es el kvecinos más cercanos que aunque es un método retardado ofrece interesantes posibilidades si se escoge un conveniente valor para k.

5. Al comparar el método anticipativo de árbol de decisión con el método retardado de kvecinos más cercanos el usuario podrá comprobar por sí mismo que el costo computacional (*tiempo de respuesta*) de este último método es considerablemente alto comparado con el primero, especialmente si se escoge un valor alto para k.

6. El aplicativo web permite al usuario visualizar hasta 4 curvas discriminantes aplicadas a la misma muestra en el mismo plano. Se recomienda que el usuario aproveche esta posibilidad comparando visualmente distintos métodos aplicados a una misma situación.

7. A excepción del método de Naive Bayes Kernel, los otros 4 métodos tienen parámetros que dependiendo del valor que se les establezca pueden hacer que el comportamiento de los métodos varíe. Se recomienda probar cada método con diferentes valores de su parámetro correspondiente para observar cuál es el cambio que el parámetro provoca en el método y tratar, si es posible, de encontrar un valor idóneo para este parámetro. En relación a esto, el usuario podrá notar que existen grados de poda (*parámetros de parada*) para los cuales la precisión del árbol de decisión es mejor. Al aplicar el árbol de decisión a una muestra con clases distribuidas exponencialmente el error de prueba bajó de 0.11 a 0.06 cuando su parámetro de parada aumentó de 2% a 8%.

8. Valiéndose de los errores de prueba que se registran en el Tablero de resultados de la página web el usuario podrá llevar a cabo experimentos globales que consisten en la aplicación de todos los métodos con distintos valores de sus parámetros. Particularmente, en un experimento global

aplicado a un conjunto de datos en el cual cada atributo de cada clase se distribuía exponencialmente, el método que obtuvo el menor error de prueba fue el árbol de decisión con el 94% de aciertos.

9. Aunque tanto los métodos de regresión como el árbol de decisión proporcionan un modelo general con el cual discriminar nuevos elementos, el sistema de reglas que produce el árbol de decisión es más comprensible desde el punto de vista humano que los modelos de regresión.

10. Aunque el método de árbol de decisión es comprensible, computacionalmente eficiente y produce en general buenos resultados, adolece de limitada expresividad en ciertos conjuntos de datos y de inestabilidad ante variaciones de la muestra. Específicamente, en 6 muestras tomadas de una población compuesta por 2 clases distribuidas exponencialmente la eficacia de este método al discriminar varió del 87% al 96% al utilizar un parámetro de parada del 16%.

11. El método de Naive Bayes Kernel clasificó correctamente el 91% de los datos de prueba al ser aplicado a una muestra con distribución de Cuadrados traslapados. En este caso Naive Bayes superó en 2% a la regresión polinómica de grado 2. Sin embargo, cuando no se está seguro

de la independencia condicional de los atributos que determinan los elementos de la muestra es aconsejable contrastar este método con otros que no tengan el limitante de la independencia.

12. Al aumentar la complejidad de los modelos de regresión no siempre se logra una mejoría en el error de prueba, antes esta complejidad puede conllevar a que el modelo se sobreajuste a los datos de entrenamiento. Específicamente, al aumentar la complejidad de la regresión polinómica en la discriminación de una muestra con clases distribuidas exponencialmente no se logró ninguna mejoría. Con la regresión de grado 1 se obtuvo el error de prueba de 0.11 mientras que con las regresiones de grado 2 y 3 este error se incrementó ligeramente a 0.12.

13. Aunque los métodos de regresión no son tan expresivos como los kvecinos más próximos, éstos a veces pueden ser muy útiles para captar el patrón de ciertos conjuntos de datos y discriminar correctamente cuando el grado de sus polinomios es idóneo. En particular, en un experimento diseñado para medir la capacidad de cada método para captar el patrón de la distribución de círculos concéntricos, el método de regresión polinomial de grado 4 captó bien el patrón logrando una eficacia del 97%.