Clasificador de Redes Neurales

Resumen

El **Clasificador Probabilístico de Redes Neurales (PNN, Probabilistic Neural Network Classifier)** ejecuta un método no paramétrico para clasificar observaciones en uno de *g* grupos basados en *p* variables cuantitativas observadas. Más que hacer supuestos sobre la naturaleza de la distribución de las variables dentro de cada grupo, construye una estimación no paramétrica de la función de densidad de cada grupo en una localización deseada basada en las observaciones colindantes a ese grupo. La estimación se construye usando una ventana de Parzen que pondera observaciones de cada grupo de acuerdo con su distancia a la localización especificada.

Las observaciones son asignadas a los grupos con base en el producto de tres factores:

1. La función de densidad estimada en la vecindad del punto.
2. Las probabilidades previas de pertenecer a cada grupo.
3. El costo de clasificar incorrectamente casos que pertenecen a un grupo dado.

La esfera de influencia de la función de ponderación de Parzen puede ser definida por el usuario u optimizada vía jackknifing.

StatFolio de Ejemplo: *neuralclassifier.sgp*

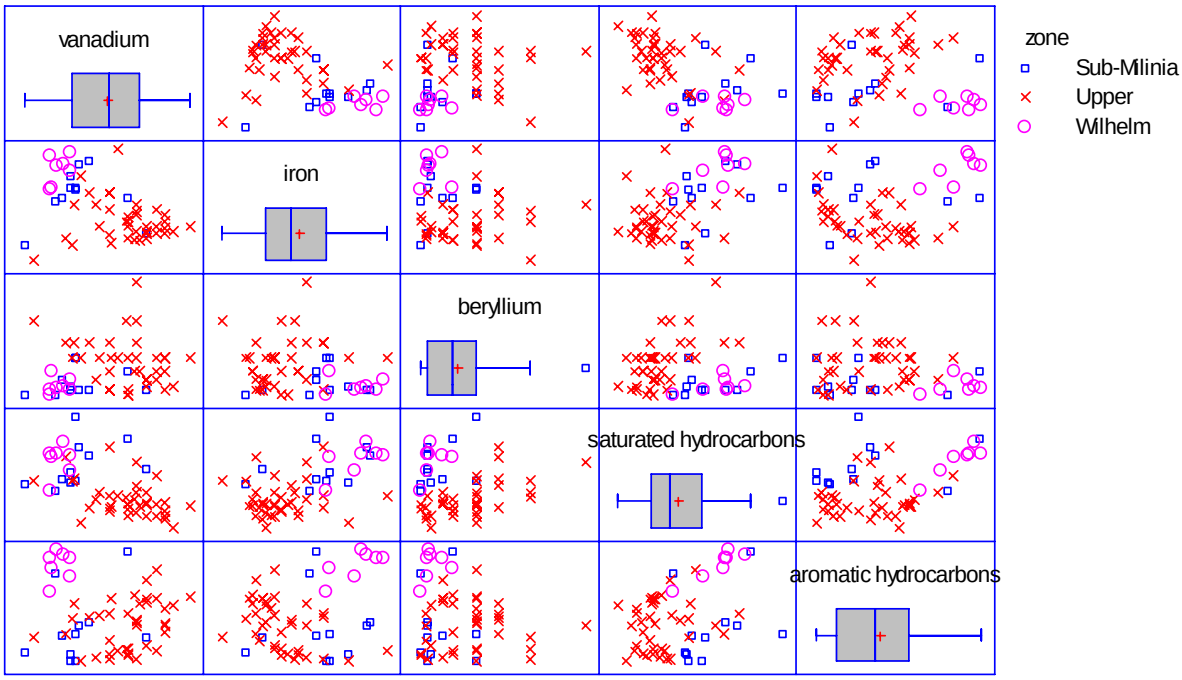
Datos de Ejemplo:

El archivo *sandstone.sf6* contienes un conjunto de datos tomados de Gerrild y Lantz (1969) que es descrito por Johnson y Wichern (2002). Los datos consisten de un total de $n = 56$ muestras de piedra arenisca provenientes de tres zonas: *Wilhelm*, *Sub-Milinia*, y *Upper*. Cada muestra ha sido analizada químicamente y se midieron el valor de cinco variables. La tabla a continuación muestra una lista parcial de los datos en ese archivo:

Sample (muestra)	*Vanadium (vanadio)*	*Iron* (hierro)	*Beryllium* (berilio)	*Saturated hydrocarbons* (hidrocarburos saturados)	*Aromatic hydrocarbons* (hidrocarburos aromáticos)	*Zone* (zona)
1	3.9	51	0.20	7.06	12.19	Wilhelm
2	2.7	49	0.07	7.14	12.23	Wilhelm
3	2.8	36	0.30	7.00	11.30	Wilhelm
4	3.1	45	0.08	7.20	13.01	Wilhelm
5	3.5	46	0.10	7.81	12.63	Wilhelm
6	3.9	43	0.07	6.25	10.42	Wilhelm
7	2.7	35	0.00	5.11	9.00	Wilhelm
8	5.0	47	0.07	7.06	6.10	Sub-Milinia
9	3.4	32	0.20	5.82	4.69	Sub-Milinia
10	1.2	12	0.00	5.54	3.15	Sub-Milinia
…	…	…	…	…	…	

Es de interés poder clasificar muestras en zonas con base en las 5 mediciones.

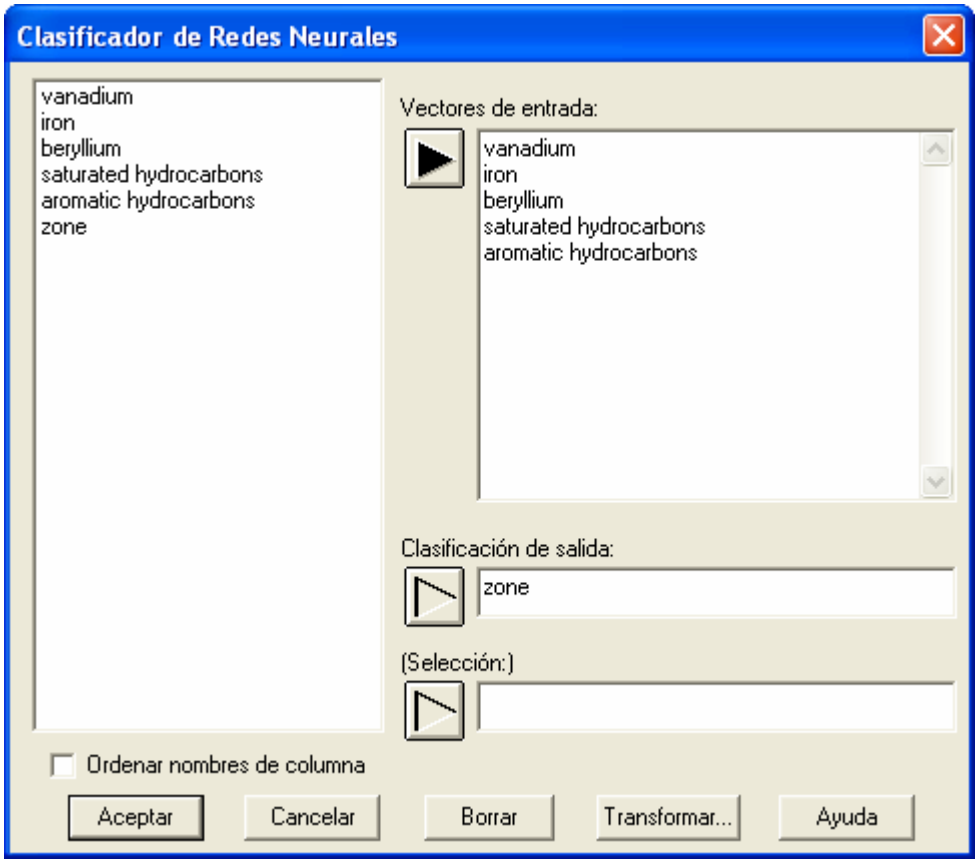
A continuación se muestra un gráfico de matriz de los datos observados:



Hay bastante agrupamiento, pero también bastante traslape.

Ingreso de Datos

La caja de diálogo del ingreso de datos solicita los nombres los nombres de *p* variables de entrada que se usarán para clasificar los casos y una sola variable de salida que define los grupos conocidos a partir de la muestra de entrenamiento:



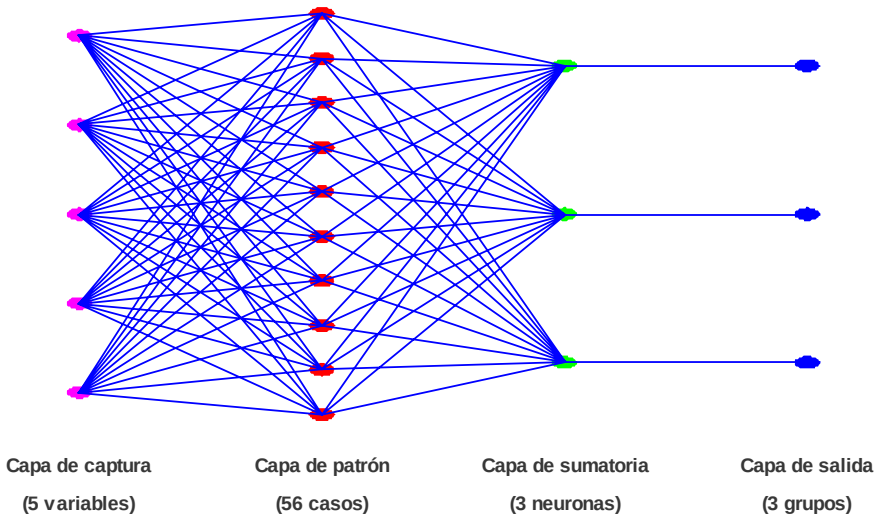
- **Vectores de Entrada:** los nombres de las *p* variables de entrada, las cuales deben ser factores cuantitativos que caractericen las muestras.

- **Factor de Clasificación:** variable de salida numérica o no numérica que contenga un identificador de a qué grupo pertenece cada observación. El número de valores diferentes en esta columna se representará con *g*.

- **Selección:** selección de un subgrupo de datos.

Diagrama de Red

El problema de clasificar casos puede formularse como una red neural. El *Diagrama Neural* ilustra la estructura básica de la red:



La red consiste de cuatro capas:

1. Una *capa de entrada* con *p* neuronas, una para cada variable de entrada.
2. Una *capa patrón* con *n* neuronas, una para cada caso que se usará para entrenar a la red.
3. Una *capa de resumen* con *g* neuronas, una para cada clase de salida.
4. Una *capa de salida*, que también tiene una neurona binaria para cada clase de salida que se enciende o apaga dependiendo de si un caso se asigna o no al grupo correspondiente.

Conceptualmente, la *capa de entrada* provee la información de las *p* variables predictoras alimentando con sus valores (estandarizados restándoles la media y dividiéndolos entre la desviación estándar) a las neuronas de la *capa patrón*. Las neuronas pasan los valores a través de una *función de activación*, que usa los valores de entrada para estimar la función de densidad de probabilidad para cada grupo en una localización determinada. Las estimaciones de densidad pasan luego a la capa de resumen, la cual combina la información de los *n* casos de entrenamiento con probabilidades previas y costos de clasificación errónea para derivar un puntaje para cada grupo. Los puntajes entonces se usan para encender la neurona binaria en la capa de salida correspondiente al grupo con el puntaje más alto y apaga todas las demás neuronas de salida.

Capa de Entrada

Las neuronas en la capa de entrada representan los valores de las *p* variables de entrada. Estos valores, denotados por X_1 a X_p, son estandarizados sustrayendo la media muestral de los *n* casos de entrenamiento y dividiendo entre la desviación estándar muestral. Los valores estandarizados pasan luego a la capa patrón.

Capa Patrón

La capa patrón toma cada variable de entrada X_i y calcula su contribución a la estimación de la función de densidad de probabilidad para el grupo al que pertenece pasándola a través de una función de activación. En esta red, la función de activación cuantifica la contribución del *i-ésimo* valor en el caso de entrenamiento a la estimación de la función de densidad para el grupo *j* y está dada por

$$g_{ij} = W\left(\frac{X - X_i}{\sigma}\right) \text{ si la observación } i \text{ pertenece al grupo } j \quad (1)$$

y

$$g_{ij} = 0 \text{ en cualquier otro caso.} \quad (2)$$

σ es un parámetro de escalamiento que controla qué tan rápido decae la influencia de un punto como una función de su distancia a *X*. Debido a su forma, frecuentemente se considera que la función W es la función Gaussiana definida por

$$\exp\left(-\frac{\|X - X_i\|^2}{\sigma^2}\right) \quad (3)$$

donde $\|X - X_i\|^2$ es el cuadrado de la distancia Euclidiana entre *X* y X_i.

Capa de Resumen

Las neuronas en la capa de resumen combinan la información de todos los miembros del grupo de entrenamiento. Siendo n_j el número de observaciones en el grupo de entrenamiento que pertenecen al grupo *j*, la función de densidad estimada para el grupo *j* en la localización X es proporcional a

$$g_j(X) = \frac{1}{n_j}\sum_{i=1}^{n} g_{ij} \quad (4)$$

Para determinar a qué grupo debe de ser asignada la observación en la localización X, se necesitan otras dos cantidades:

1. La probabilidad previa h_j de que una observación pertenezca al grupo *j*, sin considerar las variables de entrada. Esto representaría normalmente la proporción relativa de todas las muestra en la población que pertenecen al grupo *j*. Usando las *Opciones de Análisis*, se puede asumir que las probabilidades previas son iguales, se pueden representar por la fracción de la muestra de entrenamiento que viene con cada grupo, o ser ingresadas por el usuario.

2. El costo c_j de clasificar incorrectamente una observación que pertenece al grupo *j*. En algunos casos, como cuando se investiga la presencia de una enfermedad, clasificar incorrectamente a un individuo que pertenece a un grupo (tiene la enfermedad) puede ser más serio que clasificar mal a un individuo que pertenece al otro grupo (no tiene la enfermedad).

La capa de resumen asigna un puntaje a cada grupo multiplicando la función de densidad estimada por la probabilidad previa y el costo:

$$Puntaje_j = h_j c_j g_j(X) \quad (5)$$

Estos puntajes pasan luego a la capa de salida.

Capa de Salida

Las neuronas en la capa de salida son binarias y sólo pueden encenderse o apagarse. Con base en los puntajes, la neurona de salida *j* se enciende si

$$Puntaje_j > Puntaje_k \quad (6)$$

para todo *k* diferente de *j*. De otro modo, se apaga. En caso de un empate, la neurona que se enciende se determina aleatoriamente.

Entrenando a la Red

El único parámetro en la formulación de la red que puede variarse es el parámetro de escala σ, que afecta cuán rápido decae la influencia de una observación sobre la densidad en el punto X conforme aumenta su distancia al punto *X*. El procedimiento provee 3 opciones para determinar σ:

1. σ puede ser definido por el usuario. El valor por omisión es σ = 1, lo que no carece de razón ya que las variables de entrada se han estandarizado.
2. σ puede ser ignorada y una observación en el punto *X* siempre coincidirá con el grupo correspondiente a su vecino más cercano.
3. Se puede tratar con diferentes valores de σ y escoger el que maximice el porcentaje de las *n* observaciones del grupo de entrenamiento que se clasifiquen correctamente.

Cuando se selecciona el tercer caso, la red se entrena usando un procedimiento llamado *jackknifing*. Jackknifing remueve un punto a la vez del grupo de entrenamiento y determina qué tan frecuentemente se clasifica correctamente cuando *no* se usa para estimar los puntajes de los grupos. Se selecciona el valor de σ que clasifica correctamente el porcentaje más alto de puntos removidos.

Resumen del Análisis

El *Resumen del Análisis* resume el desempeño del algoritmo en el grupo de entrenamiento:

Clasificador Bayesiano de Redes Neurales - zone
Factor de clasificación: zone
Factores:
vanadium (percent ash)
iron (percent ash)
beryllium (percent ash)
saturated hydrocarbons (percent area)
aromatic hydrocarbons (percent area)

Número de casos en el conjunto de entrenamiento: 56
Número de casos en el conjunto de validación: 0

Parámetro de espaciamiento usado: vecino más cercano

Conjunto de Entrenamiento

		Porcentaje Correctamente
zone	*Miembros*	*Clasificado*
Sub-Milinia	11	63.6364
Upper	38	100.0
Wilhelm	7	85.7143
Total	56	91.0714

Conjunto de Validación

		Porcentaje Correctamente
zone	*Miembros*	*Clasificados*
Sub-Milinia	0	
Upper	0	
Wilhelm	0	
Total	0	

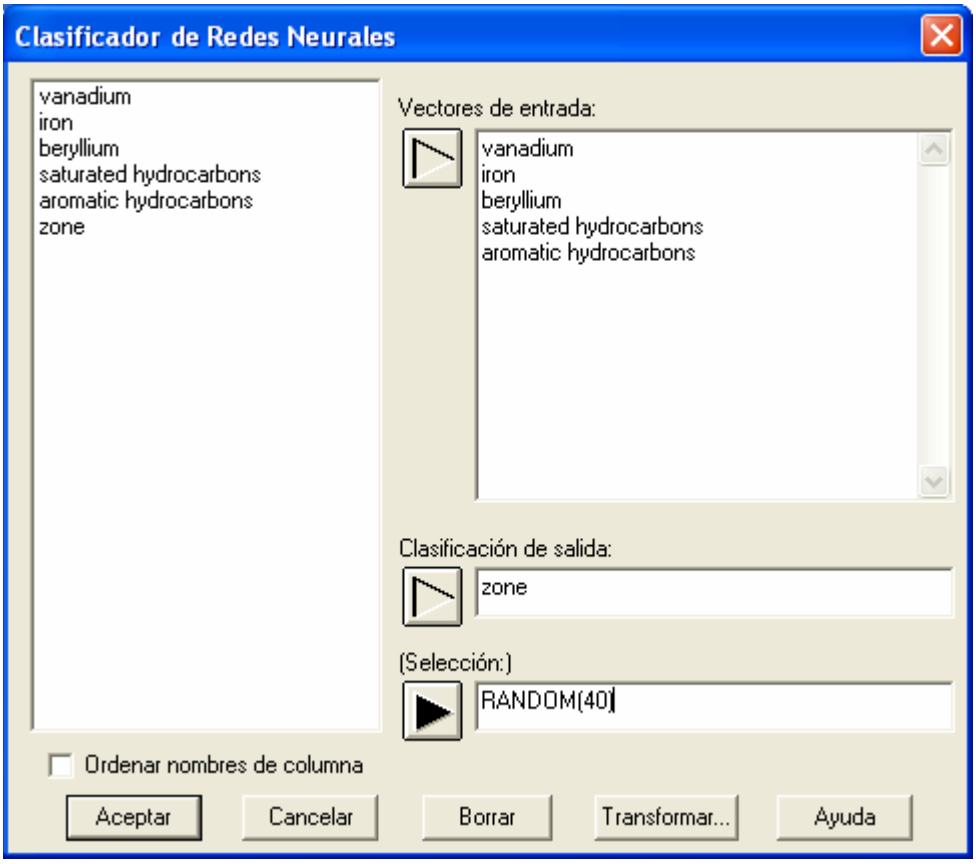
En la tabla se incluyen:

- **Factores**: identificación de las variables de entrada.
- **Número de casos en el conjunto de entrenamiento**: el número de observaciones *n* en el grupo de entrenamiento.
- **Número de casos en el conjunto de validación**: el número de casos retenidos para no formar parte del grupo de entrenamiento. Los casos pueden ser retenidos usando el campo *Selección* en la caja de diálogo del ingreso de datos.
- **Parámetro de espaciamiento usado**: método para determinar la función de densidad de probabilidad. Si se especifica σ por el usuario o se estima por jackknifing, se desplegará su valor. Si cada punto se hace coincidir con su *vecino más cercano*, así se indicará.
- **Conjunto de Entrenamiento**: número y porcentaje de observaciones del grupo de entrenamiento que fueron clasificadas correctamente.
- **Conjunto de Validación**: número y porcentaje de observaciones retenidas para no formar parte del grupo de entrenamiento que fueron clasificadas correctamente.

Por ejemplo, la tabla anterior muestra que el criterio del vecino más cercano clasificó correctamente un poco más del 91% de las observaciones. Esto podría servir como punto de referencia con el cual compara otros métodos. También es interesante advertir que el criterio del vecino más cercano funciona bien con los grupos más grandes, pero no lo es, ni con mucho, con los grupos más pequeños. Esto es de esperarse, ya que entre mayor sea el grupo provee de mayores oportunidades de tener un vecino en la cercanía. El grupo *Sub-Milinia* es particularmente difícil de clasificar, ya que tiende a estar más disperso que los otros dos grupos.

Ejemplo -Reteniendo un conjunto aleatorio de observaciones

Como se mencionó anteriormente, el campo *Selección* en la caja de diálogo del ingreso de datos puede ser usado para retener observaciones evitando que formen parte del grupo de entrenamiento. Por ejemplo, uno puede retener aleatoriamente 16 de las 56 observaciones usando la función ALEATORIO, como se muestra a continuación:



Las 40 observaciones elegidas aleatoriamente se usarán como el *grupo de entrenamiento*, mientras que las 16 restantes formarán el *grupo de validación.*

Clasificador Bayesiano de Redes Neurales - zone (RANDOM(40))

Factor de clasificación: zone
Factores:
vanadium (percent ash)
iron (percent ash)
beryllium (percent ash)
saturated hydrocarbons (percent area)
aromatic hydrocarbons (percent area)
Selección de la Variable: RANDOM(40)

Número de casos en el conjunto de entrenamiento: 39
Número de casos en el conjunto de validación: 17

Parámetro de espaciamiento usado: vecino más cercano

Conjunto de Entrenamiento

		Porcentaje Correctamente
zone	*Miembros*	*Clasificado*
Sub-Milinia	9	55.5556
Upper	25	100.0
Wilhelm	5	80.0
Total	39	87.1795

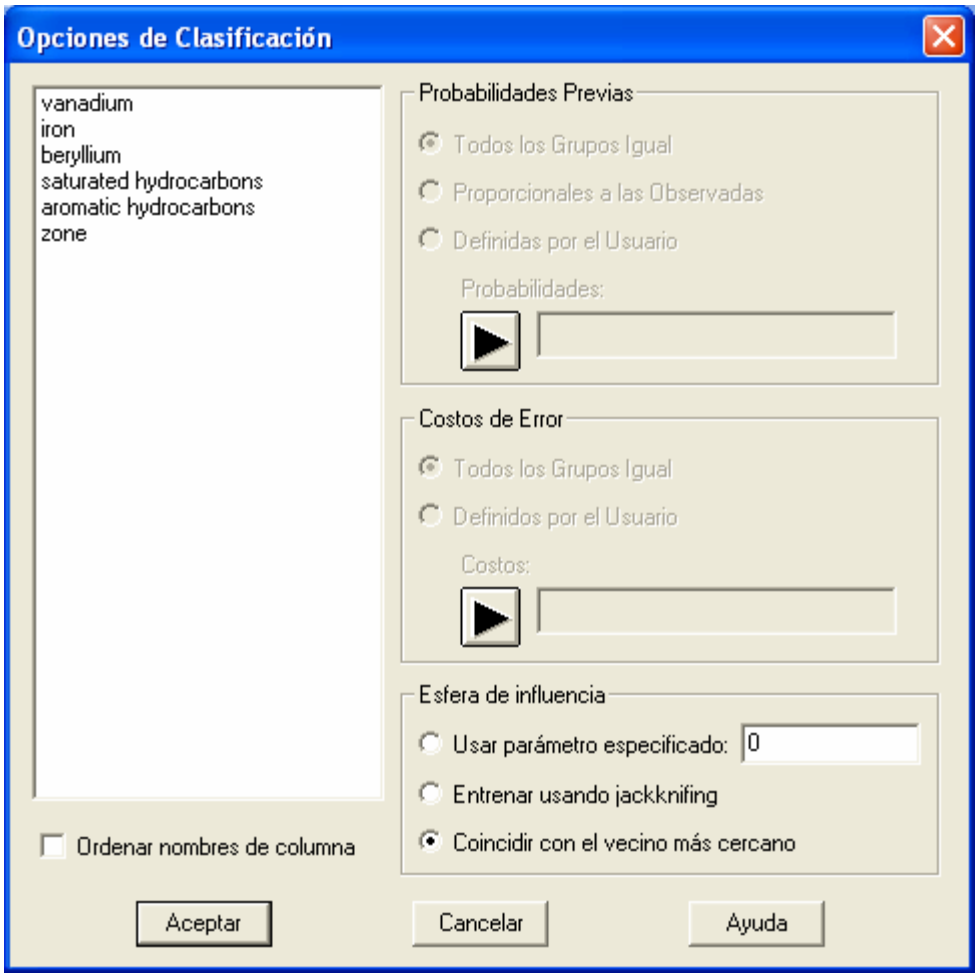
Conjunto de Validación

		Porcentaje Correctamente
zone	*Miembros*	*Clasificados*
Sub-Milinia	2	50.0
Upper	13	92.3077
Wilhelm	2	100.0
Total	17	88.2353

Cierto, el conjunto de validación es extremadamente pequeño, pero confirma que las muestras *Sub-Milinia* son difíciles de clasificar correctamente.

Opciones de Análisis

La caja de diálogo de *Opciones de Análisis* permite al usuario controlar el algoritmo de clasificación:



- **Probabilidades Previas:** método para determinar la probabilidad de la pertenencia a un grupo antes de examinar los datos. Seleccione *Todos los Grupos Igual* para suponer probabilidades iguales para todos los grupos, *Proporcional a lo Observado* para establecer las probabilidades igual a la fracción del grupo de entrenamiento representada por cada grupo, o *Definida por el Usuario* para ingresar una columna con *g* valores que sumen 1.

- **Costos del Error:** costos relativos por clasificar incorrectamente un miembro de cada grupo. *Todos los Grupos Igual* asigna costos iguales a todos los grupos. Si son *Definidos por el Usuario*, ingrese una columna con *g* valore positivos.

- **Esfera de Influencia**: método para estimar la función de densidad. *Usar el parámetro especificado* especifica el valor de σ deseado. *Entrenar usando jackknifing* retiene valores del conjunto de entrenamiento uno a la vez y determina σ con base en el porcentaje de veces que el punto retenido es clasificado correctamente. *Coincidir con el vecino más cercano* ignora todas las probabilidades previas y costos para hacer coincidir cada punto con el grupo al que corresponde su vecino más cercano en el espacio de las variables X.

Ejemplo – Optimizar sigma usando jackknifing

Las siguientes salidas muestran los resultados de optimizar σ usando el método jackknife:

Clasificador Bayesiano de Redes Neurales - zone

Factor de clasificación: zone
Factores:
- vanadium (percent ash)
- iron (percent ash)
- beryllium (percent ash)
- saturated hydrocarbons (percent area)
- aromatic hydrocarbons (percent area)

Probabilidades previas: no informativa
Costos de error: igual para todos los casos

Número de casos en el conjunto de entrenamiento: 56
Número de casos en el conjunto de validación: 0

Parámetro de espaciamiento usado: 0.0 (optimizado por jackknifing durante el entrenamiento)

Conjunto de Entrenamiento

		Porcentaje Correctamente
zone	*Miembros*	*Clasificado*
Sub-Milinia	11	63.6364
Upper	38	100.0
Wilhelm	7	85.7143
Total	56	91.0714

A pesar de tratar con un gran número de valores diferentes de σ, el mejor valor fue 0, que corresponde al método del vecino más cercano. Fijando el valor de σ, se advertirá que incluso una valor positivo pequeño de σ reduce el porcentaje de casos clasificados correctamente:

Clasificador Bayesiano de Redes Neurales - zone

Factor de clasificación: zone
Factores:
- vanadium (percent ash)
- iron (percent ash)
- beryllium (percent ash)
- saturated hydrocarbons (percent area)
- aromatic hydrocarbons (percent area)

Probabilidades previas: proporcional a la ocurrencia en el conjunto de entrenamiento
Costos de error: igual para todos los casos

Número de casos en el conjunto de entrenamiento: 56
Número de casos en el conjunto de validación: 0

Parámetro de espaciamiento usado: 0.25 (especificado por usuario)

Conjunto de Entrenamiento

		Porcentaje Correctamente
Zone	*Miembros*	*Clasificado*
Sub-Milinia	11	45.4545
Upper	38	100.0
Wilhelm	7	85.7143
Total	56	87.5

Tabla de Clasificación

La *Tabla de Clasificación* muestra el resultado de usar el criterio de clasificación obtenido para asignar casos observados y nuevos a los grupos. Para un conjunto de valores *X*, un caso es asignado al grupo que dé el puntaje más alto, donde el puntaje se basa en la función de densidad de probabilidad estimada, la probabilidad previa, y el costo del error. Dado que el tamaño de las poblaciones de las que se toman las muestra de cada grupo puede no ser el mismo, la probabilidad de que un individuo pertenezca a un grupo en particular antes de examinar los datos puede variar de grupo a grupo. Por ejemplo, investigando una enfermedad, la proporción de individuos a los que se les aplica una prueba diagnóstica que realmente tienen la enfermedad puede ser muy pequeña, un hecho que necesita tomarse en cuenta. Usando *Opciones de Ventana*, el usuario define como manejar las probabilidades previas. Puede suponerse que sean iguales para todos los grupos, que sean proporcionales a la fracción de los datos en cada grupo, o pueden ser ingresadas por el usuario.

La tabla a continuación muestra una típica salida:

Tabla de Clasificación

Actual	*Tamaño*	*Predicción para*		
zone	*de Grupo*	*Sub-Milinia*	*Upper*	*Wilhelm*
Sub-Milinia	11	7	2	2
		(63.64%)	(18.18%)	(18.18%)
Upper	38	0	38	0
		(0.00%)	(100.00%)	(0.00%)
Wilhelm	7	1	0	6
		(14.29%)	(0.00%)	(85.71%)

Porcentaje de casos de entrenamiento correctamente clasificados: 91.07%

	Probabilidad	*Costo de*
zone	*Previa*	*Error*
Sub-Milinia	0.3333	1.0
Upper	0.3333	1.0
Wilhelm	0.3333	1.0

	Grupo	*Vecino*	*Distancia*	*Vecino*	*Distancia*
Fila	*Actual*	*Más Cercano*	*Más Cercano*	*2º Más Cercana*	*2º Más Cercana*
7	Wilhelm	Sub-Milinia*	0.161634	Wilhelm	0.2889
11	Sub-Milinia	Upper*	0.227325	Sub-Milinia	0.334108
12	Sub-Milinia	Upper*	0.265937	Sub-Milinia	0.315371
16	Sub-Milinia	Wilhelm*	0.294938	Upper	0.298606
18	Sub-Milinia	Wilhelm*	0.135704	Sub-Milinia	0.397795
57		Sub-Milinia	0.202902	Upper	0.22956

* = incorrectamente clasificado.

La sección superior de la tabla muestra qué tan bien el criterio de clasificación sirvió para clasificar los datos de entrenamiento. Cada fila presenta los resultados para los casos que de hecho pertenecen a un grupo en particular. Las columnas muestran con qué frecuencia los casos se clasificaron como pertenecientes a cada grupo. En la parte inferior se muestra el porcentaje de casos clasificados correctamente.

La parte central de la tabla presenta las probabilidades previas. Para los datos del ejemplo, se supuso que las probabilidades previas eran las mismas para todos los grupos.

La parte inferior de la tabla muestra los dos grupos que recibieron los puntajes más altos para casos selectos. La tabla muestra:

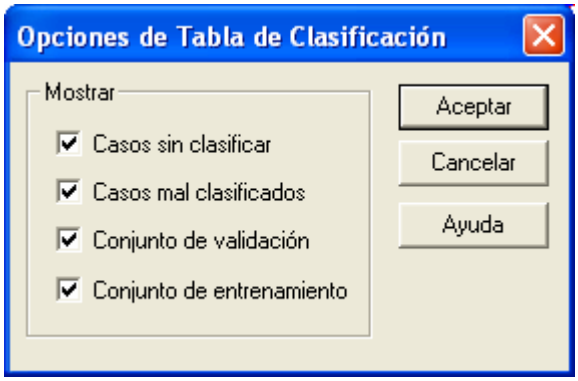
1. *Grupo más alto y segundo más alto* (o *Vecino más Cercano y segundo más cercano*) – los dos grupos con los puntajes más altos o menores distancias.
2. *Puntaje más alto y segundo más alto* (o *Distancia más cercana y segunda más cercana*) – los puntajes o distancias más cercanas de los dos grupos.

Dependiendo de las *Opciones de Ventana*, la tabla puede incluir todas las filas en la hoja de datos o solo filas seleccionadas. Se pueden incluir predicciones para observaciones para las cuales no se conozca el grupo de pertenencia agregando información adicional para las variables *X* en la hoja de datos pero dejando en blanco la celda para el indicador de grupo. Por ejemplo, suponga que se tomó una nueva muestra con las siguientes mediciones:

vanadium = 5.7 por ciento en ceniza
iron = 32 por ciento en ceniza
beryllium = 0.5 por ciento en ceniza
saturated hydrocarbons = 4.99 por ciento en área
aromatic hydrocarbons = 3.62 por ciento en área

Estos valores se pondrían en la fila #57 de la hoja de datos. La tabla muestra que el grupo con el mayor puntaje para estos valores es *Sub-Milinia*, seguido por *Upper*.

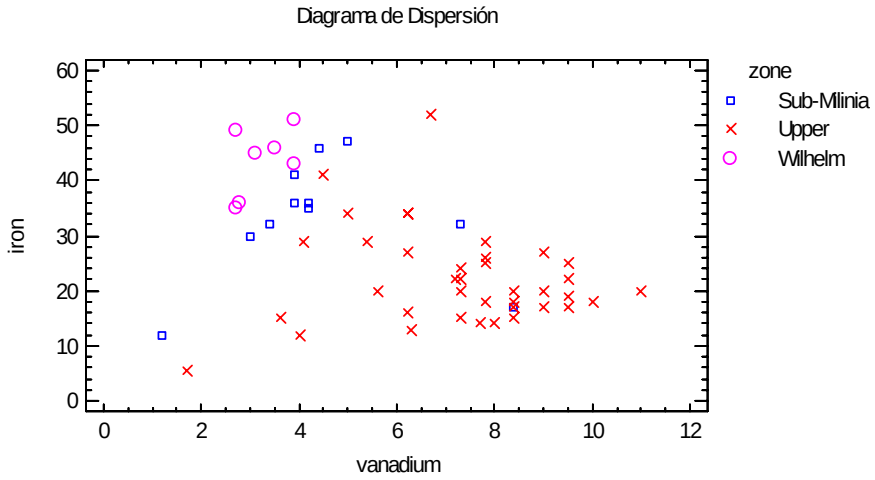
Opciones de Ventana



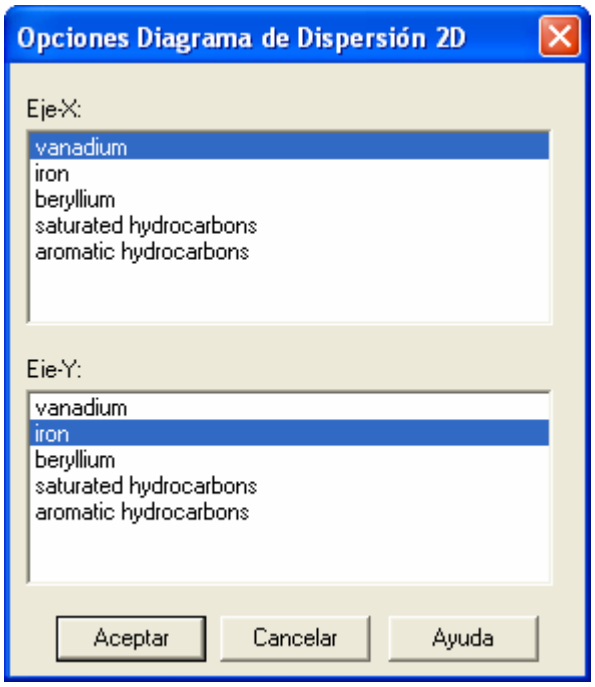
- **Mostrar**: *Casos sin Clasificar* incluirá todas las observaciones en la base de datos con información para las variables X pero sin indicador de pertenencia a un grupo. *Casos Mal Clasificados* incluirá todos los casos que sean clasificados incorrectamente. *Conjunto de Validación* incluirá todos los casos en el grupo de validación. *Conjunto de Entrenamiento* incluirá todos los casos en el grupo de entrenamiento.

Diagrama de Dispersión 2D

El *Diagrama de Dispersión 2D* grafica los datos para cualesquiera dos de las variables *X*.



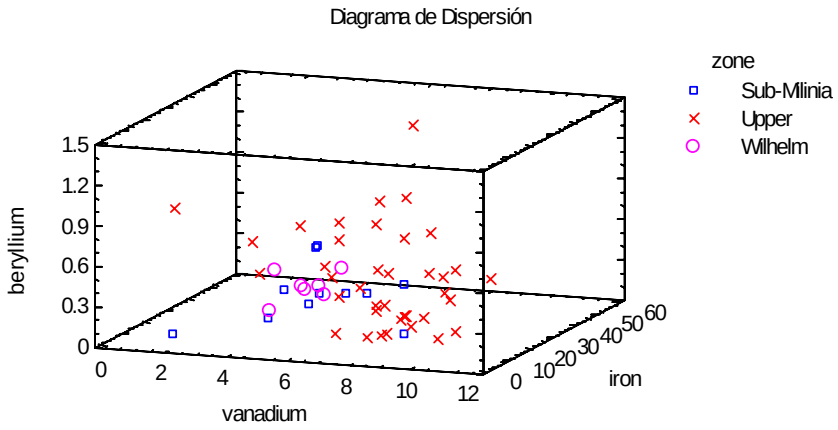
Opciones de Ventana



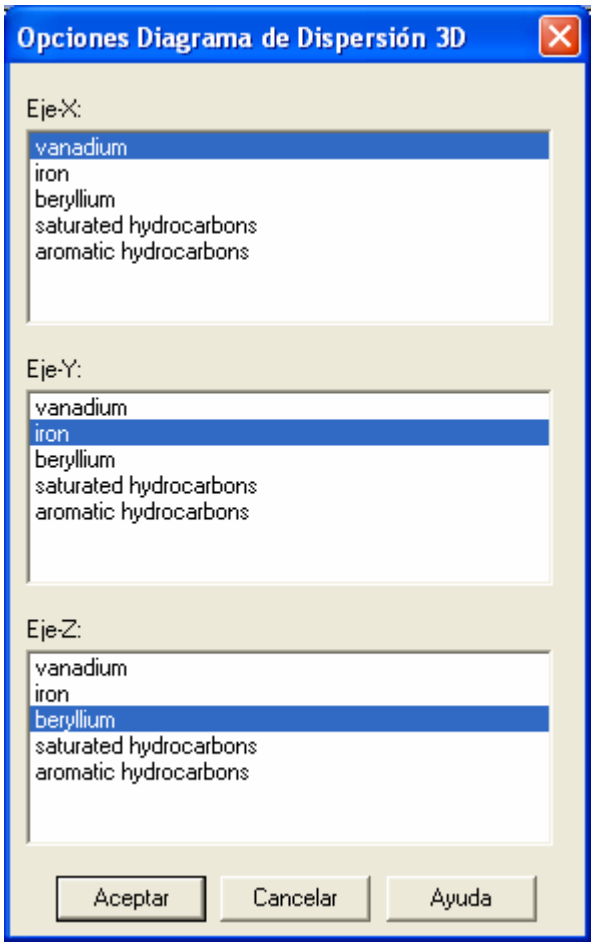
Seleccione variables para definir los ejes horizontal y vertical.

Diagrama de Dispersión 3D

El *Diagrama de Dispersión 3D* grafica los datos para cualesquiera tres de las variables *X*.



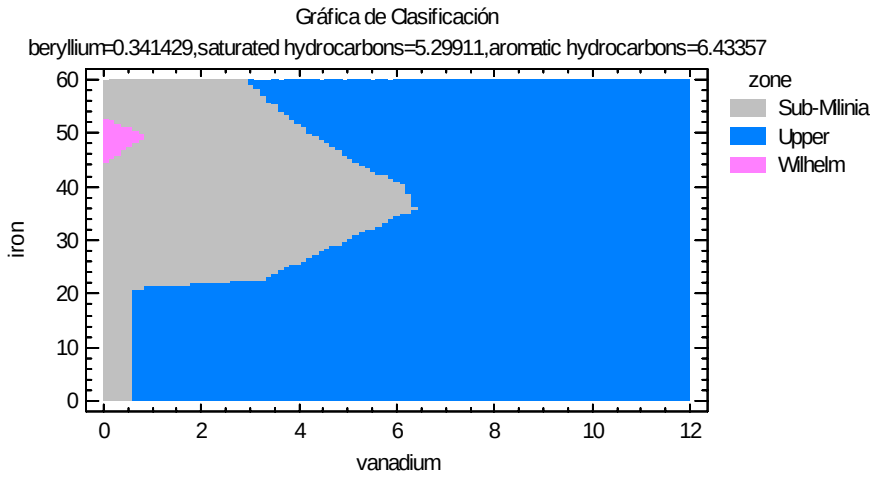
Opciones de Ventana



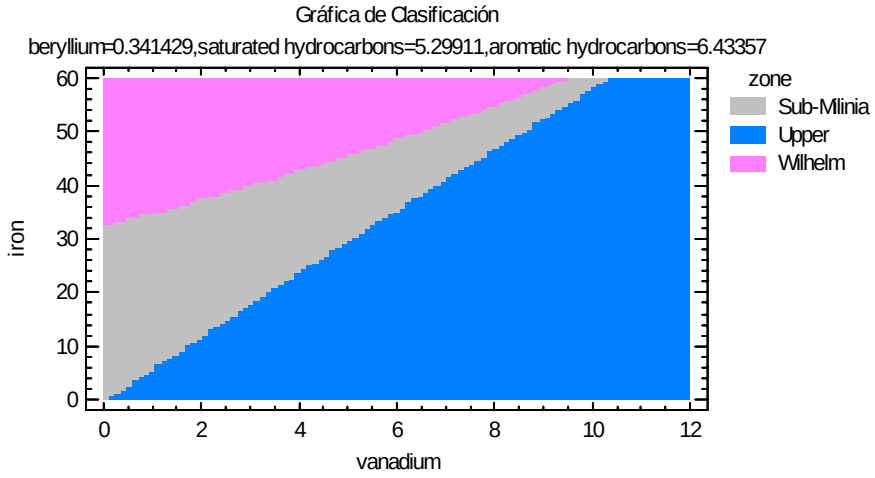
Seleccione variables para definir los tres ejes.

Gráfico de Clasificación

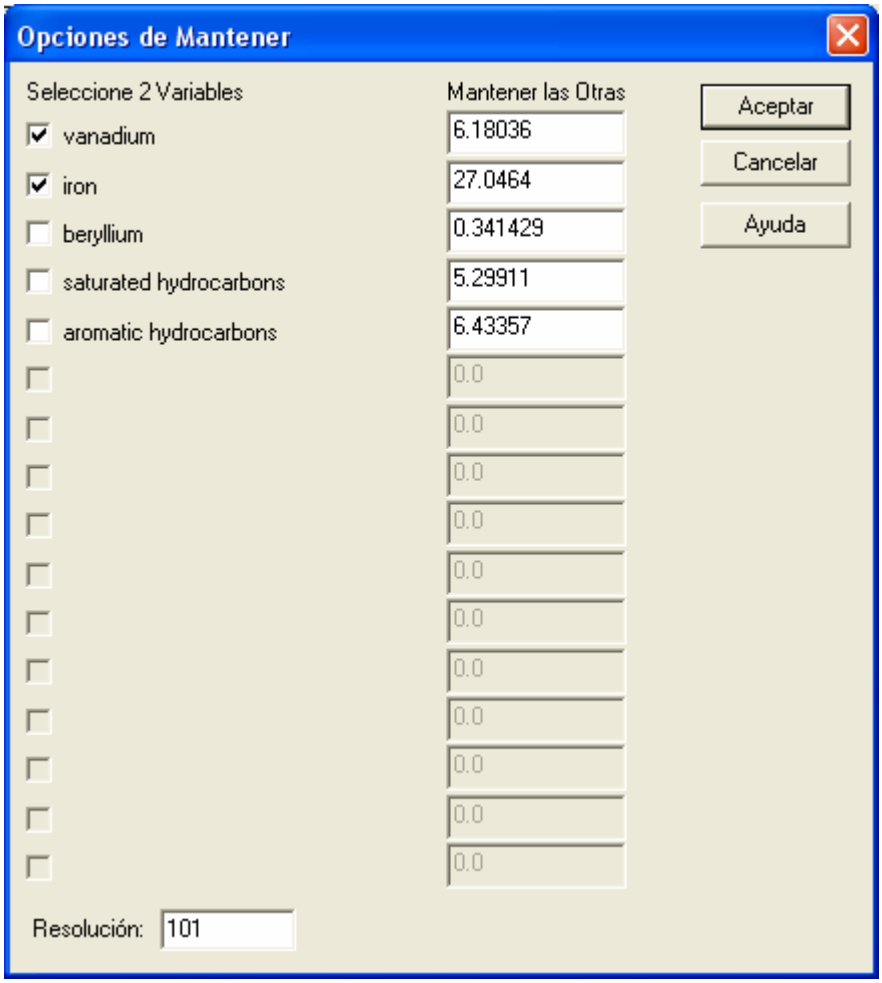
El *Gráfico de Clasificación* puede usarse para entender mejor cómo la región definida por las variables *X* se divide en áreas que hacen que las muestras se clasifiquen como pertenecientes a diferentes grupos.



Cada región de corresponde a un grupo diferente. Dos de las variables *X* se usan para definir los ejes horizontal y vertical, mientras las otras variables se mantienen en valores fijos. Advierta la naturaleza irregular de la gráfica anterior, que corresponde al uso del método del vecino más cercano. La gráfica a continuación, creada usando $\sigma = 1$, es mucho más suave.



Opciones de Ventana



- **Seleccione 2 Variables**: las variables a graficar en los ejes horizontal y vertical.

- **Mantener las Otras**: valores en los que se fijarán las variables no seleccionadas.

- **Resolución**: número de posiciones a lo largo de los ejes horizontal y vertical en las cuales se evalúa el algoritmo de clasificación. Una mayor resolución dará un gráfico más suave pero aumentará el tiempo requerido para generarlo.