COMPONENTES PRINCIPALES SUPERVISADOS PARA CLASIFICACIÓN DE DATOS DE EXPRESIÓN GENÉTICA

por:

Jaime Carlos Porras Cerrón

Tesis sometida en cumplimiento parcial de los requisitos para el grado de

MAESTRO EN CIENCIAS
en
MATEMATICA
(Estadística)

UNIVERSIDAD DE PUERTO RICO
MAYAGÜEZ CAMPUS
2005

Aprobada por:

_____	_____
Julio Quintana, Ph.D. Miembro, Comité Graduado	Fecha
_____	_____
Edgardo Lorenzo, Ph.D. Miembro, Comité Graduado	Fecha
_____	_____
Edgar Acuña, Ph.D. Presidente, Comité Graduado	Fecha
_____	_____
Gladys Ducoudray, Ph.D. Representante de Estudios Graduados	Fecha
_____	_____
Pedro Vásquez, D.Sc. Director de Departamento	Fecha

ABSTRACT

The gene expression data obtained through the technology of *microarrays* are characterized by its considerably greater amount of features in comparison to the number of observations. The direct use of traditional statistics techniques of supervised classification can give poor results in gene expression data. Therefore before analyzing this type of data is advisable to perform a dimension reduction. The present work combines two types of dimensional reduction techniques: feature selection and feature extraction. In the first step of the proposed method feature selection is applied, and in the second step principal components are formed with the selected features. This technique is called Supervised Principal Components (SPC). Three classifiers are applied to these components and the misclassification error is estimated. Two algorithms of SPC are presented; they essentially, differ in the time in which the feature selection is made. Finally, the results of this new technique are applied to nine gene expression data sets.

ABSTRACTO

Los datos de expresiones genéticas obtenidos a través de la tecnología de microarreglos tienen como característica principal contar con una cantidad considerablemente mayor de variables en comparación al número de observaciones. En estos casos utilizar directamente técnicas estadísticas tradicionales de clasificación supervisada puede brindar resultados poco satisfactorios. Por esta razón es recomendable realizar una reducción de dimensionalidad, antes de analizar este tipo de datos. El presente trabajo combina dos formas de reducción de dimensionalidad: selección de variables y extracción de variables. Como primer paso del método propuesto, se realiza una selección de variables (se usaron diferentes procedimientos) para posteriormente, con las variables seleccionadas, formar Componentes Principales, los cuales son llamados Componentes Principales Supervisados (CPS). A estos componentes se les pueden aplicar distintos clasificadores para obtener finalmente el error de mala clasificación. Se presentan dos algoritmos de CPS, que esencialmente, se diferencian en el momento en que se hace la selección de variables. Finalmente, se muestran los resultados de esta nueva técnica aplicada a nueve conjuntos de datos de expresión genética.

DEDICATORIA

A mis madres Jesús y Gloria, por todo su amor y cariño.

A mi abuelo Epifanio y mi padre Jaime, por sus sabias enseñanzas.

A mis hermanos Al y Eduardo, por su constante apoyo.

A María Beatriz, por su amor, ternura y comprensión.

AGRADECIMIENTOS

A Dios, por su compañía en muchos momentos de soledad y adversidad.

Al Dr. Edgar Acuña Fernández, presidente del Comité Graduado, por su apoyo y constantes sugerencias en el desarrollo de la presente tesis.

A todos los que fueron mis profesores del Departamento de Matemática.

A todos mis verdaderos amigos del Departamento de Matemática; en especial a Luis Daza por animarme a esta aventura.

A Frida Coaquira y Carlos López de Castilla, porque más que amigos, fueron como hermanos durante mi estadía en Mayagüez.

A todas las personas que de una u otra forma hicieron posible la culminación de la presente tesis.

Tabla de Contenido

Lista de Tablas

Tablas **Página**

Lista de Figuras

Figuras **Página**

1 INTRODUCCIÓN

El creciente desarrollo de la ciencia trae consigo la aparición de nuevas disciplinas, y muchas de ellas provenientes de la fusión de algunas ciencias ya existentes. Estas nuevas áreas de estudio hacen uso de equipos sumamente sofisticados, los cuales producen, a su vez, nuevas estructuras de datos.

En ese contexto, una disciplina relativamente nueva, y que ha logrado un amplio desarrollo en la actualidad es la bioinformática. Ésta se dedica a la investigación y desarrollo de herramientas útiles para entender el flujo de información desde los genes hasta sus estructuras moleculares, su función bioquímica, su conducta biológica; finalmente, su influencia en las enfermedades y en la salud. Los datos que se analizan en bioinformática provienen mayormente, por tanto, de las expresiones de genes (o expresión genética), las cuales pueden llegar a ser miles en una sola observación. Existen varias maneras de medir la expresión genética. Una de ellas es la tecnología de los microarreglos (*microarrays*). Ésta permite analizar simultáneamente miles de genes; sin embargo, el costo por observación (o individuo) es muy alto.

Por otro lado, la estadística es una ciencia que analiza información para posteriormente tomar decisiones sobre los resultados obtenidos, por lo tanto no puede ser ajena a estos cambios. Por esa razón constantemente se proponen metodologías para analizar estructuras de información emergentes.

1.1 Justificación

El desarrollo del presente trabajo de tesis se justifica en el marco de la aparición de nuevas estructuras de datos. En efecto, un tipo de estructura de datos relativamente nuevo es el proveniente de la aplicación de la tecnología de microarreglos. Este tipo de estructura de datos conocido comúnmente como *microarray data,* presenta como característica principal que el número de variables (genes) es considerablemente mayor (usualmente miles) en comparación a la cantidad de observaciones analizadas (usualmente menos de 100).

En muchos estudios, la expresión genética es asociada a algún tipo de enfermedad cancerígena; es decir cada conjunto de genes provenientes de un individuo es relacionado a un tipo de cáncer. Más aún, en las últimas investigaciones médicas, se le utiliza para poder diferenciar, lo que se ha denominado como subtipos de cáncer. Desde el punto de vista estadístico, se puede establecer que la medición de los genes provenientes de las expresiones genéticas se puede considerar como variables predictoras; mientras que los tipos, subtipos de cáncer o ausencia de cáncer, codificados adecuadamente, pueden ser utilizados como las clases.

En el análisis de clasificación supervisada, se dispone de un conjunto de observaciones multivariadas, para las cuales se conocen a priori las clases a las que pertenecen. El objetivo principal en este tipo de análisis es estimar el error de mala clasificación para un clasificador dado.

Si se considera que muchas de las técnicas estadísticas tradicionales han sido

diseñadas para analizar un número considerable de observaciones en comparación a la cantidad de variables en estudio; entonces cuando sucede lo contrario, como en el caso de los datos provenientes de expresiones genéticas, se pueden obtener resultados poco satisfactorios.

Si bien es cierto que en la actualidad se han desarrollado varios métodos que trabajan con datos provenientes de expresiones genéticas en clasificación supervisada; muchos de ellos presentan algoritmos complejos, lo que conlleva que el tiempo de procesamiento de la información sea alto. Asimismo, existen algunos algoritmos que son utilizados para datos procedentes de expresión genética que en su aplicación no verifican previamente algunas suposiciones estadísticas requeridas.

Por lo tanto, es este trabajo se desarrollará una nueva técnica diseñada a trabajar principalmente con datos provenientes de expresión genética y que presente un algoritmo simple. Es decir, que utilice un menor tiempo de procesamiento; así mismo que mejore la reducción del error de mala clasificación, cumpliendo con todas las suposiciones estadísticas necesarias.

Finalmente, cabe mencionar que la metodología que se expondrá se aplicó solamente a diversas bases de datos de microarreglos, pero eso no impide que pueda también aplicarse a datos que provengan de otras áreas, como es el caso de Quimiometría (*Chemometrics*), que es un área de química que trabaja con datos de estructura similar a la de los microarreglos.

1.2 Objetivos

Objetivo Principal

Desarrollar una técnica que combine la selección y extracción de variables, y consiga, en un menor tiempo, estimar el error de mala clasificación en datos provenientes de expresión genética.

Objetivos Específicos

- Comparar el tiempo de procesamiento de distintos métodos de selección de variables aplicados en clasificación.
- Determinar el efecto de diferentes métodos de selección de variables (en clasificación supervisada) sobre la formación de componentes principales para la estimación del error de mala clasificación.
- Utilizar los componentes principales supervisados para determinar cual de ellos brinda la menor tasa de error de mala clasificación, mediante la utilización de tres clasificadores (Rpart, KNN y Regresión Logística Nominal).
- Analizar la tendencia del número de componentes principales supervisados a utilizar para obtener la menor tasa de error de mala clasificación.
- Desarrollar programas en lenguaje R, en el ambiente Windows, basados en los algoritmos propuestos, que permitan realizar las estimaciones de los errores de mala clasificación.

1.3 Resumen de los capítulos siguientes

La presente tesis está estructurada en cinco capítulos. El segundo capítulo está referido, por un lado, a una revisión de los procedimientos de reducción de dimensionalidad: selección de variables y extracción de variables. Se describen, también varios métodos de selección de variables aplicados a datos provenientes de expresión genética. Por otro lado, se brinda información sobre los clasificadores no paramétricos utilizados para estimar el error de mala clasificación; así como diferentes formas para poder estimar este error.

En el tercer capítulo se brinda una descripción detallada de la metodología de la técnica propuesta a la que se ha denominado Componentes Principales Supervisados (CPS), la cual muestra dos variantes. Se brinda un mayor detalle a los métodos de selección de variables utilizados en la presente tesis; así como la descripción de la selección del número de componentes a utilizar para la estimación del error de mala clasificación.

En el cuarto capítulo, se presenta la aplicación de los CPS a nueve conjunto de datos comúnmente utilizadas en este tipo de investigación, los resultados obtenidos, así como sus respectivas discusiones.

En el quinto y último capítulo, se presentan las conclusiones de los resultados obtenidos y recomendaciones para futuros trabajos relacionados a clasificación en datos provenientes de expresiones genéticas.

2 REDUCCIÓN DE DIMENSIONALIDAD

2.1 Introducción

La tecnología de microarreglos permite medir los niveles de expresión de miles de genes simultáneamente contenidos en una lámina o dispositivo. La lámina constituye un arreglo de secuencias de fragmentos de DNA inmovilizadas, ordenadas y que están adheridas a una superficie generalmente de cristal. Cada secuencia corresponde a un gen diferente.

Desde el punto de vista estadístico un gen sería una variable y una lámina sería una observación. En estudios de cáncer, cada lámina mide los niveles de expresión genética de pacientes con diferentes tipos de cáncer. En todas las aplicaciones donde se use tecnología de microarreglos; el número de variables (genes) p es mucho mayor que el número de observaciones n; un estudio típico incluye de 1000 a 20000 genes para sólo 10 a 100 láminas.

Diversos métodos de clasificación son extensamente utilizados para diagnosticar tipos de cáncer a través de los genes. La clasificación supervisada es la aplicación más importante en datos provenientes de tecnología de microarreglos. En este caso, muchos autores creen que este tipo de datos requiere del desarrollo de nuevos y sofisticados métodos estadísticos; sin embargo otros reportan que se han obtenido buenos resultados cuando se realizan tratamientos previos de los datos antes de aplicar los métodos estadísticos tradicionales. Uno de ellos es la reducción de dimensionalidad, es decir aplicar técnicas estadísticas que permitan: seleccionar variables, realizar transformaciones de las variables o una

combinación de los dos criterios antes mencionados, antes de aplicar un clasificador.

Los métodos de reducción de dimensionalidad se dividen en selección de variables y extracción de variables. Dentro de cada una de estas formas de reducción de dimensionalidad no existe una subdivisión definida. En los trabajos de Kohavi y John (1997), Guyon et al. (2003) y Boulesteix (2004) se presenta una subdivisión de los métodos de selección de variables; en el caso de los métodos de extracción de variables, Boulesteix (2004), brinda una interesante clasificación, la cual será presentada posteriormente.

Básicamente, nosotros utilizaremos tres métodos de selección de variables: la prueba de Kruskal-Wallis, RFE (*Recursive Feature Elimination*) y Relief. A éstos, los combinaremos con el método de extracción de variables de Componentes Principales.

La literatura menciona que los Componentes Principales no necesariamente mejoran la predicción cuando se usa en regresión o en clasificación. En clasificación no supervisada, trabajos como el de Yeung y Ruzzo (2001) demuestran que el uso de componentes principales en lugar de las variables predictoras originales, no necesariamente mejora y en muchos casos degrada la calidad esperada de clasificación. Una manera de mejorar la estimación del error de mala clasificación es realizando un pre-procesamiento al conjunto de datos mediante selección de variables, es decir eliminando variables redundantes e irrelevantes. Las variables seleccionadas servirán para formar los componentes, por lo que se espera que esto mejore su poder predictivo.

Un buen ejemplo que demostraría este punto sería, hacer un Análisis de Componentes Principales (ACP) incluyendo todas las variables predictoras, y luego hacer PCA incluyendo solo las variables relevantes y mostrar cómo afecta en la estimación de los errores de mala clasificación.

2.2 Selección de Variables

El primer criterio de reducción de dimensionalidad consiste en seleccionar un subconjunto de genes a los cuales posteriormente se le aplicará un método de clasificación.

En la literatura de microarreglos, este criterio es frecuentemente conocido como: selección de genes (*gene selection*), o selección de subconjuntos (*subset selection*). Distintas publicaciones mencionan que después de la selección de genes, se aplican métodos clásicos de clasificación como: análisis discriminante lineal, análisis discriminante cuadrático, discriminante lineal de Fisher, clasificación por vecinos más cercanos (Dudoit et al., 2002), redes neurales artificiales (Kahn et al., 2001), *Support Vector Machin*es (Furey et al., 2000). Algunos métodos clásicos de clasificación, como clasificación por vecinos más cercanos (*K Nearest Neighborhood*, *KNN*), no requiere necesariamente que $n > p$, pero su uso directo cuando el número de variables es mucho más grande que el número de observaciones, (como es el caso de datos provenientes de microarreglos), hace que se obtenga un pobre nivel de correcta clasificación. Otros métodos no pueden ser aplicados si $n < p$: como el análisis discriminante

clásico, dado que las matrices de variancia-covariancia tienen que ser invertidas para estimar la función de discriminación; esto no puede ser realizado cuando n es pequeño en comparación a un valor grande de p.

Adicionalmente, selección de variables es aplicada a datos provenientes de microarreglos por las siguientes razones:

- Puede ser utilizado como un paso preliminar de clasificación, porque la mayoría de los métodos de clasificación trabajan sólo con un pequeño subconjunto de variables.
- Es de crucial interés para biólogos, identificar los genes que estén asociados a una enfermedad específica. A estos genes los denominan genes marcadores (*marker gene*).
- Se pueden eliminar variables (genes) redundantes e irrelevantes que no contribuyen significativamente a la clasificación de nuevas observaciones.
- Seleccionar un subconjunto de variables implica un menor costo computacional, en cuanto al tiempo de procesamiento y por lo tanto de obtención de los resultados.

Como ya se mencionó anteriormente, existen muchas formas de clasificar los métodos de selección de variables. Una forma sencilla consiste en dividirlos en tres grupos: Métodos de Filtro (*Filters Methods*), Métodos de Envoltura (*Wrapper Methods*) y Métodos de Encaje (*Embedded Methods*).

2.2.1 Métodos de Filtro (Filters Methods)

Estos tipos de métodos de selección de variables no se basan en ningún clasificador para realizar la selección de variables. Cada variable se toma individualmente y se calcula una medida de puntuación para utilizarla posteriormente como indicador discriminatorio de las variables. Éstas son *rankeadas* de acuerdo a su puntuación; por lo que se puede escoger sólo las $\tilde{p}$ mejor *rankeadas* (donde $\tilde{p} < p$) ó las variables cuya puntuación exceda un valor preestablecido.

Una de las puntuaciones más utilizadas es la prueba F de Fisher; que para cada variable j, se define de la siguiente manera:

$$F_j = \frac{\sum_{k=1}^{K} \sum_{i:Y_i=k} (\hat{\mu}_{kj} - \hat{\mu}_j)^2 / (K-1)}{\sum_{k=1}^{K} \sum_{i:Y_i=k} (x_{ij} - \hat{\mu}_{kj})^2 / (n-K)} \qquad \begin{matrix} j = 1,2,\ldots,p \\ k = 1,2,\ldots,K \end{matrix} \qquad 2.1$$

donde K es el número de clases.

En esta fórmula se supone que:

- X_j está normalmente distribuido en cada clase k, con media μ_{kj} y variancia constante σ^2.
- Las observaciones x_{ij}, $i = 1,\ldots,n$ son independientes.

Para la j-ésima variable se desea probar:

$H_0 : \mu_{1j} = \cdots = \mu_{Kj}$

versus

$H_1 : \mu_{1j} \neq \mu_{kj}$ para al menos un k,

Bajo H_0, F_j tiene una distribución F de Fisher con K-1 y $n-K$ grados de libertad. El valor descriptivo de la prueba (*p-value*) o la prueba estadística correspondiente puede usarse como puntuación para seleccionar las variables relevantes. El estadístico *T*, también es un criterio de selección muy utilizado cuando las observaciones están clasificadas solamente en dos clases, (éste es un caso especial del estadístico *F*).

Existen variantes de la prueba anterior. Una de las primeras es la propuesta por Golub et al. (1999), que se usa para dos clases. En este caso, el criterio está definido por:

$$\rho_j = \left| \frac{\bar{x}_j^{(1)} - \bar{x}_j^{(2)}}{\sigma_j^{(1)} + \sigma_j^{(2)}} \right| \qquad 2.2$$

donde $\bar{x}_j^{(k)}$ y $\sigma_j^{(k)}$ indican el promedio y la desviación estándar de cada variable para todas las observaciones pertenecientes a la clase k.

La otra modificación de la prueba *F*, muy utilizada para este tipo de datos, es la razón BSS/WSS, propuesta por Dudoit et. al. (2002); en la cual la selección de variables se basa en la razón de las sumas de cuadrados entre grupos (*Between Sum Square* - *BSS*) y dentro de grupos (*Within Sum Square – WSS*). Para la variable (gen) j, la razón estaría dada por:

$$\frac{BSS(j)}{WSS(j)} = \frac{\sum_i \sum_k I(y_i = k)(\bar{x}_{kj} - \bar{x}_{.j})^2}{\sum_i \sum_k I(y_i = k)(x_{kj} - \bar{x}_{kj})^2} \qquad 2.3$$

donde $\bar{x}_{.j}$ denota el nivel promedio de la expresión genética j a través de todas las muestras y $\bar{x}_{kj}$ denota el nivel promedio del gen j en toda la muestra que

pertenece a la clase k.

Por otro lado, se sabe que los datos provenientes de microarreglos contienen muchos valores atipicos (*outliers)* y pocas observaciones. Algunos autores (como Dettling y Bühlmann (2002)) prefieren usar estadísticos más robustos como la prueba no paramétrica de ordenamiento de Wilcoxon para el caso de dos clases $(K = 2)$. Para cada variable j, sólo se necesita el supuesto de que las observaciones $x_{ij},\ldots,x_{nj}$ sean independientes. Si $rank(x_{ij})$ denota el rango de x_{ij} en la sucesión $x_{ij},\ldots,x_{nj}$, la prueba estadística para la variable j está dada por:

$$W_j = \sum_{i:Y_i=1} rank(x_{ij}) \qquad 2.4$$

y se utiliza para probar la hipótesis

$$H_0 : mediana\left(X_j \middle| Y = 1\right) = mediana\left(X_j \middle| Y = 2\right)$$

versus

$$H_1 : mediana\left(X_j \middle| Y = 1\right) \neq mediana\left(X_j \middle| Y = 2\right)$$

Bajo H_0, W_j tiene una distribución de Wilcoxon con grados de libertad n_1 y n_2. El valor descriptivo de la prueba (*p-value*) o la prueba estadística correspondiente para cada variable j puede ser usado como una medida de relevancia.

Otra prueba no paramétrica, cuando sólo hay dos clases, es la propuesta por Park et. al (2001), la cual brinda resultados muy similares a los de la prueba de Wilcoxon. Dettling (2004) utilizó esta prueba como método de preselección previo al uso del método LogitBoost.

La prueba de Park consiste en asignar el valor 0 a las n_1 observaciones

pertenecientes al primer grupo y 1 a los n_2 observaciones pertenecientes al segundo grupo. Para cada variable, de manera independiente, se ordenan los valores de las observaciones (en forma ascendente), con lo que se permutarían los 0's y 1's. En cada columna de manera independiente se calcula una puntuación consistente en el número de desplazamientos que se tienen que realizar para tener todas las observaciones pertenecientes al primer grupo en la parte superior de la columna y todas las observaciones pertenecientes al segundo grupo en la parte inferior de la columna. Esta puntuación se expresa por:

$$Puntuación = \sum_{i \in N_2} \sum_{j \in N_1} h\left(x_j - x_i\right) \qquad 2.5$$

donde N_i representa el conjunto de índices pertenecientes al grupo i y $h\left(x\right)$ es la función indicadora

$$h\left(x\right) = \begin{cases} 0, & si\ x \le 0 \\ 1, & si\ x > 0 \end{cases} \qquad 2.6$$

Sin embargo en muchos conjuntos de datos la variable clasificadora presenta más de dos clases $\left(K > 2\right)$. En estos casos, para cada variable (evaluada de manera independiente) se puede hacer uso de otras pruebas no paramétricas como es el caso de la prueba de la Mediana (*Mood Median Test*) ó la prueba de Kruskal Wallis.

La prueba de la Mediana consiste en elaborar una tabla de contingencia considerando dos columna para cada clase K, en la primera columna se contabiliza el número de observaciones que sean menores ó iguales a la mediana

de la clase K y en la otra columna se contabiliza el número de observaciones mayores a la mediana de la clase K. Posteriormente a esta tabla de contingencia se le aplica una prueba X^2.

La prueba de Kruskal-Wallis, es una generalización de la prueba estadística de rangos de Wilcoxon. Esta prueba fue usada en esta tesis y será discutida en el siguiente capítulo.

Los métodos de filtro de selección de variables antes mencionados, son considerados como métodos de *rankeo* univariado y son criticados debido a que en algunos casos el subconjunto de las variables mejor *rankeadas* no es el mejor subconjunto en términos de exactitud de clasificación porque:

- Las variables con mejor posición de acuerdo al ordenamiento podrían estar fuertemente correlacionadas, lo que implica que se consideraría información redundante.
- No se considera la interacción entre variables cuando se usa métodos de *rankeo* univariado y dos variables que pudieran tener una puntuación relevante individual baja podrían separar mejor las diferentes clases cuando son consideradas juntas.

Algunos Métodos de Filtro y los Métodos de Envoltura (que se presentarán en la siguiente sección) tratan de superar los problemas antes mencionados mediante:

- El uso de las puntuaciones relevantes para evaluar los subconjuntos de variables.
- El uso de un algoritmo que permita explorar el espacio de los posibles subconjuntos.

Algunos de los métodos de filtro que realizan las acciones mencionadas anteriormente son:

- El Método "Las Vegas Filter" (LVF) - (Liu y Setiono - 1997):

  La selección del subconjunto de variables se hace de manera aleatoria y la función evaluadora que se usa es una medida de inconsistencia la cual es aplicada a variables categóricas. Por lo tanto, si el conjunto de datos posee variables continuas, éstas deben ser discretizadas previamente antes de aplicar el método LVF.

  El método Relief también es un Método de Filtro, este fue utilizado en la presente tesis y será discutido en el siguiente capítulo.

2.2.2 Métodos de Envoltura (Wrapper Methods)

Los métodos "*Wrapper*", realizan la selección de variables usando como criterio de evaluación las estimaciones del error de clasificación basadas en algún clasificador. Entre algunos de los métodos de envoltura se tienen:

- El Método de Selección secuencial hacia adelante (*SFS*)

  Es un método heurístico de selección de variables que evalúa la contribución de una variable en la clasificación de acuerdo a las variables previamente seleccionadas. El proceso comienza con un conjunto vacío de variables. En el primer paso, se realiza la clasificación con cada una de las variables, se selecciona aquella variable que haya realizado mejor la clasificación. En el segundo paso, se prueba con subconjuntos de tamaño dos (la primera variable seleccionada más cada una de las otras) y se selecciona el par de variables que haya producido la

tasa de clasificación correcta más alta, y asimismo se prueba con los grupos de tres variables para seleccionar la tercera variable. El proceso continúa hasta que al incrementarse el conjunto de variables ya seleccionadas con cada una de las variables restantes no produce incremento en la tasa de clasificación correcta.

- Método de selección secuencial flotante hacia adelante (*SFFS*)

  Este método fue introducido por Pudil et al. en 1994 y es muy similar al método "*Stepwise*" de selección de variables en regresión. En cada paso se incluye una nueva variable por medio de un procedimiento secuencial hacia adelante, pero luego se realiza la exclusión de las variables menos significativas, una por una, hasta que la tasa de error de clasificación correcta disminuya. Una vez que ya no se puede seguir excluyendo variables se hace otro paso hacia delante para incluir otra variable y nuevamente se realiza la exclusión de variables, si es posible. El proceso recurrente termina cuando ya no se pueden efectuar más pasos hacia adelante porque la tasa de clasificación correcta ya no se incrementa.

RFE (*Recursive Feature Elimination)*, también es de método de envoltura y se discutirá más extensamente en el siguiente capítulo, debido a que es uno los métodos de selección de variables que se utilizó en la presente tesis.

Las definiciones de los métodos antes presentados corresponden a Coaquira (2002). En su propuesta se desarrollaron comparaciones de estos dos tipos de métodos de selección óptima de subconjuntos para clasificadores basados en

estimadores de densidad por Kernel.

2.2.3 Métodos de Encaje (Embedded Methods)

Estos métodos tienen como característica principal que realizan la selección de variables durante el proceso de clasificación. Se dividen en dos grupos:

- Métodos de Subconjuntos Anidados

  Algunos métodos de encaje guían su búsqueda en la estimación de cambios en el error de mala clasificación cuando se hacen movimientos en el espacio de un subconjunto de variables. Combinando lo anterior con estrategias de búsqueda (*backward elimination* o *forward elimination*) se producen los subconjuntos de variables anidadas. Algunos de estos métodos son: CART, *neural networks*, *The Gram-Schmidt orthogonolización* y OBD ("*Optimum Brain Damage*").

- Optimización de Objetivo Directo

  El objetivo de este método es satisfacer dos condiciones de manera simultánea:

  La primera es que la estimación del clasificador sea la mejor posible; es decir que la tasa de clasificación correcta sea máxima, y la segunda es optimizar el número de variables, es decir utilizar el mínimo número de variables en la estimación del error de clasificación. Un método de este tipo es el propuesto por Weston et al. (2000).

2.3 Extracción de Variables

La selección de variables es un criterio de reducción de dimensionalidad muy utilizado en análisis de datos provenientes de microarreglos. Sin embargo, presenta dos desventajas. Primero, una gran parte de la información contenida en los datos llega a perderse, dado que muchas variables se eliminan por los procedimientos de selección. Segundo, las interacciones y las correlaciones entre variables, casi siempre son ignoradas.

Los métodos de extracción de variables son una alternativa a los métodos de selección de variables cuando se trabaja con datos de gran dimensión.

Los métodos de extracción de variables presentan ciertas ventajas sobre los métodos de selección de variables. Entre ellas se pueden mencionar las siguientes:

- Permiten la visualización de los datos dado que se trabaja en un espacio de baja dimensión.
- Se incorporan interacciones y correlaciones entre variables.
- Aunque se usa la información de miles de variables (genes) y dado que la dimensión se reduce a unas pocas variables transformadas, se pueden emplear diversos métodos estadísticos que pueden trabajar con pocas variables.
- En un caso ideal, los nuevos componentes pueden interpretarse en aplicaciones científicas.

Los métodos de extracción de variables se pueden usar para diferentes propósitos, como por ejemplo: para hallar conglomerados, en regresión, en clasificación supervisada. El enfoque en esta tesis será en clasificación supervisada.

A continuación presentamos la clasificación de los métodos de extracción de variables propuesta por Boulesteix (2004).

2.3.1 Métodos Lineales y No Lineales de extracción de variables

Los métodos de extracción de variables pueden ser transformaciones lineales o no lineales de las variables originales. Estos datos transformados reciben usualmente el nombre de componentes o factores. Los métodos lineales son usualmente más rápidos, robustos y mejor interpretables que los métodos no lineales. En contraste, los métodos no lineales pueden, algunas veces, descubrir estructuras complicadas que los métodos lineales no detectan.

Los métodos no lineales para la reducción de dimensión como: *Isomap* o *Sammon's non linear mapping*, son computacionalmente muy intensivos para datos de gran dimensión; más aún, se sabe que ellos no brindan buenos resultados cuando el número de observaciones es pequeño, como es el caso de datos provenientes de microarreglos.

2.3.2 Métodos Supervisados y No Supervisados de Extracción de Variables

Los métodos supervisados de extracción de variables se caracterizan por usar la información de la clase Y para construir los nuevos componentes, contrariamente a los métodos no supervisados. Se sabe que métodos no supervisados de reducción de dimensionalidad son herramientas útiles para la representación gráfica para descubrir estructuras de datos.

La clasificación de los métodos de extracción de variables presentada anteriormente no es excluyente. Es decir, existen métodos que pertenecen a ambos grupos. Por ejemplo: El Análisis de Componentes Principales (ACP) es un método lineal y no supervisado; el método de Mínimos Cuadrados Parciales (PLS) es lineal y supervisado.

2.4 Clasificadores

Para estimar los errores de mala clasificación se utilizaron los tres clasificadores siguientes: Rpart, KNN y Regresión Logística Nominal.

2.4.1 Rpart

El uso inicial de diagramas de árboles en estadística estuvo a cargo de Breiman, Friedman, Olsen y Stone (1984) quienes introdujeron nuevos algoritmos para su construcción y los aplicaron a problemas de regresión y clasificación. Este método es conocido como CART (*Classification and Regression Trees*).

El término árboles o árbol binario es por la gráfica, formada por nodos y arcos los cuales son mostrados creciendo de arriba hacia abajo. La raíz es el nodo superior, en cada nodo se hace una partición hasta llegar a un nodo terminal u hoja.

La metodología a seguir para construir un árbol binario resulta de conjugar varios elementos:

- Un criterio para evaluar la ventaja derivada de la división de un nodo.
- Una especificación del espacio de búsqueda.
- La forma de estimar la tasa de mala clasificación.

- Un criterio para decidir cuando detener el crecimiento del árbol, (o como poder podar el árbol cuando ha crecido en exceso).
- Un criterio para asignar un valor a cada hoja.

El algoritmo de construcción de un árbol implica la estimación de una medida de "impureza". Siguiendo la notación de Breiman et al. se denotará la impureza del nodo t por $i(t)$. En el caso de árboles de clasificación, (en los que la variable respuesta es de tipo cualitativo), la impureza de un nodo debería estar relacionada con las proporciones en que se presentan los elementos de las diferentes clases. Si la variable Y puede tomar K valores, sea $p(k \mid t)$ la proporción de elementos de la clase k en la muestra de entrenamiento que han ido al nodo t. Se desea que $i(t)$ sea mínima si:

$$\begin{aligned} &p(\ell \mid t) = 1 \\ &p(k \mid t) = 0 \qquad\qquad \forall k \neq \ell \end{aligned} \tag{2.7}$$

Ello, en efecto, correspondería a un nodo "puro" y todos los elementos que van a él son de la clase ℓ. Por el contrario, desearíamos que la función $i(t)$ fuera máxima cuando

$$p(k \mid t) = K^{-1} \qquad\qquad \forall k \tag{2.8}$$

pues un nodo en el que todas las clases aparecen equi-representadas es en cierto sentido máximamente impuro.

Hay varias elecciones de $i(t)$ que satisfacen las propiedades anteriores y otras más que también son deseables. Tenemos así la función entropía dada por:

$$i(t) = -\sum_{k=1}^{K} p(k \mid t) \log p(k \mid t) \qquad 2.9$$

y el índice de Gini

$$i(t) = \sum_{i \neq k} p(i \mid t)\, p(k \mid t) \qquad 2.10$$

Lo que se desea es valorar la ganancia en términos de impureza de una división del nodo t. Una posibilidad intuitivamente atractiva es:

$$\Delta(s,t) = i(t) - p_L i(t_L) - p_R i(t_R) \qquad 2.11$$

en que la mejora en términos de impureza resultante de elegir la división s del nodo t se evalúa como la diferencia entre la impureza de dicho nodo y la de sus hijos, t_L y t_R, ponderadas por las respectivas proporciones p_L y p_R de elementos de la muestra que la división s hace ir a cada uno de ellos.

La impureza total $I(T)$ de un árbol T se define como la suma ponderada de impurezas de sus hojas. Si $\tilde{T}$ es el conjunto formado por las hojas de T, entonces

$$I(T) = \sum_{t \in \tilde{T}} p(t) i(t) \qquad 2.12$$

2.4.2 KNN

El método K-NN (*K Nearest Neighborhood*) propuesto por Fix y Hodges (1951) es un método no paramétrico de estimación de la función de densidad. Sea $x_1, x_2, \ldots, x_n$ una muestra con función de densidad desconocida $f(x)$. Se estima $f(x)$ a partir de una celda de centro en x, que crece hasta contener k elementos, donde el valor de k se define arbitrariamente o como una función de n. Estas

observaciones son los k vecinos más cercanos a x. Se tiene entonces:

$$\hat{f}(x) = \frac{k/n}{V_k(x)} \qquad 2.13$$

donde $V_k(x)$ es el volumen de un elipsoide centrado en x y de radio la distancia de x al k-ésimo vecino más cercano.

En el caso de clasificación, si la función de clase condicional $F_{y|k}(x)$ es estimada por 2.13, para cada una de las clases la regla de decisión adopta un aspecto más simple para clasificar a la clase C_i tal que:

$$\frac{k_i P_i / n_i}{V_k(x)} > \frac{k_j P_j / n_j}{V_k(x)} \qquad \forall\ j \neq i \qquad 2.14$$

Usualmente las probabilidades a priori P_i son estimadas como la proporción de n_i/n, por tanto (2.14) se reduce a

$$k_i > k_j \quad \forall\ j \neq i \qquad 2.15$$

Luego el proceso de clasificación sería de la siguiente manera:

a) Hallar los k objetos que están a una distancia más cercana a x, k es usualmente un número impar.
b) Si la mayoría de esos k objetos pertenecen a la clase C_i entonces el objeto considerado también pertenece a ella. En caso de empate se clasifica al azar.

2.4.3 Regresión Logística Nominal

En regresión logística, cada fila de la matriz de variables predictoras corresponde a las observaciones del vector p-dimensional $x = (x_1\ x_2 \ldots x_p)^T$. Las entradas del

vector de respuesta Y, corresponden a la observación de la variable y, la cual representa una categoría codificada dentro del conjunto $\{1,2,\ldots,K\}$, que se llamará grupo o clase para efectos de clasificación supervisada. En nuestro caso esta variable es de tipo nominal; debido a que no hay un orden natural en las categorías de la variable respuesta. Aquí una categoría es elegida arbitrariamente como la categoría de referencia. Supongamos que ésta es la primera categoría, entonces la probabilidad de clasificar una observación en una de las K clases se obtiene del modelo:

$$\log\left(\frac{P(y=k)}{p(y=1)}\right) = c_k + \beta_{1k}x_1 + \beta_{2k}x_2 + \ldots + \beta_{pk}x_p \quad k = 2,3,\ldots,K \qquad 2.16$$

Después de estimar los parámetros de la regresión logística se puede hacer la predicción de una observación $x = (x_1\ x_2 \ldots x_p)^T$, la cual consiste en la clasificación de dicha observación en una de las K clases. Para lograr este objetivo se estiman las probabilidades de pertenecer a cada una de las K clases y se aplica la siguiente regla:

$$x \in clase\ k^* \Leftrightarrow k^* = \arg\max_k P(y=k) \qquad 2.17$$

2.5 Estimación del error de mala clasificación

El error de mala clasificación (EMC), error verdadero o error real se define como la probabilidad de que la regla de clasificación (o clasificador) clasifique incorrectamente una observación de prueba (observación que pertenece a la

muestra en la cual se evaluará el clasificador).

El EMC es un buen indicador para la elección de un buen método de clasificación. Se dice que un método de clasificación *A* es mejor que otro método de clasificación *B*, si el error de mala clasificación que se obtiene mediante el método *A* es inferior al error de mala clasificación obtenido con el método *B*. Sin embargo, el error de mala clasificación puede obtenerse a través de diferentes criterios, los cuales se encuentran basados en el uso de la muestra inicialmente recolectada. Algunos de esos criterios se presentan a continuación:

Criterio 1: Es conocido como el criterio de estimación de la tasa de error por resustitución o error aparente (Smith, 1947); aquí se usa todo el conjunto de datos $(x_i, Y_i)_{i=1,\ldots,n}$ para construir la función de decisión d. La clase a la que pertenecen las observaciones de $(x_i, Y_i)_{i=1,\ldots,n}$ son predichas usando d.

En otras palabras, se utiliza la muestra inicialmente recolectada para obtener la función de decisión y se aplica esa función en la misma muestra para estimar el error de mala clasificación.

Una crítica a este criterio es que cuando se usa un método complejo, éste podría construir una función de decisión la cual clasifique correctamente un conjunto de datos; sin embargo, tal función de decisión podría no ser útil para aplicarlo a nueva información.

Este criterio no es muy recomendado, pues subestima el error de mala clasificación, favorece más a métodos complejos y puede conducir a falsas conclusiones si el tamaño de la muestra no es muy grande al compararse con el número de variables utilizadas en el clasificador.

Criterio 2: Una alternativa consiste en dividir el conjunto de datos $(x_i, Y_i)_{i=1,..,n}$ en dos conjuntos de datos excluyentes conocidos como: muestra de entrenamiento o aprendizaje (*training o learning data set*) $\mathcal{L}$ y muestra de prueba (*test data set*) $\mathcal{T}$. $\mathcal{L}$ es usado para construir la función de decisión d, la cual se aplica en $\mathcal{T}$. Es decir, se construye la función de decisión en una parte de la muestra inicial y se aplica la función en la parte restante de la muestra para estimar el error de mala clasificación.

La partición de la muestra en muestra de entrenamiento y muestra de prueba se fija, algunas veces, por razones experimentales. Por ejemplo: las muestras fueron realizadas en dos diferentes momentos (t_1 *y* t_2) ó lugares (l_1 *y* l_2) (donde las observaciones correspondientes a t_1 (l_1) pertenecerían a la muestra de entrenamiento y las observaciones de t_2 (l_2) pertenecerían a la muestra de prueba). Si esto no ocurriese es generalmente mejor dividir la muestra original de manera aleatoria. El principal inconveniente de este criterio es que es muy sensitivo a los cambios en la selección de la muestra de entrenamiento y muestra de prueba.

Criterio 3: Una mejor opción es repetir r veces el Criterio 2 y calcular el promedio de los errores de mala clasificación obtenidos de manera independiente. El incremento de r hace decrecer la variancia de la media; sin embargo r, debería ser lo más grande técnicamente posible. Escoger la razón entre el tamaño de la muestra de entrenamiento con respecto al tamaño de toda la muestra inicial es importante. Los valores comúnmente utilizados para esta razón son: 2/3, 7/10 y 9/10. Reducir esa razón generalmente incrementa el valor de razón medio, debido

a que las reglas de decisión son construidas utilizando menos observaciones. El incremento de esta razón aumenta la correlación entre los errores estimados obtenidos con las *N* particiones.

Criterio 4: Validación cruzada (Stone, 1974) es una opción muy conocida; que consiste en dividir la muestra $(x_i, Y_i)_{i=1,\ldots,n}$ en f subconjuntos excluyentes de (aproximadamente) igual tamaño $S_1,\ldots,S_i,\ldots,S_f$. Para cada subconjunto S_i el siguiente procedimiento es repetido. S_i es considerada la muestra de prueba. La muestra de entrenamiento está formado por las f-1 muestras restantes $S_1,\ldots,S_{i-1},S_{i+1},\ldots,S_f$ que es usada para construir la función de decisión d. Las clases de las observaciones de S_i son predecidas utilizando d. Después este procedimiento es repetido para $i = 1,\ldots,f$. Dos valores comunes utilizados para f son $f = 10$ y $f = n$. Cuando esto último sucede el procedimiento es llamado validación cruzada dejando uno afuera (*leave-out- cross validation*), el cual fue planteado por Lanchenbruch, (1965).

Comparado con otros métodos de estimación de errores, este método produce un estimador con poco sesgo, pero con mucha variabilidad.

El tercer criterio podría preferirse al de validación cruzada para muestras pequeñas como son las que provienen de expresiones genéticas. Un análisis riguroso fue realizado por Braga-Neto y Dougherty (2004).

3 COMPONENTES PRINCIPALES SUPERVISADOS

3.1 Introducción

La base principal para este trabajo de tesis es la publicación "Predicción mediante Componentes Principales Supervisados" cuyos autores son: Bair, Hastie, Paul y Tibshirani (2004). Ellos desarrollaron una técnica similar a la de la presente tesis, pero su aplicación se realizó en análisis de regresión y análisis de supervivencia. Aquí se utilizó en el análisis de clasificación supervisada. En su técnica, ellos realizan una selección de variables para posteriormente aplicar a esas variables seleccionadas el Análisis de Componentes Principales (A.C.P.); con lo que se formaría lo que denominaron como Componentes Principales Supervisados (C.P.S.); los cuales se utilizaron para estimar el modelo de regresión o supervivencia. En el caso de clasificación supervisada, se les aplicará a los C.P.S. un clasificador para obtener finalmente el error de mala clasificación.

La diferencia fundamental del trabajo realizado por los autores antes mencionados, con respecto al nuestro, es la forma de selección de variables. Ellos proponen que las variables seleccionadas (tomadas una a una de manera independiente) sean aquellas cuyo coeficiente de regresión sea superior a un valor límite óptimo, el cual se estima utilizando un tipo de validación cruzada, explicado detalladamente por Tibshirani y Efron (2002).

Nuestra selección de variables se basa en el uso de pruebas no paramétricas; Es decir, las variables predictoras que formarán los componentes principales supervisados serán aquellas que indiquen la mayor diferencia significativa de las clases en la variable dependiente.

Las pruebas no paramétricas de selección de variables no requieren la verificación del supuesto de normalidad, aleatoriedad de residuales y homogeneidad de variancia entre las clases, para cada variable predictora evaluada independientemente. Sin embargo existen estudios como el de Nguyen y Rocke (2002), quienes en análisis de clasificación supervisada de datos provenientes de expresión genética; y con el fin de reducir la dimensionalidad, aplican selección de variables con técnicas estadísticas como diferencia de medias y análisis de variancia de una vía (*one way* ANOVA). Estos procedimientos requieren obligatoriamente la verificación de los supuestos antes mencionados; sin embargo los autores en mención sólo asumen el cumplimiento de dichos supuestos.

Considerando el trabajo presentado por Bair, et al. (2004), se puede afirmar que nuestra metodología está referida primero a una selección de variables (*feature selection*); para, posteriormente, realizar una extracción de variables (*feature extraction*) o transformación de las variables seleccionadas mediante el Análisis de Componentes Principales.

Por otro lado, investigaciones como la de Vega (2004) sólo realizan extracción de variables. Él trabajó con componentes estimados mediante PLS (*Partial Least Square*) y regresión logística aplicado a clasificación supervisada, cuando la

variable dependiente es de tipo nominal. Es decir, generalizó los trabajos de Fort y Lambert-Lacroix (2003) y el de Ding y Gentleman (2004) quienes aplicaron componentes estimados mediante PLS a clasificación supervisada a conjuntos de datos de expresión genética donde la variable dependiente sólo presenta dos clases. Boulesteix (2004), en su tesis doctoral también utilizó la técnica de PLS, pero combinado con el clasificador discriminante lineal.

Otra publicación donde se utilizan técnicas más sofisticadas es la presentada por Dettling y Bühlmann (2004); quienes realizan primero selección de variables utilizando el método propuesto por Park et al. (2001). Luego aplican Bagging y Boosting (que son métodos que tiene como características el uso del remuestreo y un costo computacional muy alto) con el clasificador CART.

Finalmente, se han publicado muchas técnicas novedosas para clasificación en datos provenientes de expresión genética, las cuales combinan selección de variables y otros métodos estadísticos tradicionales. Un caso es el método HykGene de Wang et al. (2004) quienes realizan primero selección de variables por diferentes métodos para, luego aplicar un análisis de conglomerados (*Cluster Analysis*).

3.2 Métodos utilizados para obtener los CPS

3.2.1 La Prueba de Kruskal-Wallis

Cada variable x_v es evaluada de manera independiente, donde $v = 1,2,\ldots, p$. En cada una de las x_v los datos consisten de K muestras independientes (correspondiente a las K clases) de tamaños $n_j \quad (j = 1,\ldots,K)$, donde n es el total de las observaciones:

$$n = \sum_{j=1}^{K} n_j \tag{3.1}$$

Se ordenan las observaciones en forma ascendente y se le asigna el rango correspondiente. En el caso de empates, se utiliza la media de los rangos correspondientes.

Si $R(x_{ij})$ es el rango asignado a la observación x_{ij} y R_j la suma de los rangos asignados a la muestra j:

$$R_j = \sum_{i=1}^{n_i} R(x_{ij}) \tag{3.2}$$

Se calcula R_j para cada muestra

El estadístico de prueba está dado por:

$$H = \frac{1}{S^2}\left[\sum_{j=1}^{K}\frac{R_j^2}{n_j} - \frac{n(n+1)^2}{4}\right] \tag{3.3}$$

donde:

$$S^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n_i}\sum_{j=1}^{K} R(x_{ij})^2 - \frac{n(n+1)^2}{4}\right] \tag{3.4}$$

Si no hay empates, S^2 se simplifica a:

$$S^2 = \frac{n(n+1)}{12} \tag{3.5}$$

y el estadístico de prueba se reduce a:

$$H = \frac{12}{n(n+1)} \sum_{j=1}^{K} \frac{R_j^2}{n_j} - 3(n+1) \tag{3.6}$$

Si el número de empates es moderado, la diferencia entre ambas expresiones de H será pequeña.

Los supuestos que requieren esta prueba son:

- Las k muestras son seleccionadas aleatoriamente desde sus respectivas poblaciones, lo que implica independencia de las observaciones dentro de cada muestra.
- Las k muestras son independientes entre sí.
- La escala de medida es al menos ordinal.

3.2.2 El Método RFE (Recursive Feature Elimination)

Este método de selección de variables es descrito por Guyon et al. (2002) y se basa en la eliminación recurrente de variables. En cada paso de los procedimientos de "iteración", un clasificador es usado con todas las variables presentes, un criterio de *rankeo* es calculado para cada variable, y la variable con el criterio de *rankeo* más pequeño es eliminada. Un criterio de *rankeo* que se utiliza comúnmente es:

$$\Delta P_j = \frac{1}{2}\left(\frac{\partial^2 P}{\partial b_j^2}\right) b_j^2 \qquad 3.7$$

donde, P es una función de pérdida calculada en los datos de entrenamiento, y b_j es el coeficiente correspondiente a la variable j en el modelo. ΔP_j aproxima la sensitividad de P para la variable j. Para el clasificador SVM (*Support Vector Machine*) y con la función de pérdida cuadrática media $P = \left\| y - \vec{b}^T \vec{x} \right\|^2$, se tiene que $\Delta P_j = b_j^2 \left\| x_j \right\|^2$ es parecido a la prueba estadística T elevada al cuadrado para la variable j, donde x_j es el n-vector de muestra para la variable j. Suponiendo que los valores de cada variable tienen rangos similares, entonces $\Delta P_j = b_j^2$ se usa con frecuencia. Por razones computacionales, podría ser más eficiente eliminar un gran número de variables a la vez, pero se corre el riesgo de degradar la clasificación.

En esta tesis hemos usado la librería RFE en R, desarrollado por Ambroise y McLachlan, para seleccionar variables con el método RFE.

3.2.3 El Método de selección de variables RELIEF

Este método genera subconjuntos de variables de manera heurística. La idea principal es seleccionar aquellas variables de acuerdo a cómo se distinguen las clases a través de las distancias entre observaciones próximas. En el método se asignan pesos W_j a las variables, midiendo de esta manera la relevancia de las variables a base de distancias. Se calculan las distancias de las observaciones con respecto a una observación x que fue seleccionada aleatoriamente de la muestra

de entrenamiento, y se identifica la observación más cercana a x, y que pertenezca a la misma clase, a ésta se le llama "*Nearhit*"; mientras que la observación más cercana a la elegida pero que pertenezca a la otra clase, se le llama "*Nearmiss*". En caso de empate en las distancias se sugiere tomar el promedio de ellas (las coordenadas de las observaciones que empatan) para el cálculo de los pesos.

Si en el cálculo la diferencia entre el x_j y el "*Nearmiss*" es mayor a la diferencia entre x_j y el "*Nearhit*", entonces esto indica que la variable es buena y esto hace que se incremente el valor de su peso W_j. El método RELIEF seleccionará aquellas variables cuyos pesos finales sean mayores que un umbral ("*threshold*") predeterminado. Usualmente se sugiere umbrales iguales o cercanos a cero.

La cantidad de muestras aleatorias debe ser aproximadamente igual al número total de observaciones del conjunto de datos original.

Kononenko (1994) extendió la aplicación del RELIEF a problemas con más de dos clases. El algoritmo propuesto por Kononenko, llamado RELIEF-F, consiste en encontrar un "Nearmiss" para cada clase diferente y luego promediar su contribución en la estimación de los pesos.

Para obtener los resultados bajo este método de selección de variables, se utilizó la función reliefcont de la librería dprep de R, elaborada por Edgar Acuña.

3.2.4 Análisis de Componentes Principales

Sea X la matriz de observaciones de dimensión $n \times p$ con matriz variancia-

covariancia estimada $\hat{\Sigma}$, de dimensión $p \times p$, donde se cumple que:

$$tr\left(\hat{\Sigma}\right) = \hat{\sigma}_{11} + \hat{\sigma}_{22} + \ldots + \hat{\sigma}_{pp} \qquad 3.8$$

El objetivo del Análisis de Componentes Principales (Hotelling, 1933) o PCA por sus siglas en inglés (*Principal Components Analysis*) es hacer una reducción de dimensionalidad. Es decir, la información contenida en p variables predictoras $X = \left(X_1, \ldots, X_p\right)$ reducirla a $Z = \left(Z_1, \ldots, Z_{p'}\right)$, con $p' < p$; las nuevas variables Z_i son llamados los componentes principales y son no correlacionados entre si. Geométricamente hablando, la aplicación de componentes principales equivale a hacer una rotación de los ejes coordenados.

El primer paso en el PCA es estandarizar todas las variables predictoras X_j.

Así, sea X^* la matriz estandarizada, obtenida usando:

$$X_j^* = \frac{x_{ij} - \bar{x}_j}{s_j} \quad \forall \quad \begin{matrix} i=1,2,\ldots,n \\ j=1,2,\ldots,p \end{matrix} \qquad 3.9$$

donde:

x_{ij} es la *i-ésima* observación correspondiente a la *j-ésima* variable;

$\bar{x}_j$ es la media muestral de la *j-ésima* variable y

s_j es la desviación estándar de la *j-ésima* variable.

Luego, $X^{*T}_{pn} X^*_{np}$ es la matriz de correlaciones de las variables predictoras X_j.

Para determinar los componentes principales, hay que hallar una matriz ortogonal V tal que $Z_{nm} = X^*_{np} V_{pm}$, para la cual $\left(Z^T Z\right)_{mm} = \left(X^* V\right)^T \left(X^* V\right) = diag\left(\lambda_1, \ldots, \lambda_m\right)$, de tal forma que $V^T_{mp} V_{pm} = \mathrm{I}_{mm}$ y $V_{pm} V^T_{mp} = I_{pp}$, donde los λ_j para $j = 1, 2, \ldots, m$

son los valores propios (o autovalores) de la matriz de correlación $X^{*T}X^{*}$.

El número máximo m de componentes principales que se puede construir es igual a $m = \min(n, p)$, el mínimo entre el número de observaciones y variables predictoras. Aplicando una propiedad de los valores propios se tiene:

$$tr\left(\hat{\Sigma}\right) = \sum_{i=1}^{p} \lambda_i \qquad 3.10$$

Por lo tanto, la *j-ésima* componente principal Z_j tiene desviación estándar igual a $\sqrt{\lambda_j}$ y puede ser escrita como:

$$Z_j = v_{j1}X_1^{*} + v_{j2}X_2^{*} + \ldots + v_{jp}X_p^{*} \qquad 3.11$$

donde $v_{j1}, v_{j2}, \ldots, v_{jp}$ son los elementos de la *j-ésima* fila de V.

La matriz V es llamada la matriz de cargas ("*loadings*") y contiene los coeficientes de las variables en cada componente principal.

Los valores calculados de los componentes principales Z_j son llamados los valores rotados o simplemente "*scores*".

Cuando sólo se utiliza PCA y no se hace otro análisis posterior, decidir sobre la cantidad de componentes principales que se deben utilizar es un gran problema. Sin embargo cuando, posteriormente, se utilizan los componentes en un análisis de clasificación supervisada, se puede considerar que dicha cantidad de componentes será la que permita reducir el error de mala clasificación. Una descripción más específica se brindará en la Sección 3.4 del presente capítulo.

3.3 Descripción del Algoritmo

Como se mencionó anteriormente nuestro método realiza selección de variables y posteriormente con las variables seleccionadas se hace extracción de variables mediante el método de Componentes Principales.

La idea de la selección de variables induce a que ésta puede realizarse de dos formas:

- A toda la muestra (dado que estamos trabajando con muestras pequeñas) o
- A parte de ella (muestra de entrenamiento) de acuerdo a la forma de estimación de error de mala clasificación presentado en la Sección 2.5.

A la primera propuesta la denominaremos CPS1 y a la segunda CPS2.

Para ambas propuestas, se tiene una matriz de variables predictoras continuas X $(n \times p)$ y un vector de respuesta de tipo nominal $Y\ (n \times 1)$.

3.3.1 Primera Propuesta: Algoritmo CPS1

La descripción de la primera propuesta se presenta a continuación:

1. A todo el conjunto de datos aplíquese un método de selección de variables para determinar las p_1 variables que mejor discriminan las clases; p_1 puede tomar diferentes valores (en nuestro caso $p_1 = 100$). Este paso dará como resultado la matriz reducida $X^R\ (n \times p_1)$, donde $p_1 < p$.

2. Dividir $X^R\ (n \times p_1)$ e $Y\ (n \times 1)$ en muestra de entrenamiento $\mathcal{L}$, formada por X_L^R (n_L, p_1) y $Y_L\ (n_L, 1)$ (donde, $n_L = 2n/3$); y muestra de prueba $\mathcal{T}$, formada por X_T^R (n_T, p_1) y $Y_T\ (n_T, 1)$; (donde, $n_T = n/3$). De tal forma que para cada clase en $\mathcal{L}$ y

$\mathcal{T}$. también aparezca la proporción 2/3 y 1/3.

3. Hallar los vectores de medias y de desviaciones estándar de $X_L^R\left(n_L, p_1\right)$ los cuales serán utilizados para estandarizar $X_T^R\left(n_T, p_1\right)$.
4. Estandarizar la matriz $X_L^R\left(n_L, p_1\right)$ (de acuerdo a 3.9), obteniéndose $X_L^{R*}\left(n_L, p_1\right)$.
5. Aplicar Análisis de Componentes Principales a $X_L^{R*}\left(n_L, p_1\right)$; de donde se obtendrán los datos transformados $Z_L\left(n_L, m\right)$ (*scores*) y las cargas $C\left(p_1, m\right)$ (*loadings*).
6. Aplicar a $Z_L\left(n_L, m\right)$ y $Y_L\left(n_L, 1\right)$ los clasificadores para obtener el error de mala clasificación de tal manera que permita determinar el número de componentes óptimo m a utilizar en $\mathcal{T}$.
7. Estandarizar $X_T^R\left(n_T, p_1\right)$ con las medias y desviaciones estándar obtenidas en el paso 3 (de acuerdo a 3.9), lo que da como resultado $X_T^{R*}\left(n_T, p_1\right)$. Multiplicar $X_T^{R*}\left(n_T, p_1\right)$ con las cargas $C\left(p_1, m\right)$ obtenidas en el paso 5 para obtener $Z_T\left(n_T, m\right)$.
8. Aplicar los clasificadores a $Z_T\left(n_T, m\right)$ y $Y_T\left(n_T, 1\right)$ para obtener el error de mala clasificación en la muestra de entrenamiento.
9. Repetir los pasos 2 al 8, r veces; dado que de acuerdo al paso 2 se pueden obtener diferentes muestras aleatorias de entrenamiento y de prueba.

   En esta tesis se usó $r = 50$, que es un número de repeticiones consideradas en varias publicaciones.

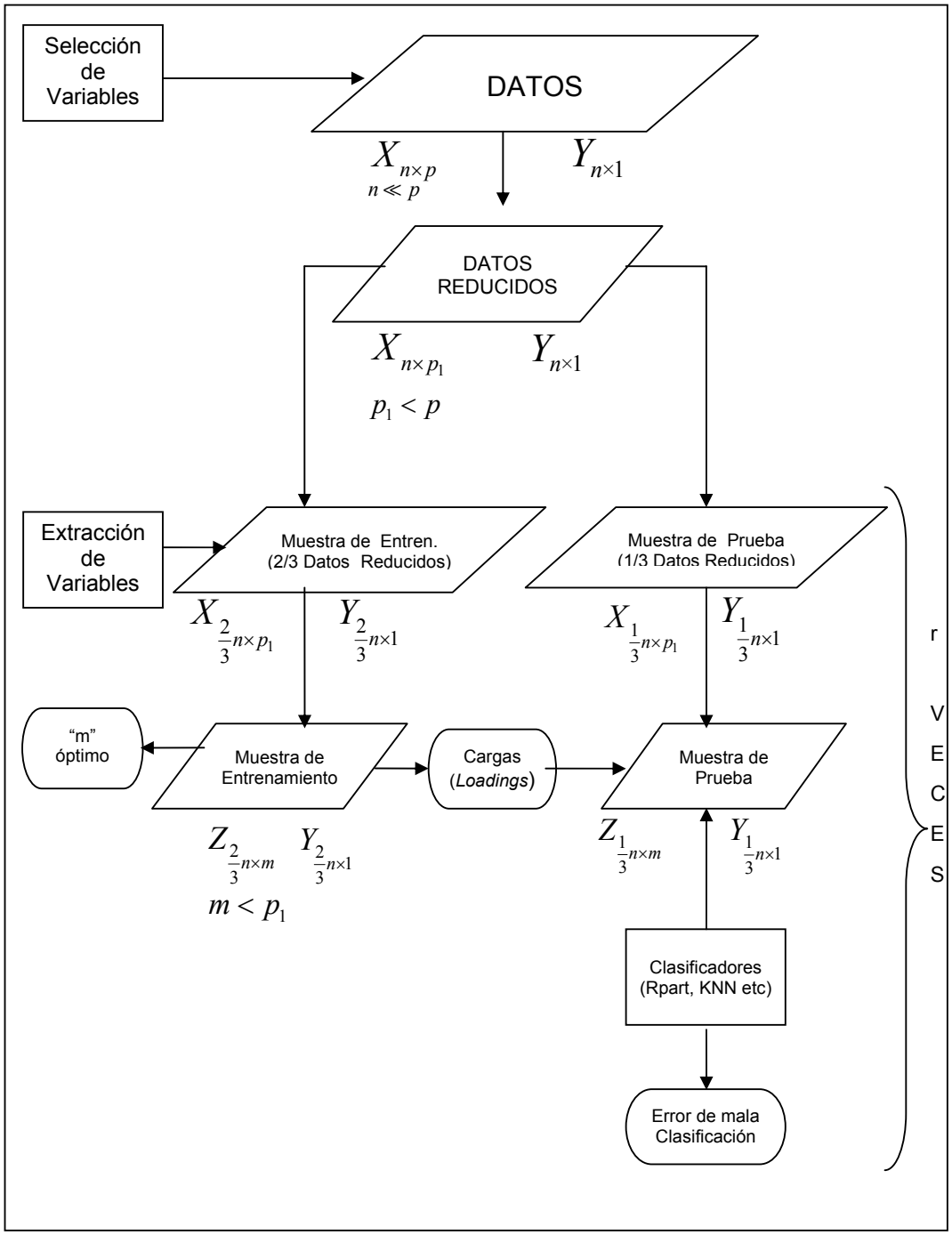


Figura 3.1 Primera Propuesta Algoritmo CPS1

3.3.2 Segunda Propuesta: Algoritmo CPS2

1. Dividir $X\ (n\times p)$ e $Y\ (n\times 1)$ en la muestra de entrenamiento $\mathcal{L}$, formada por $X_L\ (n_L, p)$ y $Y_L\ (n_L, 1)$ (donde, $n_L = 2n/3$); y la muestra de prueba $\mathcal{T}$, formada por $X_T\ (n_T, p)$ y $(n_T, 1)\ Y_T$, (donde $n_T = n/3$). De tal forma que para cada clase en $\mathcal{L}$ y $\mathcal{T}$ también aparezca la proporción 2/3 y 1/3.
2. Aplicar un método de selección de variables a $\mathcal{L}$ para determinar las variables que mejor discriminan las clases. Este paso dará como resultado matrices reducidas $X_L^R\ (n_L \times p_1)$ y $X_T^R\ (n_T \times p_1)$ donde, $p_1 < p$; p_1 puede tomar diferentes valores (en nuestro caso $p_1 = 100$).
3. Se aplican igualmente los pasos 3 al 8 de la propuesta anterior, (con lo que en esta propuesta también se tendría 8 pasos).
4. El paso 9 para esta propuesta consiste en repetir los pasos 1 al 8, *r* veces; dado que de acuerdo al paso 1 se pueden obtener diferentes muestras aleatorias de entrenamiento y de prueba.

Gráficamente el algoritmo CPS2 se presenta a continuación:

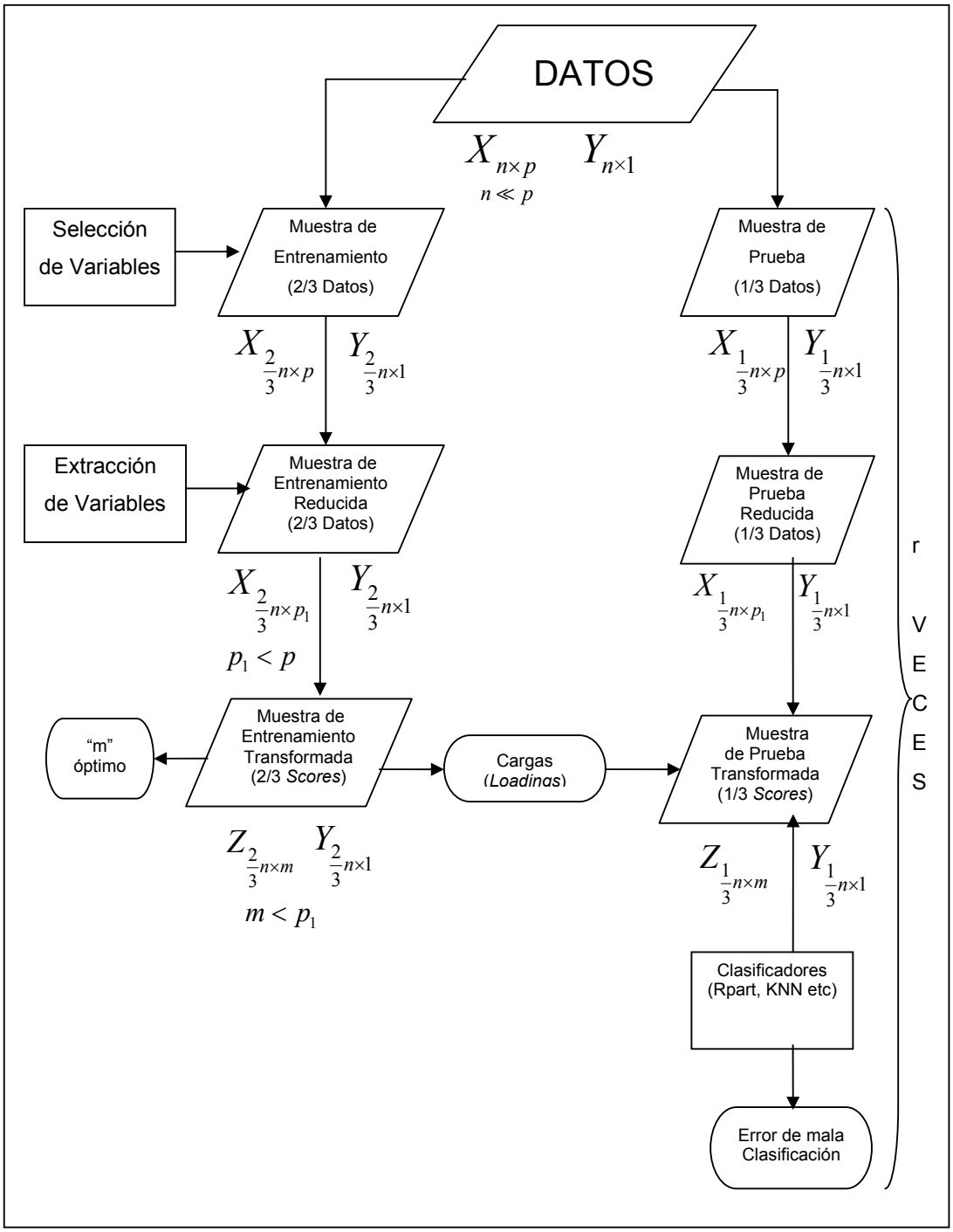


Figura 3.2 Segunda Propuesta Algoritmo CPS2

3.4 Elección del número de componentes

Cuando se trabaja con Análisis de Componentes Principales con el único objetivo de realizar una reducción de dimensión, (sin utilizar los componentes para un análisis posterior), no existe un procedimiento único que permita determinar el número correcto de componentes principales a utilizarse; sin embargo algunos de los más usados son:

a) Elegir el número de componentes hasta donde se ha acumulado por lo menos el 75% de la proporción de los valores propios.

b) Elegir hasta la componente cuyo valor propio sea mayor que 1. Para esto se puede utilizar el gráfico "*Scree Plot*".

Un método simple propuesto por Boulesteix (2004) indica que cuando la muestra original ha sido dividida en muestra de entrenamiento (ó aprendizaje) $\mathcal{L}$ (*learning set*) y muestra de prueba $\mathcal{T}$ (*test set*); solo la muestra de entrenamiento $\mathcal{L}$ debe utilizarse para determinar m (el número de componentes a utilizar para estimar el error de mala clasificación).

El siguiente procedimiento debe ser repetido r veces: "el clasificador δ_{CPS} se construye usando sólo un φ% de las observaciones de $\mathcal{L}$ y se aplica al resto de las observaciones, para diferentes valores de m. Para cada una de las r repeticiones el error de mala clasificación se calcula usando sólo el resto de las observaciones de $\mathcal{L}$. Después de las r repeticiones el error promedio de mala clasificación de las r repeticiones se calcula para cada valor de m. El valor de m

que minimice el error promedio de mala clasificación se usa para predecir la clase de las observaciones de $\mathcal{T}$; este valor se denotará por m_{opt}".

En resumen, se realiza la validación cruzada a la muestra de entrenamiento para diferentes números de componentes a utilizar, el proceso se repite *r* veces y se escogerá el número de componentes que minimice el error de mala clasificación.

En la presente tesis se utilizó este método sugerido por Boulesteix pero se tomaron diferentes valores de φ dependiendo del tamaño del conjunto de datos y se consideró un valor de r igual a cincuenta. Boulesteix indica que el valor de m_{opt} no depende del parámetro φ.

4 RESULTADOS Y DISCUSIONES

4.1 Conjunto de datos

- **Colon**: El conjunto de datos "colon" se describe extensamente en Alon et al. (1999). El conjunto de datos contiene los niveles de expresión de $p = 2000$ genes para $n = 62$ pacientes de dos clases; 22 pacientes se encuentran sanos y 40 pacientes tienen cáncer del colon. Este conjunto de datos se encuentra disponible gratuitamente como un archivo binario de R en http://stat.ethz.ch/~dettling/bagboost.html.

- **Leukemia**: Estos datos fueron introducidos por Golub et al. (1999) y contiene los niveles de expresión de 7129 genes para 47 pacientes con ALL-leukemia y 25 pacientes de AML-leukemia. Los datos originales se encuentra disponibles en http://www.genome.wi.mit.edu/MPR y también están incluidos en la librería de R golubEsets. Estos datos han sido preprocesados por Dudoit et al. (2002) y sólo se requieren 3571 genes. Los datos preprocesados están disponibles en http://stat.ethz.ch/~dettling/bagboost.html.

- **Prostate**: Este conjunto de datos presenta los niveles de expresión de 12600 genes para 50 tejidos normales y 52 tejidos con cáncer de próstata. Después de aplicar data pre-procesamiento descrito en Singh et al. (2002) se redujo a 6033 genes. Este conjunto de datos se encuentra disponible en:
  http://stat.ethz.ch/~dettling/bagboost.html.

- **Carcinoma**: Este conjunto de datos comprende los niveles de expresión de 7457 genes para 18 tejidos normales y 18 tejidos con carcinoma. Una descripción más extensa de este conjunto de datos, se encuentra en Notterman et al. (2001). Los datos antes de ser analizados necesitan ser estandarizados por muestra para tener media cero y variancia uno. Los datos se encuentran en: http://microarray.princeton.edu/oncology/carcinoma.html.

- **Breast cancer (BRCA)**: Este conjunto de datos descrito por Hedenfalk (2001); fue preprocesado por Simon et al. (2003); contiene los niveles de expresión de 3226 genes para pacientes con cáncer de pecho para tres tipos de tumores: sporadic, *BRCA1* y *BRCA2*.

- **Lymphoma**: El conjunto de datos fue presentado por Alizadeh et al. (2000) incluye los niveles de expresión de 4026 genes para 62 pacientes de tres diferentes clases: 42 muestras de linfoma célula-B (B-CLL) 9 observaciones de linfoma folicular (FL) y 11 observaciones de leucemia linfocitica crítica (DLBCL). Los datos originales contienen 18000 genes de 96 pacientes clasificados en 9 clases. Los datos se encuentran disponibles en http://llmpp.nih.gov/lymphoma/data/figure1.

- **SRBCT**: Este conjunto de datos de expresión de genes es presentado en Kahn et al. (2001). Contiene los niveles de expresión de 2308 genes para 63 pacientes con tumores de pequeñas celular redondas azuladas "*Small Round Blue Cells Tumor*" (SRBCT) que pertenecen a una de las 4 clases siguientes: *Ewing family of tumors* (EWS), *non-Hodgkin lymphoma* (BL), *neuroblastoma* (NB) y *rhabdomyosarcoma*

(RMS). Los datos se encuentran disponibles en: http://stat.ethz.ch/~dettling/bagboost.html.

- **Brain**: Este conjunto de datos presentado por Pomeroy et al. (2002), contiene 5597 niveles de expresión genética para $n = 42$ en cinco tipos de tumores del sistema central nervioso: 10 medulloblastomas, 10 gliomas malignos, 10 AT/RT's, 8 PNETs y 4 cerebelo humano. Los datos originales obtenidos aplicando la tecnología Afflymetrix están disponibles en: http://www.genome.wi.mit.edu/MPR/CNS. En esta tesis se usaron los datos preprocesados disponibles en: http://stat.ethz.ch/~dettling/bagboost.html.

- **NCI**: Este conjunto de datos comprende los niveles de expresión de 5244 genes para 61 pacientes con 8 diferentes tipos de tumores: 9 de pecho, 5 del sistema nervioso central, 7 de colon, 8 de leucemia, 8 de melanoma, 9 de pequeñas células del pulmón con carcinoma, 6 del ovario, 9 renales (Ross et al.-2000). El preprocesamiento de los datos es descrito en Dudoit et al. (2002). Una descripción más detallada de este conjunto puede encontrarse en: http://genome-www.stanford.edu/nci60. En esta tesis se usaron los datos proporcionados por la Dra. Boulesteix.

La Tabla 4.1 muestra el resumen de los conjuntos de datos que han sido utilizadas en la presente tesis y la Tabla 4.2 presenta la división de la muestra en $\mathcal{L}$ y $\mathcal{T}$ dentro de cada clase para las nueve bases de datos.

TABLA 4.1: Descripción de los Conjuntos de Datos Utilizados

Conjunto de Datos	Publicación	n	p	Número de Observaciones dentro de cada clase								Descripción
				1	2	3	4	5	6	7	8	
Colon	Alon (1999)	62	2000	40	22							Tejidos: tumor / normal
Leukemia	Golub (1999)	72	3571	47	25							Subtipos de Leucemia: ALL / AML
Prostate	Singh (2002)	102	6033	50	52							Tejidos: tumor / normal
Carcinoma	Notterman (2001)	36	7457	18	18							Tejidos: tumor / normal
Breast cáncer (BRCA)	Hendenfalk (2001)	22	3226	7	8	7						Subtipos de tumores: sporadic/BRCA1/BRCA2
Lymphoma	Alizadeh (2000)	62	4026	42	9	11						Subtipos de Lymphoma B-CLL/FL/DLBCL
SRBCT	Khan (2001)	63	2308	23	20	12	8					Diferentes tipos de tumores
Brain	Pomeroy (2002)	42	5597	10	10	10	4	8				Diferentes tumores del sistema nervioso central
NCI	Ross (2000)	61	5244	9	5	7	8	8	9	6	9	Diferentes tipos de tumores

TABLA 4.2: Distribución de la muestra de entrenamiento y de prueba

Conjunto de Datos	n	K	Muestra de Entrenamiento	Muestra de Prueba
Colon	62	$k_1 = 22$	15	7
		$k_2 = 40$	27	13
Leucemia	72	$k_1 = 47$	31	16
		$k_2 = 25$	17	8
Prostate	102	$k_1 = 50$	33	17
		$k_2 = 52$	35	17
Carcinoma	36	$k_1 = 18$	12	6
		$k_2 = 18$	12	6
Breast cáncer (BRCA)	22	$k_1 = 7$	5	2
		$k_2 = 8$	5	3
		$k_3 = 7$	5	2
Lymphoma	62	$k_1 = 42$	28	14
		$k_2 = 9$	6	3
		$k_3 = 11$	7	4
SRBCT	63	$k_1 = 23$	15	8
		$k_2 = 20$	13	7
		$k_3 = 12$	8	4
		$k_4 = 8$	5	3
Brain	42	$k_1 = 10$	7	3
		$k_2 = 10$	7	3
		$k_3 = 10$	7	3
		$k_4 = 4$	3	1
		$k_5 = 8$	5	3
NCI	61	$k_1 = 9$	6	3
		$k_2 = 5$	3	2
		$k_3 = 7$	5	2
		$k_4 = 8$	5	3
		$k_5 = 8$	5	3
		$k_6 = 9$	6	3
		$k_7 = 6$	4	2
		$k_8 = 9$	6	3

4.2 Resultados

Para poder aplicar la metodología propuesta en este trabajo se requirió de la implementación de funciones que permitan llevar a cabo los cálculos necesarios para realizar las tareas computacionales requeridas. Estas funciones consideran desde el proceso de selección de variables hasta la obtención del error de mala clasificación a través de los Componentes Principales Supervisados.

La programación y la obtención de los resultados se llevó a cabo usando el lenguaje R, en el ambiente Windows, en computadoras con doble procesador Pentium Xeon de 2.80 GHz y con 3 GB de memoria RAM. En las funciones implementadas se usaron funciones pertenecientes a librerías en R como: pps, dprep, rpart, class, rfe y nnet.

4.2.1 Número de Variables Seleccionadas y tiempo de procesamiento

Como una primera etapa de experimentación, se deseaba conocer que métodos de selección de variables serían utilizados en la presente tesis. Por esa razón se evaluaron varios métodos de selección de variables que incluyeron de Filtro y de Envoltura. A estos métodos se les estimó el número de variables seleccionadas, así como el tiempo de procesamiento en el que incurrían. Entre los métodos utilizados tenemos: Kruskal-Wallis (con un nivel de significación de 0.05), Relief (evaluado con un número de muestra n igual al tamaño de muestra original y con un valor límite para los pesos de las variables iguala a 0), Las Vegas Filter (cuyas variables fueron previamente

discretizadas bajo el método Chi-Merge y analizadas posteriormente con un valor de inconsistencia igual a 0), SFS y SFFS (ambos para el caso del clasificador KNN con un vecino más cercado). Los resultados obtenidos se presentan en las siguientes tablas:

TABLA 4.3: Número de Variables seleccionadas

Conjunto de datos	Método de Filtro			Método de Envoltura			
	K - W	Relief	LVF	SFS		SFFS	
				1NN	Rpart	1NN	Rpart
Colon	289	1410	941	3	3	4	2
Leucemia	1101	2805	1713	1	1	1	2
Prostate	1843	3927	2900	3	1	3	2
Carcinoma	1506	4417	3603	1	1	1	1
Breastcc	349	2239	1532	3	2	3	1
Lymphoma	2603	3688	1921	3	2	2	2
SRBCT	1020	2135	1096	5	4	4	3
Brain	2313	4573	2663	4	3	4	2
NCI	3088	4868	2507	9	3	4	5

TABLA 4.4 Tiempo (en segundos) requerido para la selección de variables

Conjunto de datos	Método de Filtro			Método de Envoltura				
	K - W	Relief	LVF	RFE	SFS		SFFS	
					1NN	Rpart	1NN	Rpart
Colon	16.78	147.31	23.61	7.89	2763.61	8919.90	309.15	1501.17
Leucemia	29.64	377.41	49.97	6.41	432.55	1576.33	359.83	1694.81
Prostate	56.50	1031.42	165.56	93.47	523.92	4412.00	471.08	3435.02
Carcinoma	57.14	302.89	45.71	12.85	656.31	2615.69	659.38	3166.88
Breastcc	23.39	92.07	11.47	5.73	592.94	3020.27	230.05	3158.59
Lymphoma	32.49	390.47	37.96	24.27	301.39	4921.61	274.95	4221.75
SRBCT	18.69	258.44	21.88	10.07	288.90	2869.05	294.40	1984.88
Brain	43.38	461.47	47.81	13.12	764.17	8645.38	560.86	6321.24
NCI	43.15	949.62	74.69	17.95	968.46	5033.16	804.97	6456.18

En la Tabla 4.3 se puede observar que en cada conjunto de datos, los métodos de selección de variables seleccionan diferentes cantidades de variables (p_1). Debido a esto, se decidió fijar el número de variables seleccionadas en $p_1 = 100$. Este valor fue elegido debido a que esa cantidad de variables pareció un valor adecuado; sin embargo existe la posibilidad que en un estudio posterior, se pueda encontrar un p_1 óptimo, es decir aquel que permita reducir el error de mala clasificación.

También, se puede apreciar en la Tabla 4.3 que los métodos *Wrapper* seleccionan pocas variables, lo que indicaría que no sería necesario aplicar posteriormente el método de Componentes Principales a esas variables seleccionadas. Sin embargo, estos métodos son los que producen el mayor tiempo de procesamiento (a excepción del RFE) (Tabla 4.4). Por estas dos razones se descartaron a los métodos *Wrapper* SFS y SFFS como métodos de selección de variables.

Las Vegas *Filter* es un método que presenta la desventaja en la elección de un límite de selección y que las variables si son continuas necesitan discretizarse; así mismo cuenta como parámetro al número de "iteraciones" que se realizarán para la selección de las variables; cuando este valor se incrementa trae como consecuencia un alto costo computacional; por lo que este método de selección fue descartado.

El método de selección RFE no fue incluido en la tabla 4.3 debido a que es un método que presenta un parámetro que representa el número mínimo de variables que serán seleccionadas. Es decir el número de variables seleccionadas ya se encuentra

predeterminado. En comparación a los otros métodos presentados, éste es el que utiliza el menor tiempo de procesamiento en la selección de variables. Guyon et al. (2001), Zhu y Hastie (2004) lo utilizan como método de selección de variables para datos provenientes de expresión genética. En esta tesis este método se uso con el fin de conocer el efecto que causa en la formación de los componentes principales.

Nguyen y Rocke (2002) utilizan pruebas paramétricas como el Análisis de Variancia de una vía (*one-way* ANOVA) o la prueba *T* como métodos de selección de variables, (obteniendo buenos resultados); sin embargo estos tipos de pruebas requieren la verificación de supuestos; los cuales en muchos casos no son realizadas. Debido a que los datos provenientes de expresión genética presentan mucha variabilidad se utilizó aquí la prueba de Kruskal-Wallis que es la contraparte no paramétrica al Análisis de Varianza de una vía.

Finalmente, en el trabajo de Wang et al. (2004) se menciona que el método de selección de variables Relief brinda buenos resultados, afirmación que se verificó en este trabajo.

4.2.2 Error de Mala clasificación

Como ya se mencionó anteriormente, se cuenta con dos algoritmos para realizar la estimación del error de mala clasificación mediante Componentes Principales Supervisados a los que hemos denominado: CPS1 y CPS2. Antes de la formación de estos componentes se necesita realizar la selección de variables, para lo cual se utilizaron en

forma independiente tres métodos: Kruskal-Wallis, RFE y Relief.

Después de obtenidos los CPS se les aplicarán tres tipos diferentes de clasificadores: Rpart, KNN y Regresión Logística Nominal; en el caso del clasificador KNN se utilizó el valor de $k=3$, debido a que se hizo varias pruebas con diferentes valores de k y con ese valor fue el que se obtuvo los mejores resultados.

Lo anterior implica que se proponen dieciocho maneras distintas de estimar el error de mala clasificación las cuales se aplicaron a nueve conjuntos de datos. El objetivo fue encontrar aquella combinación que minimizara el error de mala clasificación.

Cabe recalcar nuevamente que en la literatura que trabaja con selección de variables en datos provenientes de expresión genética se encontró que se desea tener un número de fijo de variables seleccionadas (*top features*), el método RFE no presenta este inconveniente, pero los métodos Kruskal-Wallis y Relief sí; por lo tanto se decidió uniformizar el criterio y seleccionar una cantidad predeterminada de variables. Los resultados que se presentarán son para las 100 "mejores" variables; sin embargo las funciones que se elaboraron permiten hacer la selección de diferentes cantidades de variables.

A continuación se muestran los errores de mala clasificación en la muestra de prueba para los cinco primeros Componentes Principales Supervisados, así como un gráfico que indique la tendencia de estos errores, estimados bajo los dos algoritmos en los nueve conjuntos de datos (Tablas 4.5 – 4.22). Las Tablas 4.23 y 4.24 muestran un resumen de las tablas anteriores. Asimismo en el Apéndice C se presentan los gráficos que muestran la clasificación de las observaciones en dos y tres dimensiones para una muestra de prueba.

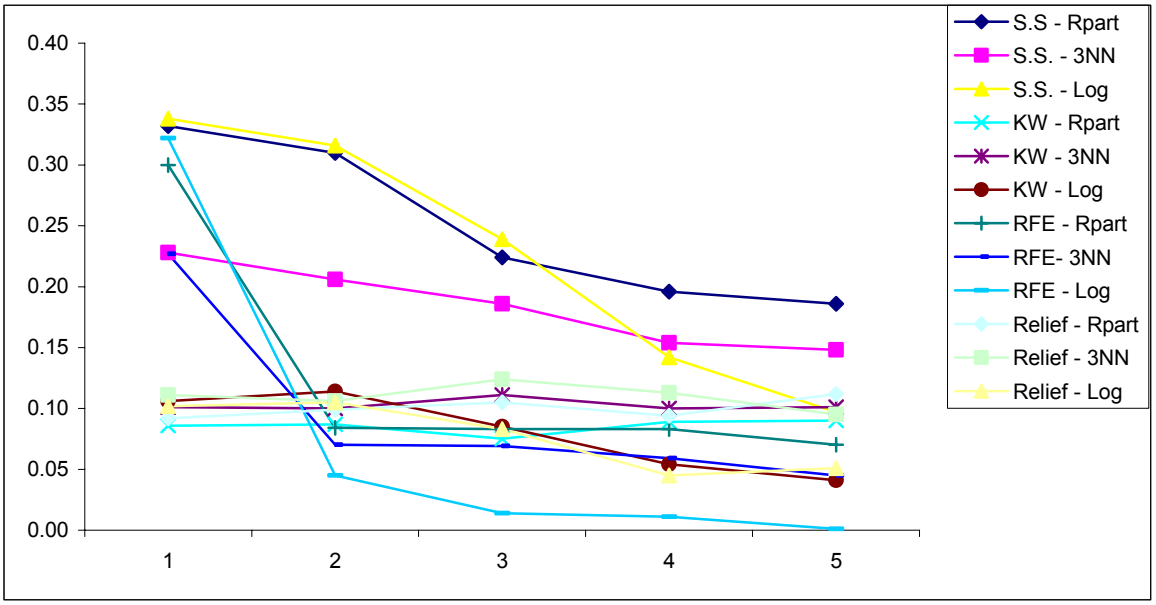
a1. Conjunto de Datos Colon: Algoritmo CPS1

TABLA 4.5 EMC estimado en Colon mediante CPS1

Método de Selección	Clasificador	Número de CPS				
		1	2	3	4	5
Sin selección	R-part	0.332	0.310	0.224	0.196	**0.186**
	3NN	0.228	0.206	0.186	0.154	**0.148**
	Logística	0.338	0.316	0.239	0.142	**0.097**
KW	R-part	0.086	0.087	0.075	**0.089**	0.090
	3NN	0.101	0.100	**0.111**	0.100	0.101
	Logística	0.106	0.114	0.085	**0.054**	0.041
RFE	R-part	0.300	**0.084**	0.083	0.083	0.070
	3NN	0.227	0.070	0.069	**0.059**	0.045
	Logística	0.322	0.045	**0.014**	0.011	0.001
Relief	R-part	0.092	0.099	0.105	0.094	**0.112**
	3NN	0.111	**0.106**	0.124	0.113	0.095
	Logística	0.102	0.105	**0.083**	0.045	0.051

* Los valores en negrita indican el número de componentes con el que se obtuvo el menor error de mala clasificación en la muestra de entrenamiento.

Figura 4.1 Tendencia de EMC en Colon con algoritmo CPS1



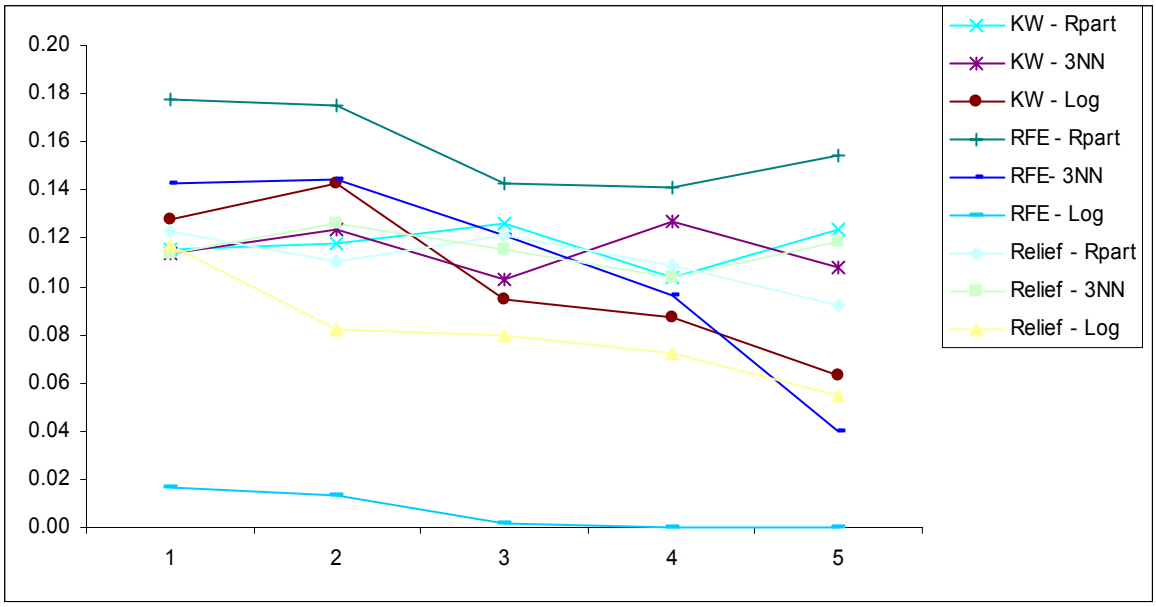
a2. Conjunto de Datos Colon: Algoritmo CPS2

TABLA 4.6 EMC estimado en Colon mediante CPS2

Métodos de Selección	Clasificador	Número de CPS				
		1	**2**	**3**	**4**	**5**
K-W	R-part	0.115	0.118	**0.126**	0.104	0.124
	3NN	0.114	**0.124**	0.103	0.127	0.108
	Logística	**0.128**	0.143	0.095	0.087	0.063
RFE	R-part	0.178	0.175	0.143	**0.141**	0.154
	3NN	0.143	**0.144**	0.121	0.096	0.040
	Logística	0.017	0.013	**0.002**	0.000	0.000
Relief	R-part	0.123	0.110	0.121	**0.109**	0.092
	3NN	0.114	**0.126**	0.115	0.104	0.119
	Logística	**0.117**	0.082	0.080	0.072	0.055

* Los valores en negrita indican el número de componentes con el que se obtuvo el menor error de mala clasificación en la muestra de entrenamiento.

Figura 4.2 Tendencia de EMC en Colon con algoritmo CPS2



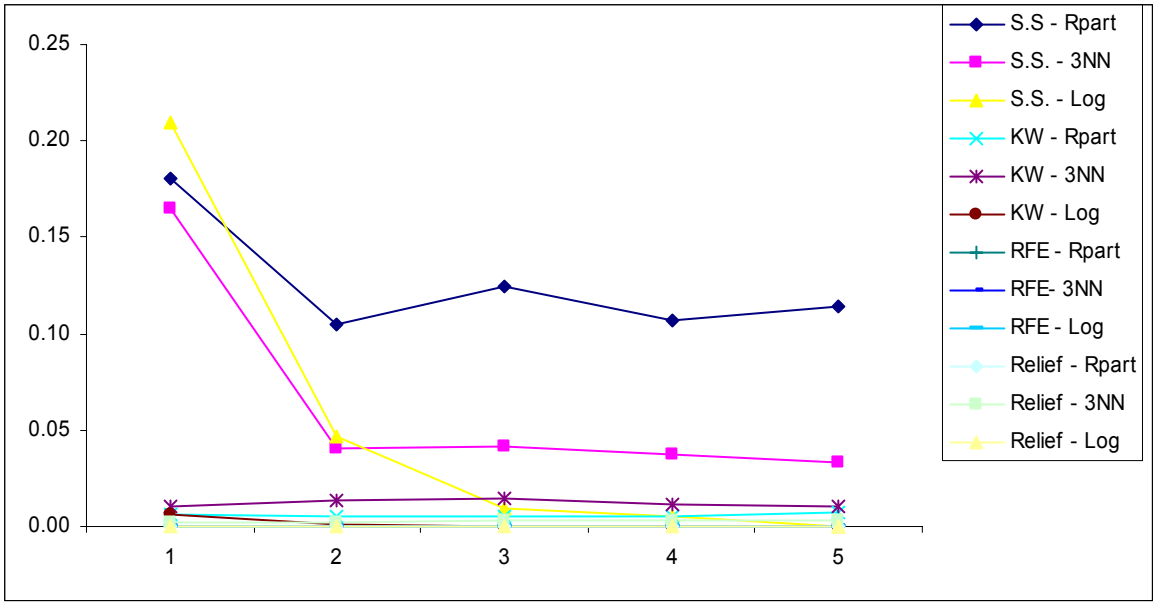
b.1) Conjunto de Datos Leukemia: Algoritmo CPS1

TABLA 4.7 EMC estimado en Leukemia mediante CPS1

Métodos de Selección	Clasificador	Número de CPS				
		1	2	3	4	5
Sin selección	R-part	0.180	0.105	0.125	0.107	**0.114**
	3NN	0.165	0.040	0.042	0.037	**0.033**
	Logística	0.210	0.047	0.009	0.005	**0.000**
KW	R-part	**0.006**	0.005	0.005	0.005	0.007
	3NN	**0.010**	0.013	0.015	0.011	0.010
	Logística	0.006	**0.001**	0.000	0.000	0.000
RFE	R-part	**0.000**	0.000	0.000	0.000	0.000
	3NN	**0.000**	0.000	0.000	0.000	0.000
	Logística	**0.000**	0.000	0.000	0.000	0.000
Relief	R-part	**0.000**	0.000	0.000	0.000	0.000
	3NN	**0.002**	0.002	0.003	0.003	0.003
	Logística	**0.000**	0.000	0.000	0.000	0.000

* Los valores en negrita indican el número de componentes con el que se obtuvo el menor error de mala clasificación en la muestra de entrenamiento.

Figura 4.3 Tendencia de EMC en Leukemia con algoritmo CPS1



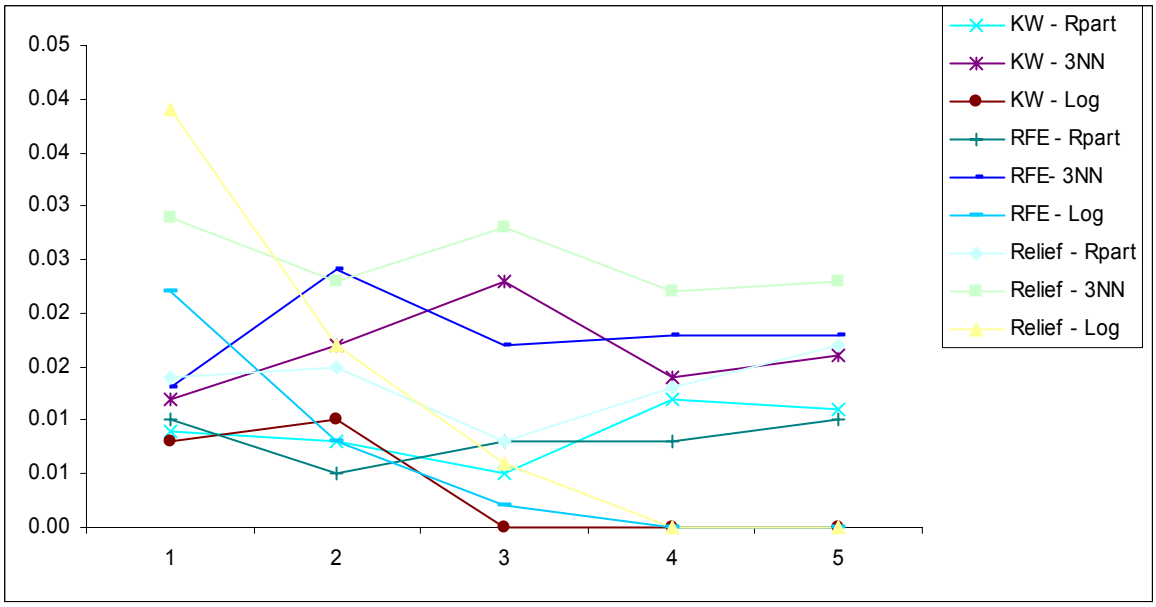
b.2) Conjunto de Datos Leukemia: Algoritmo CPS2

TABLA 4.8 EMC estimado mediante Leukemia CPS2

Métodos de Selección	Clasificador	Número de CPS				
		1	**2**	**3**	**4**	**5**
K-W	R-part	0.009	0.008	0.005	**0.012**	0.011
	3NN	0.012	0.017	**0.023**	0.014	0.016
	Logística	0.008	**0.010**	0.000	0.000	0.000
RFE	R-part	**0.010**	0.005	0.008	0.008	0.010
	3NN	**0.013**	0.024	0.017	0.018	0.018
	Logística	0.022	0.008	**0.002**	0.000	0.000
Relief	R-part	0.014	0.015	**0.008**	0.013	0.017
	3NN	0.029	**0.023**	0.028	0.022	0.023
	Logística	0.039	0.017	**0.006**	0.000	0.000

* Los valores en negrita indican el número de componentes con el que se obtuvo el menor error de mala clasificación en la muestra de entrenamiento.

Figura 4.4 Tendencia de EMC en Leukemia con algoritmo CPS2



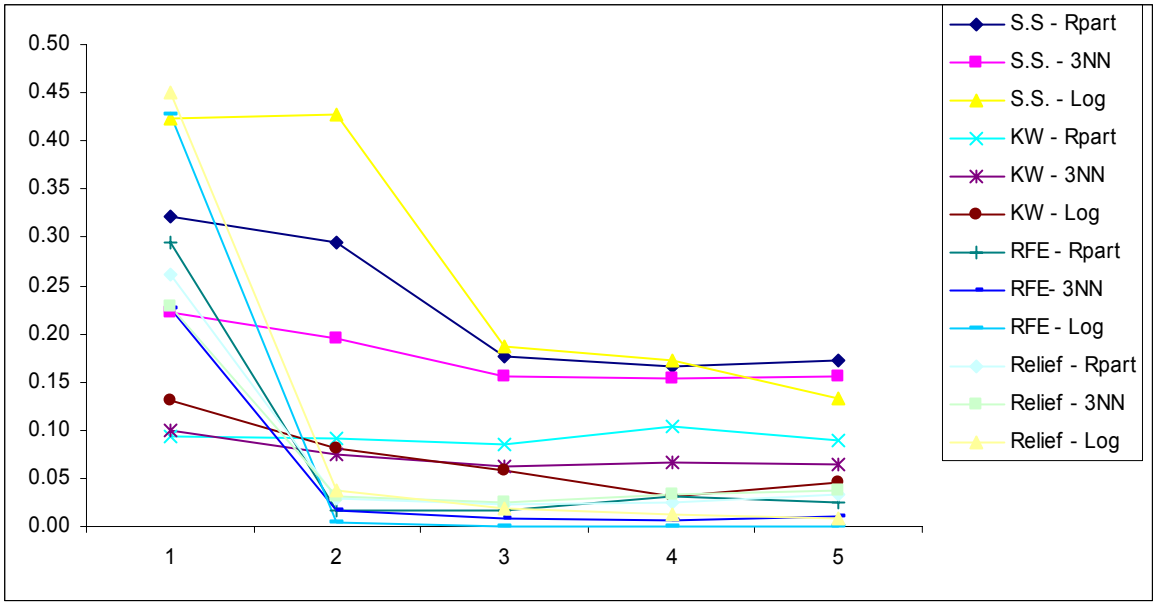
c.1) Conjunto de Datos Prostate: Algoritmo CPS1

TABLA 4.9 EMC estimado en Prostate mediante CPS1

Método de Selección	Clasificador	Número de CPS				
		1	**2**	**3**	**4**	**5**
Sin selección	R-part	0.321	0.295	0.176	**0.167**	0.172
	3NN	0.221	0.196	0.156	**0.154**	0.156
	Logística	0.423	0.427	0.187	0.173	**0.132**
KW	R-part	**0.093**	0.092	0.085	0.103	0.089
	3NN	0.100	0.075	**0.063**	0.067	0.064
	Logística	0.131	0.081	0.058	**0.031**	0.046
RFE	R-part	0.295	0.016	**0.016**	0.031	0.024
	3NN	0.227	0.017	**0.008**	0.006	0.010
	Logística	0.428	0.005	**0.000**	0.000	0.000
Relief	R-part	0.261	**0.030**	0.023	0.024	0.033
	3NN	0.229	**0.032**	0.024	0.034	0.037
	Logística	0.450	**0.037**	0.018	0.012	0.008

* Los valores en negrita indican el número de componentes con el que se obtuvo el menor error de mala clasificación en la muestra de entrenamiento.

Figura 4.5 Tendencia de EMC en Prostate con algoritmo CPS1



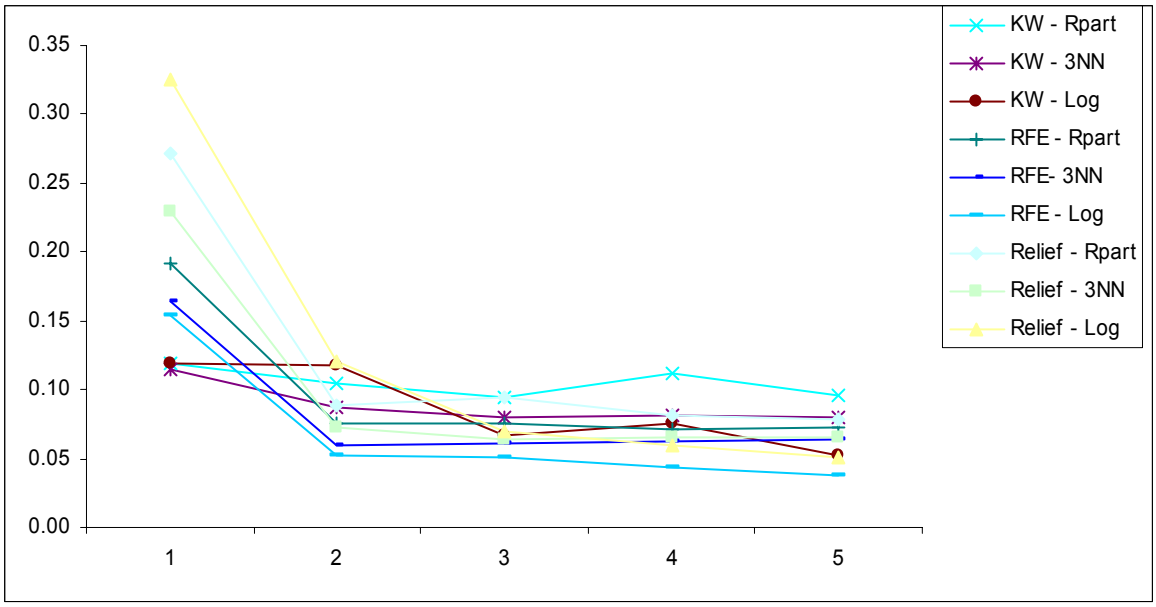
c.2) Conjunto de Datos Prostate: Algoritmo CPS2

TABLA 4.10 EMC estimado en Prostate mediante CPS2

Métodos de Selección	Clasificador	Número de CPS				
		1	**2**	**3**	**4**	**5**
K-W	R-part	0.119	0.105	0.094	**0.112**	0.096
	3NN	0.115	0.087	0.080	0.081	**0.080**
	Logística	0.119	0.117	**0.067**	0.076	0.052
RFE	R-part	0.191	0.076	0.076	**0.071**	0.073
	3NN	0.164	**0.060**	0.061	0.063	0.064
	Logística	0.154	0.053	**0.051**	0.044	0.038
Relief	R-part	0.272	0.089	**0.095**	0.081	0.078
	3NN	0.230	**0.073**	0.064	0.066	0.065
	Logística	0.325	0.121	**0.069**	0.059	0.051

* Los valores en negrita indican el número de componentes con el que se obtuvo el menor error de mala clasificación en la muestra de entrenamiento.

Figura 4.6 Tendencia de EMC en Prostate con algoritmo CPS2



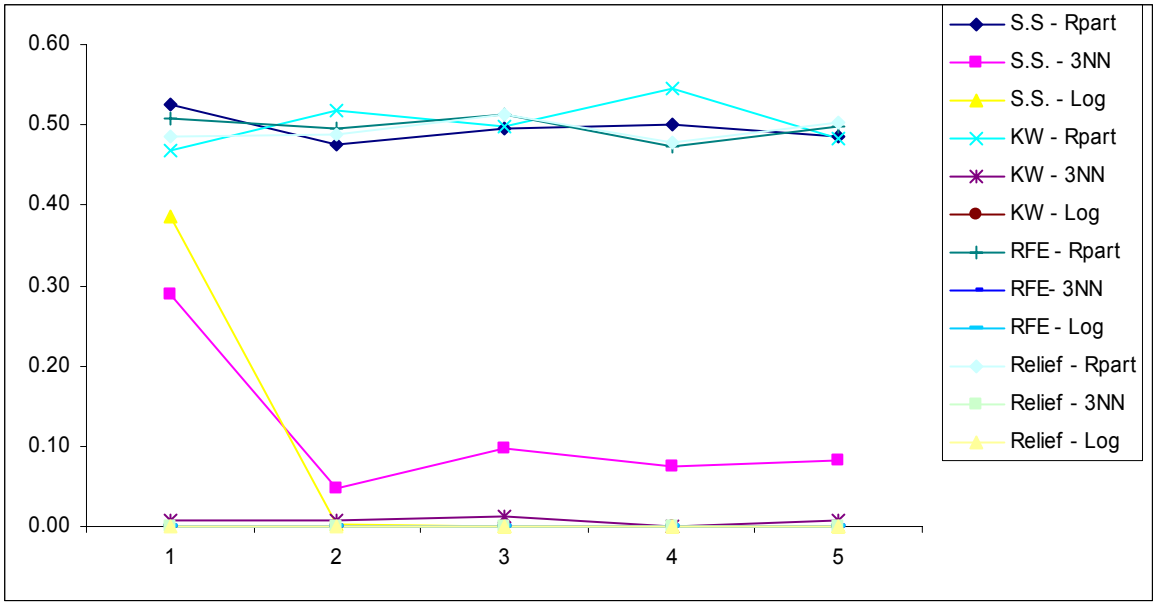
d.1) Conjunto de Datos Carcinoma: Algoritmo CPS1

TABLA 4.11 EMC estimado en Carcinoma mediante CPS1

Método de Selección	Clasificador	Número de CPS				
		1	2	3	4	5
Sin selección	R-part	0.525	0.475	0.495	**0.501**	0.485
	3NN	0.290	**0.048**	0.096	0.075	0.081
	Logística	0.387	0.003	**0.000**	0.000	0.000
KW	R-part	0.468	**0.517**	0.498	0.545	0.483
	3NN	0.007	0.008	**0.012**	0.000	0.007
	Logística	**0.000**	0.000	0.000	0.000	0.000
RFE	R-part	**0.508**	0.495	0.512	0.473	0.497
	3NN	**0.000**	0.000	0.000	0.000	0.000
	Logística	**0.000**	0.000	0.000	0.000	0.000
Relief	R-part	0.485	0.487	**0.513**	0.478	0.503
	3NN	**0.000**	0.000	0.000	0.000	0.000
	Logística	**0.000**	0.000	0.000	0.000	0.000

* Los valores en negrita indican el número de componentes con el que se obtuvo el menor error de mala clasificación en la muestra de entrenamiento.

Figura 4.7 Tendencia de EMC en Carcinoma con algoritmo CPS1



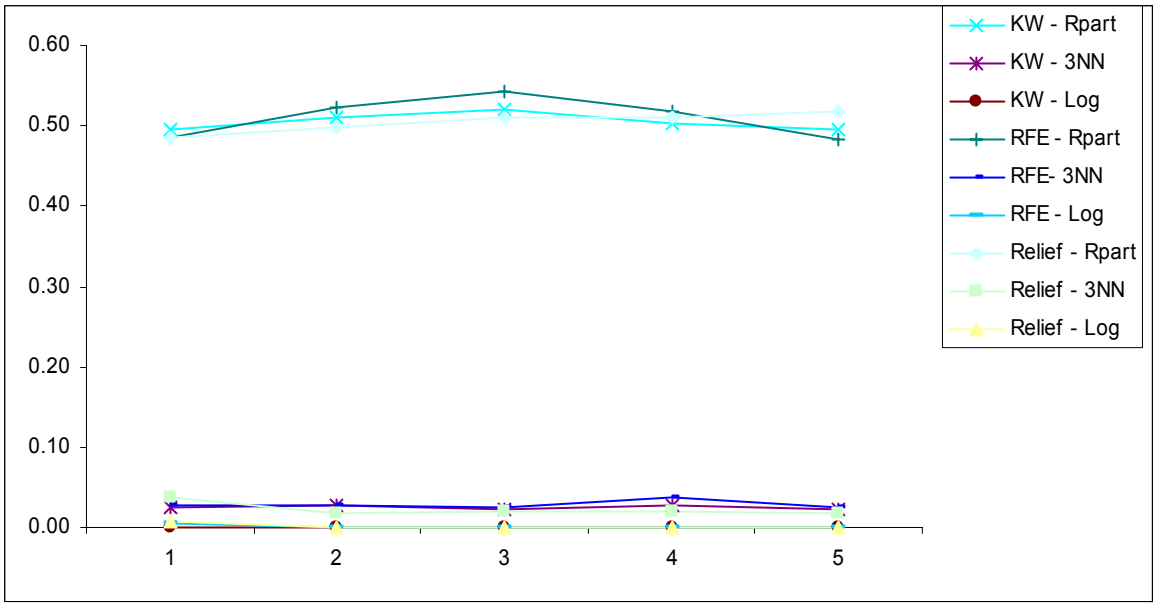
d.2) Conjunto de Datos Carcinoma: Algoritmo CPS2

TABLA 4.12 EMC estimado en Carcinoma mediante CPS2

Métodos de Selección	Clasificador	Número de CPS				
		1	**2**	**3**	**4**	**5**
K-W	R-part	0.495	0.510	**0.521**	0.503	0.495
	3NN	**0.026**	0.028	0.022	0.028	0.023
	Logística	**0.000**	0.000	0.000	0.000	0.000
RFE	R-part	0.485	0.523	0.543	0.518	**0.483**
	3NN	**0.028**	0.027	0.025	0.037	0.025
	Logística	0.004	**0.000**	0.000	0.000	0.000
Relief	R-part	0.485	0.498	0.510	0.510	0.517
	3NN	0.037	**0.018**	0.020	0.020	0.018
	Logística	0.008	**0.000**	0.000	0.000	0.000

* Los valores en negrita indican el número de componentes con el que se obtuvo el menor error de mala clasificación en la muestra de entrenamiento.

Figura 4.8 Tendencia de EMC en Carcinoma con algoritmo CPS2



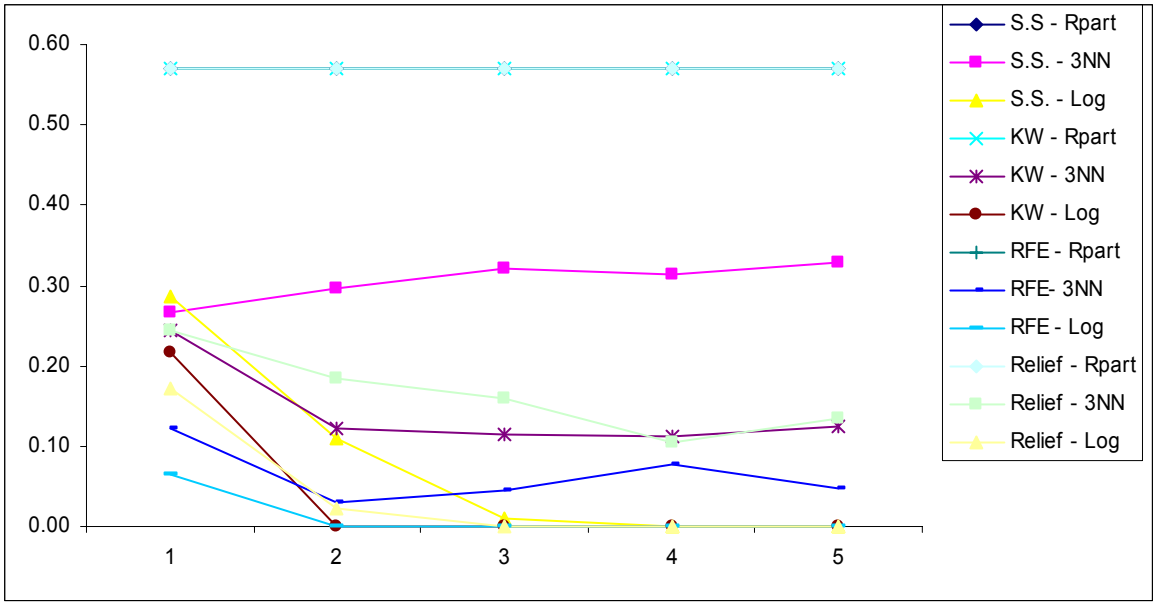
e.1) Conjunto de datos BRCA: Algoritmo CPS1

TABLA 4.13 EMC estimado en BRCA mediante CPS1

Método de Selección	Clasificador	Número de CPS				
		1	**2**	**3**	**4**	**5**
Sin selección	R-part	0.571	0.571	**0.571**	0.571	0.571
	3NN	0.266	0.297	**0.320**	0.314	0.329
	Logística	0.286	0.109	0.009	**0.000**	0.000
KW	R-part	**0.571**	0.571	0.571	0.571	0.571
	3NN	0.245	**0.122**	0.114	0.111	0.125
	Logística	0.217	**0.000**	0.000	0.000	0.000
RFE	R-part	**0.571**	0.571	0.571	0.571	0.571
	3NN	0.122	**0.031**	0.045	0.077	0.048
	Logística	0.065	**0.000**	0.000	0.000	0.000
Relief	R-part	**0.571**	0.571	0.571	0.571	0.571
	3NN	0.245	0.185	**0.160**	0.105	0.134
	Logística	0.171	0.022	**0.000**	0.000	0.000

* Los valores en negrita indican el número de componentes con el que se obtuvo el menor error de mala clasificación en la muestra de entrenamiento.

Figura 4.9 Tendencia de EMC en BRCA con algoritmo CPS1



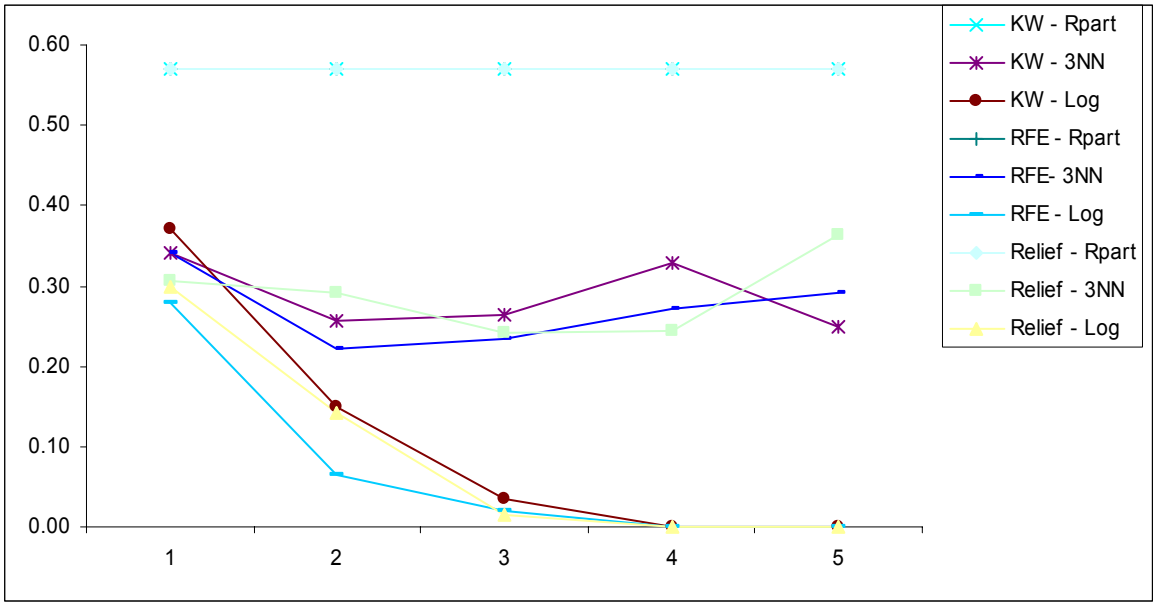
e.2) Conjunto de datos BRCA: Algoritmo CPS2

TABLA 4.14 EMC estimado en BRCA mediante CPS2

Métodos de Selección	Clasificador	Número de CPS				
		1	**2**	**3**	**4**	**5**
K-W	R-part	0.571	0.571	**0.571**	0.571	0.571
	3NN	0.342	**0.257**	0.264	0.328	0.250
	Logística	0.3714	0.150	**0.036**	0.000	0.000
RFE	R-part	0.571	0.571	0.571	0.571	**0.571**
	3NN	0.342	0.221	**0.235**	0.271	0.292
	Logística	0.279	0.064	**0.021**	0.000	0.000
Relief	R-part	0.571	**0.571**	0.571	0.571	0.571
	3NN	0.307	**0.292**	0.242	0.243	0.364
	Logística	0.300	0.143	**0.014**	0.000	0.000

* Los valores en negrita indican el número de componentes con el que se obtuvo el menor error de mala clasificación en la muestra de entrenamiento.

Figura 4.10 Tendencia de EMC en BRCA con algoritmo CPS2



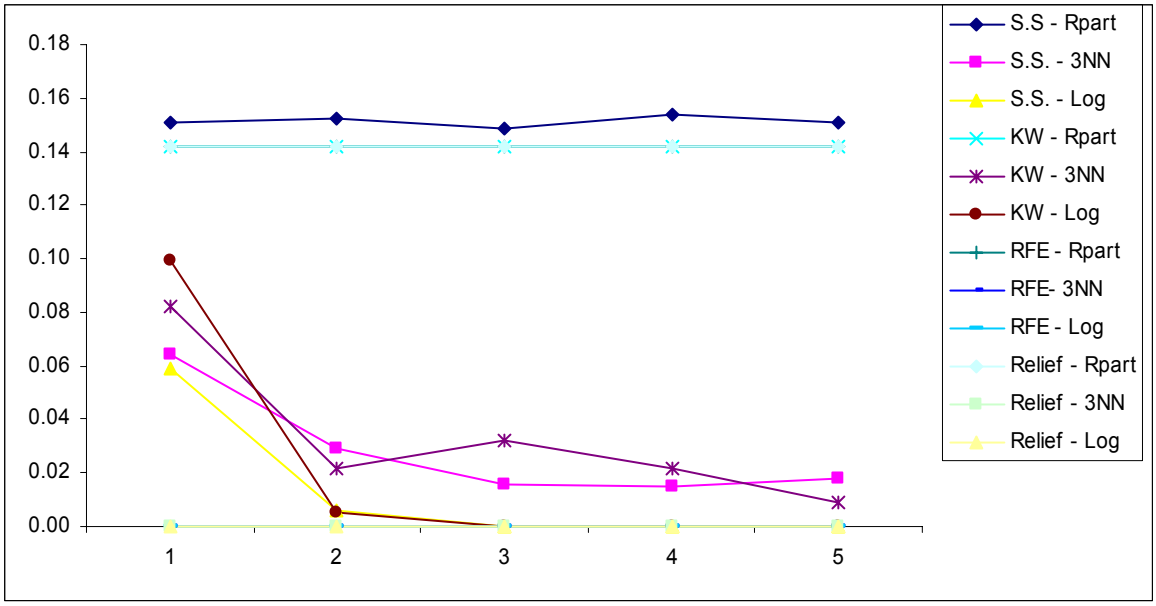
f.1) Conjunto de datos Lymphoma: Algoritmo CPS1

TABLA 4.15 EMC estimado en Lymphoma mediante CPS1

Método de Selección	Clasificador	Número de CPS				
		1	2	3	4	5
Sin selección	R-part	0.151	0.152	0.149	**0.154**	0.151
	3NN	0.064	0.029	0.016	0.015	**0.018**
	Logística	0.059	0.006	0.000	**0.000**	0.000
KW	R-part	0.142	**0.142**	0.142	0.142	0.142
	3NN	0.082	0.022	0.032	0.022	**0.009**
	Logística	0.099	0.005	**0.000**	0.000	0.000
RFE	R-part	0.142	**0.142**	0.142	0.142	0.142
	3NN	**0.000**	0.000	0.000	0.000	0.000
	Logística	**0.000**	0.000	0.000	0.000	0.000
Relief	R-part	0.142	**0.142**	0.142	0.142	0.142
	3NN	0.000	**0.000**	0.000	0.000	0.000
	Logística	**0.000**	0.000	0.000	0.000	0.000

* Los valores en negrita indican el número de componentes con el que se obtuvo el menor error de mala clasificación en la muestra de entrenamiento.

Figura 4.11 Tendencia de EMC en Lymphoma con algoritmo CPS1



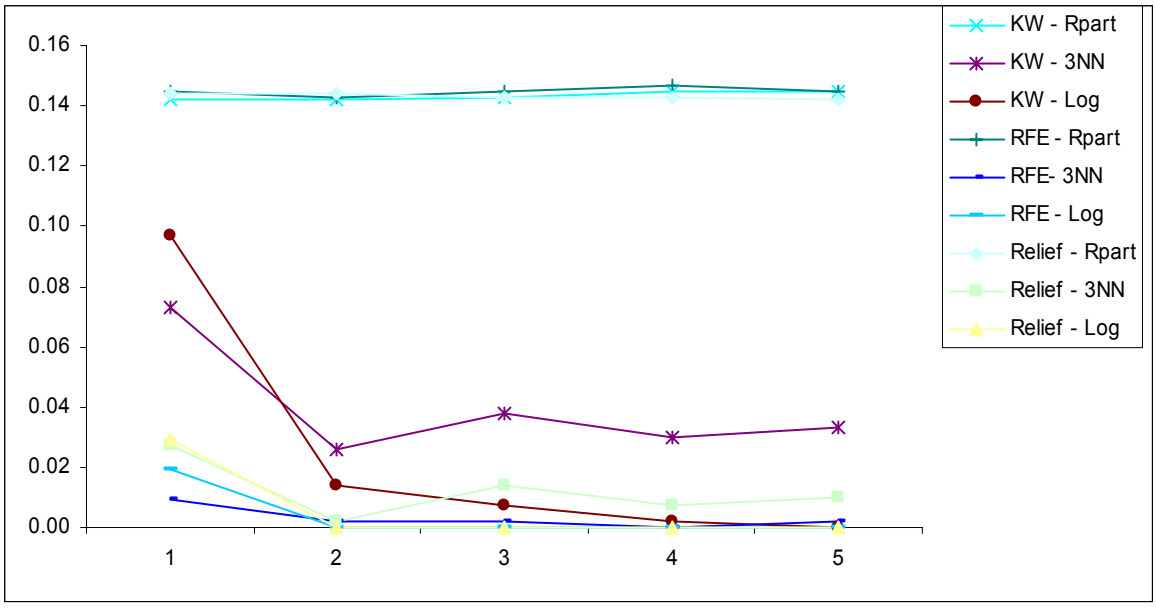
f.2) Conjunto de datos Lymphoma: Algoritmo CPS2

TABLA 4.16 EMC estimado en Lymphoma mediante CPS2

Métodos de Selección	Clasificador	Número de CPS				
		1	**2**	**3**	**4**	**5**
K-W	R-part	0.142	**0.142**	0.143	0.145	0.145
	3NN	0.073	0.026	0.038	0.030	**0.033**
	Logística	0.097	**0.014**	0.007	0.002	0.000
RFE	R-part	0.145	**0.143**	0.145	0.147	0.145
	3NN	0.009	**0.002**	0.002	0.000	0.002
	Logística	0.019	**0.000**	0.000	0.000	0.000
Relief	R-part	0.144	0.144	0.143	**0.143**	0.142
	3NN	**0.027**	0.002	0.014	0.007	0.010
	Logística	0.029	**0.000**	0.000	0.000	0.000

* Los valores en negrita indican el número de componentes con el que se obtuvo el menor error de mala clasificación en la muestra de entrenamiento.

Figura 4.12 Tendencia de EMC en Lymphoma con algoritmo CPS2



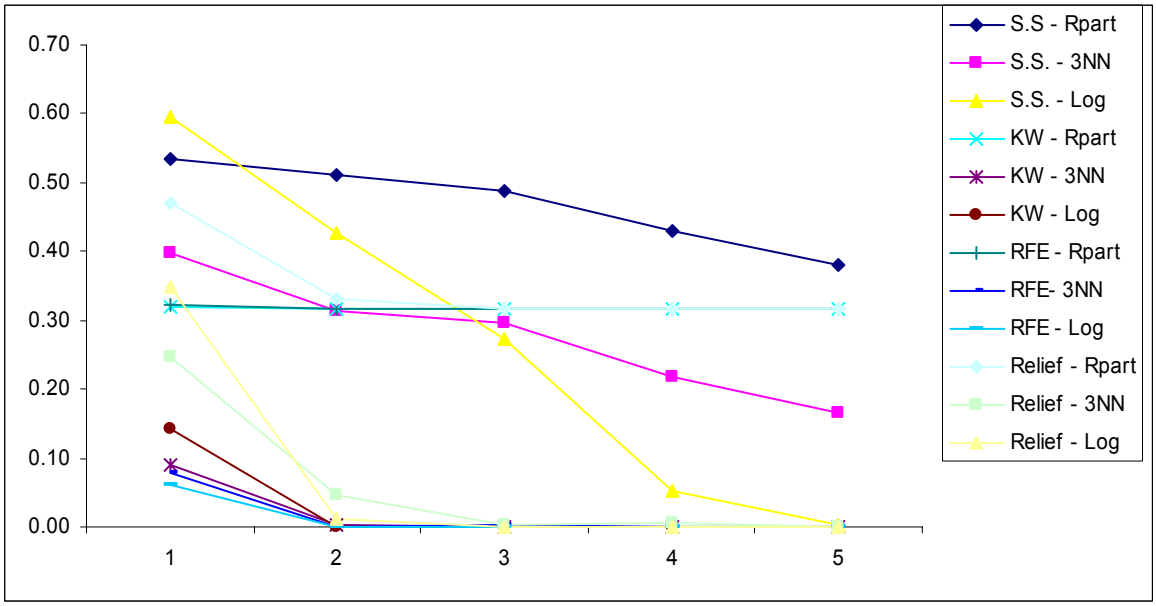
g.1) Conjunto de datos SRBCT: Algoritmo CPS1

TABLA 4.17 EMC estimado en SRBCT mediante CPS1

Método de Selección	Clasificador	Número de CPS				
		1	**2**	**3**	**4**	**5**
Sin selección	R-part	0.535	0.511	0.489	0.430	**0.380**
	3NN	0.399	0.314	0.296	0.219	**0.165**
	Logística	0.596	0.426	0.272	0.053	**0.003**
KW	R-part	**0.319**	0.318	0.318	0.318	0.318
	3NN	0.090	0.002	0.000	**0.001**	0.001
	Logística	0.142	**0.000**	0.000	0.000	0.000
RFE	R-part	0.321	0.318	0.318	**0.318**	0.318
	3NN	0.079	0.001	0.003	**0.001**	0.000
	Logística	0.062	**0.000**	0.000	0.000	0.000
Relief	R-part	0.471	**0.330**	0.318	0.318	0.318
	3NN	0.246	0.046	**0.004**	0.006	0.001
	Logística	0.350	0.011	**0.000**	0.000	0.000

* Los valores en negrita indican el número de componentes con el que se obtuvo el menor error de mala clasificación en la muestra de entrenamiento.

Figura 4.13 Tendencia de EMC en SRBCT con algoritmo CPS1



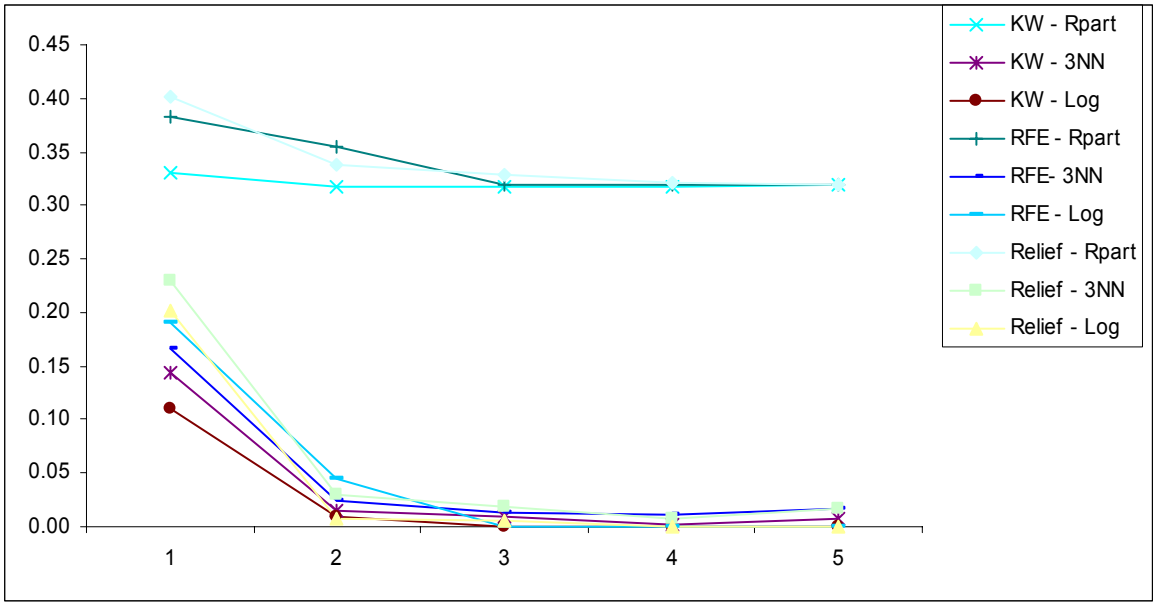
g.2) Conjunto de datos SRBCT: Algoritmo CPS2

TABLA 4.18 EMC estimado en SRBCT mediante CPS2

Métodos de Selección	Clasificador	Número de CPS				
		1	**2**	**3**	**4**	**5**
K-W	R-part	0.331	**0.318**	0.318	0.318	0.320
	3NN	0.143	0.015	**0.009**	0.002	0.007
	Logística	0.111	**0.009**	0.000	0.000	0.000
RFE	R-part	0.382	0.355	0.320	**0.320**	0.320
	3NN	0.166	0.025	**0.013**	0.011	0.016
	Logística	0.191	0.045	**0.000**	0.000	0.000
Relief	R-part	0.402	0.338	**0.329**	0.322	0.320
	3NN	0.229	0.029	0.018	**0.007**	0.016
	Logística	0.202	**0.007**	0.005	0.000	0.000

* Los valores en negrita indican el número de componentes con el que se obtuvo el menor error de mala clasificación en la muestra de entrenamiento.

Figura 4.14 Tendencia de EMC en SRBCT con algoritmo CPS2



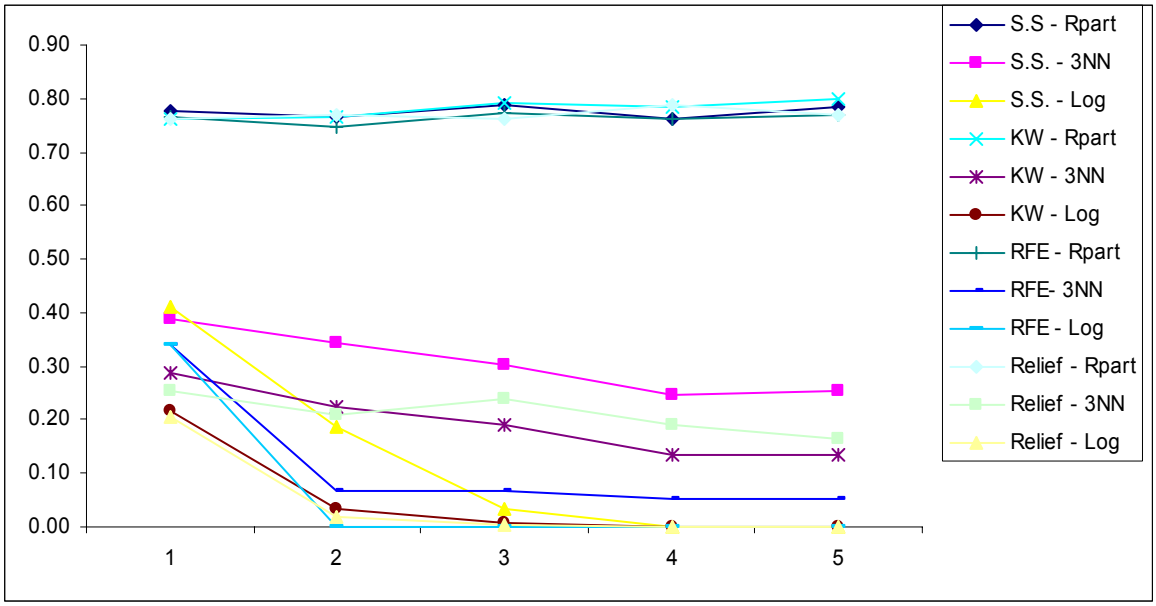
h.1) Conjunto de datos Brain: Algoritmo CPS1

TABLA 4.19 EMC estimado en Brain mediante CPS1

Método de Selección	Clasificador	Número de CPS				
		1	2	3	4	5
Sin selección	R-part	0.775	0.767	**0.787**	0.760	0.783
	3NN	0.390	0.344	0.301	**0.247**	0.253
	Logística	0.410	0.186	0.033	**0.001**	0.000
KW	R-part	0.761	**0.766**	0.793	0.786	0.800
	3NN	0.286	0.224	0.190	**0.135**	0.135
	Logística	0.215	0.035	0.009	**0.000**	0.000
RFE	R-part	0.766	0.747	**0.773**	0.760	0.770
	3NN	0.340	0.069	0.069	0.053	**0.053**
	Logística	0.341	**0.000**	0.000	0.000	0.000
Relief	R-part	0.761	0.770	0.760	**0.787**	0.770
	3NN	0.253	0.209	0.240	**0.190**	0.164
	Logística	0.206	0.020	0.003	**0.000**	0.000

* Los valores en negrita indican el número de componentes con el que se obtuvo el menor error de mala clasificación en la muestra de entrenamiento.

Figura 4.15 Tendencia de EMC en Brain con algoritmo CPS1



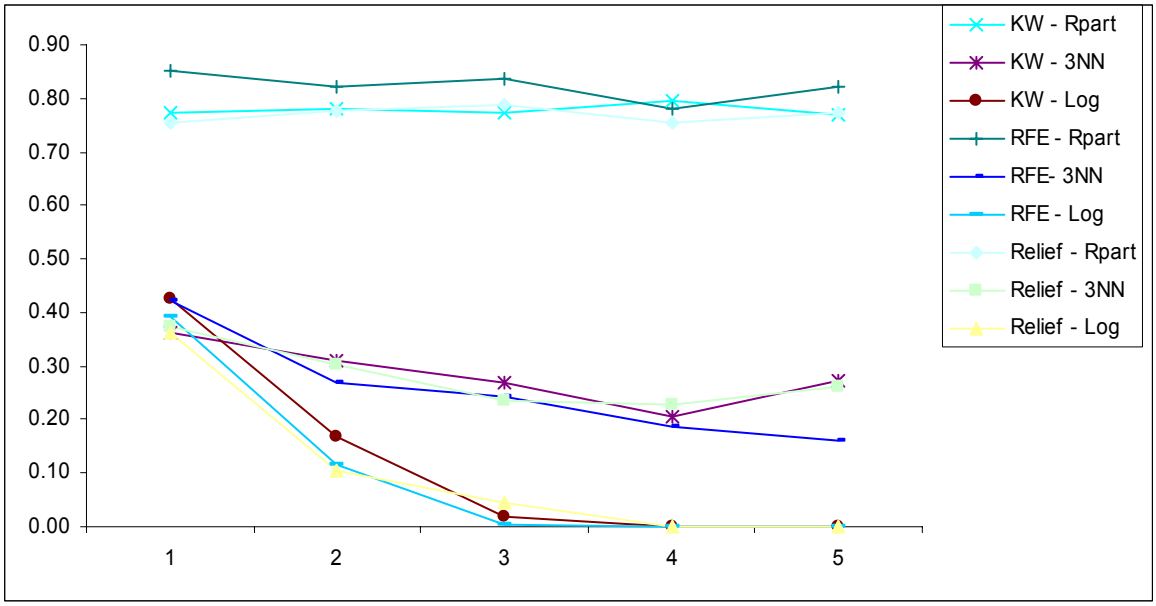
h.2) Conjunto de datos Brain: Algoritmo CPS2

TABLA 4.20 EMC estimado en Brain mediante CPS2

Métodos de Selección	Clasificador	Número de CPS				
		1	**2**	**3**	**4**	**5**
K-W	R-part	0.773	0.781	0.773	0.796	**0.769**
	3NN	0.361	0.311	0.269	**0.207**	0.273
	Logística	0.426	0.169	**0.019**	0.000	0.000
RFE	R-part	0.850	**0.823**	0.838	0.780	0.821
	3NN	0.423	0.269	0.242	0.188	**0.161**
	Logística	0.392	0.115	**0.004**	0.000	0.000
Relief	R-part	0.753	**0.777**	0.788	0.753	0.773
	3NN	0.373	0.303	**0.234**	0.227	0.262
	Logística	0.362	0.104	**0.046**	0.000	0.000

* Los valores en negrita indican el número de componentes con el que se obtuvo el menor error de mala clasificación en la muestra de entrenamiento.

Figura 4.16 Tendencia de EMC en Brain con algoritmo CPS2



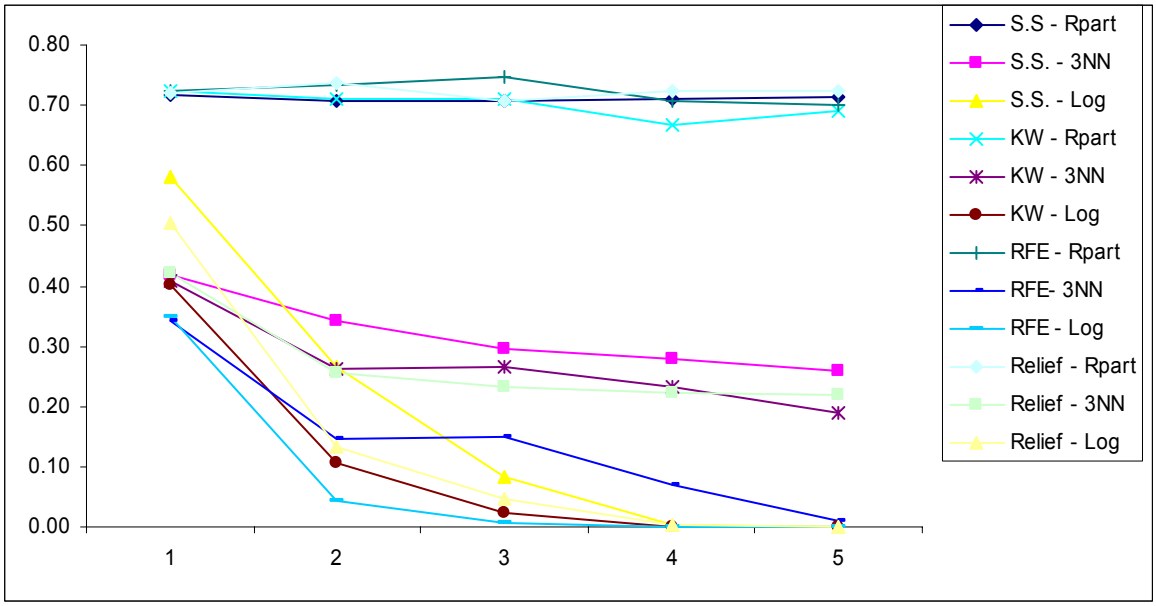
i.1) Conjunto de datos NCI: Algoritmo CPS1

TABLA 4.21 EMC estimado en NCI mediante CPS1

Métodos de Selección	Clasificador	Número de CPS				
		1	2	3	4	5
Sin selección	R-part	0.718	**0.708**	0.707	0.709	0.715
	3NN	0.417	0.341	**0.296**	0.280	0.260
	Logística	0.580	0.264	0.082	0.003	**0.000**
KW	R-part	0.723	0.711	0.710	0.667	**0.690**
	3NN	0.409	0.263	0.267	0.232	**0.190**
	Logística	0.403	0.105	0.022	**0.001**	0.000
RFE	R-part	0.722	**0.735**	0.746	0.708	0.702
	3NN	0.341	0.146	0.150	0.069	**0.010**
	Logística	0.350	0.042	0.005	**0.000**	0.000
Relief	R-part	0.720	0.737	**0.707**	0.722	0.724
	3NN	0.420	0.255	0.234	**0.222**	0.220
	Logística	0.503	0.132	0.048	0.003	**0.000**

* Los valores en negrita indican el número de componentes con el que se obtuvo el menor error de mala clasificación en la muestra de entrenamiento.

Figura 4.17 Tendencia de EMC en NCI con algoritmo CPS1



i.2) Conjunto de datos NCI: Algoritmo CPS2

TABLA 4.22 EMC estimado en NCI mediante CPS2

Métodos de Selección	Clasificador	Número de CPS				
		1	**2**	**3**	**4**	**5**
K-W	R-part	**0.704**	0.707	0.750	0.726	0.709
	3NN	0.438	0.350	0.304	0.321	**0.309**
	Logística	0.573	0.242	**0.100**	0.014	0.000
RFE	R-part	**0.697**	0.719	0.738	0.723	0.709
	3NN	0.454	0.385	0.326	**0.295**	0.254
	Logística	0.559	0.267	**0.119**	0.000	0.000
Relief	R-part	0.711	**0.685**	0.707	0.738	0.726
	3NN	0.467	0.355	0.302	**0.264**	0.302
	Logística	0.531	0.286	**0.076**	0.033	0.000

* Los valores en negrita indican el número de componentes con el que se obtuvo el menor error de mala clasificación en la muestra de entrenamiento.

Figura 4.18 Tendencia de EMC en NCI con algoritmo CPS2

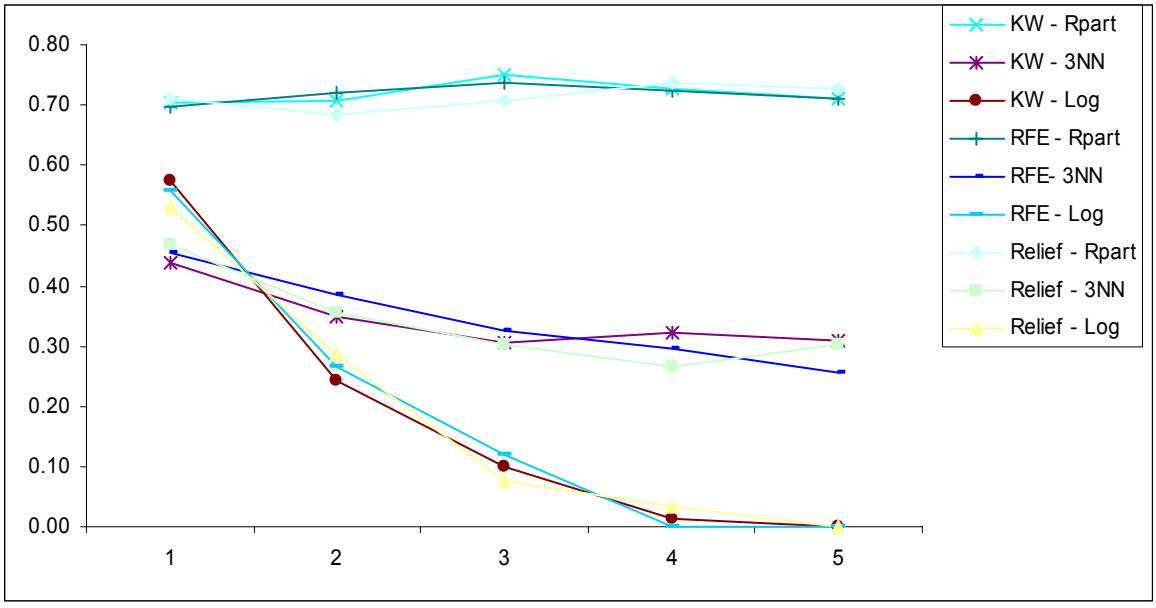


TABLA 4.23 Resumen de Resultados: EMC estimado con Algoritmo CPS1

Conjunto de datos	KW			RFE			Relief		
	Rpart	3NN	Log	Rpart	3NN	Log	Rpart	3NN	Log
Colon	0.089 [4]	0.111 [3]	0.054 [4]	0.084 [2]	0.059 [4]	**0.014 [3]**	0.112 [5]	0.106 [2]	0.083 [3]
Leukemia	0.006 [1]	0.010 [1]	0.001 [2]	**0.000 [1]**	**0.000 [1]**	**0.000 [1]**	**0.000 [1]**	0.002 [1]	**0.000 [1]**
Prostate	0.093 [1]	0.063 [3]	0.031 [4]	0.016 [3]	0.008 [3]	**0.000 [3]**	0.030 [2]	0.032 [2]	0.037 [2]
Carcinoma	0.517 [2]	0.012 [3]	0.000 [1]	0.508 [1]	**0.000 [1]**	**0.000 [1]**	0.513 [3]	**0.000 [1]**	**0.000 [1]**
BRCA	0.571 [1]	0.122 [2]	**0.000 [2]**	0.571 [1]	0.031 [2]	**0.000 [2]**	0.571 [1]	0.160 [3]	0.000 [3]
Lymphoma	0.142 [2]	0.009 [5]	0.000 [3]	0.142 [2]	**0.000 [1]**	**0.000 [1]**	0.142 [2]	0.000 [2]	**0.000 [1]**
SRBCT	0.319 [1]	0.001 [4]	**0.000 [2]**	0.318 [4]	0.001 [4]	**0.000 [2]**	0.330 [2]	0.004 [3]	0.000 [3]
Brain	0.766 [2]	0.135 [4]	0.000 [3]	0.773 [3]	0.053 [5]	**0.000 [2]**	0.787 [4]	0.190 [4]	0.000 [4]
NCI	0.690 [5]	0.190 [5]	0.001 [4]	0.735 [2]	0.010 [5]	**0.000 [4]**	0.707 [3]	0.222 [4]	0.000 [5]

TABLA 4.24 Resumen de Resultados: EMC estimado con Algoritmo CPS2

Conjunto de datos	KW			RFE			Relief		
	Rpart	3NN	Log	Rpart	3NN	Log	Rpart	3NN	Log
Colon	0.126 [3]	0.124 [2]	0.128 [1]	0.141 [4]	0.144 [2]	**0.017 [1]**	0.109 [4]	0.126 [2]	0.117 [1]
Leukemia	0.012 [4]	0.023 [3]	**0.000 [3]**	0.010 [1]	0.013 [1]	0.002 [3]	0.008 [3]	0.023 [2]	0.000 [4]
Prostate	0.112 [4]	0.080 [5]	0.067 [3]	0.100 [4]	0.060 [2]	**0.051 [3]**	0.081 [4]	0.073 [2]	0.069 [3]
Carcinoma	0.521 [3]	0.026 [1]	**0.000 [1]**	0.483[5]	0.028 [1]	0.000 [2]	0.485 [1]	0.037 [1]	0.000 [2]
BRCA	0.571 [3]	0.257 [2]	0.036 [3]	0.571 [4]	0.235 [3]	0.021 [3]	0.571 [2]	0.292 [2]	**0.014 [3]**
Lymphoma	0.142 [2]	0.033 [5]	0.007 [3]	0.143 [2]	0.002 [2]	**0.000 [2]**	0.143 [4]	0.027 [1]	**0.000 [2]**
SRBCT	0.318 [2]	0.009 [3]	0.009 [2]	0.320 [4]	0.013 [3]	0.045 [2]	0.329 [3]	0.007 [4]	**0.005 [3]**
Brain	0.769 [5]	0.207 [4]	0.019 [3]	0.823 [2]	0.188 [4]	**0.004 [3]**	0.753 [1]	0.262 [5]	0.046 [3]
NCI	0.704 [1]	0.309 [5]	0.100 [3]	0.697 [1]	0.295 [4]	0.119 [3]	0.685 [2]	0.264 [3]	**0.076 [3]**

* El valor entre corchetes indica el número de componentes principales supervisados utilizados para estimar el error de mala clasificación.

* Los valores en negrita indican la combinación con la que se obtuvo el menor error de mala clasificación para ese conjunto de datos.

El segundo algoritmo efectúa un mayor gasto computacional dado que realiza una selección de variables para cada una de las repeticiones. Para este algoritmo, la combinación que implica el mayor tiempo de procesamiento se da cuando se utiliza el método de selección de variables Relief, y se aplica el clasificador Rpart.

Según la mayoría de resultados obtenidos, los métodos Kruskal-Wallis, RFE y Relief realizan una aceptable selección de variables.

Por otro lado, cabe mencionar la rapidez de procesamiento cuando se utiliza el método RFE, y se emplea el clasificador KNN. Así mismo, cuando se hace uso de este método de selección de variable, con el clasificador de regresión logística nominal se obtiene el menor error menor de clasificación bajo el algoritmo CPS1.

En cuanto a los clasificadores utilizados, a pesar de las críticas que se le hace al Rpart, éste brindó buenos resultados en algunos conjuntos de datos. Pero obtuvo malos resultados sobre todo en los conjuntos que presentan más de dos clases o pocas observaciones como por ejemplo en los conjuntos de datos: Carcinoma, BRCA, Lymphoma, SRBCT, Brain y NCI. Los clasificadores KNN y Regresión Logística Nominal son los que presentaron los menores errores de mala clasificación, destacándose más aún entre estos dos el de Regresión Logística Nominal.

En general, es lógico pensar que a mayor cantidad de componentes utilizados, menor será el error de mala clasificación estimado (el error de mala clasificación tiende a disminuir), debido a que el porcentaje de variabilidad explicada es mayor. Es decir, se utiliza más

información para estimar el error de mala clasificación. La tendencia de disminución del error de mala clasificación es visible cuando no se realiza selección de variables. Sin embargo, cuando se realiza selección de variables, esta tendencia no ocurre en todos los conjuntos de datos utilizados, pues en muchas de ellas se obtuvo la menor tasa de error utilizando menos de tres CPS.

A pesar de eso, sobre todo en aquellos conjuntos de datos que presentaron un error de mala clasificación alto, se realizó la experimentación con más de cinco componentes, encontrándose que el error de mala clasificación no disminuye considerablemente. Un criterio estadístico que permite corroborar tal afirmación, consiste en elaborar una tabla de contingencia que permita probar si la proporción de observaciones mal clasificadas difiere al utilizar distintas cantidades de componentes y posteriormente aplicar una prueba X^2.

Como se puede observar en la Tabla 4.23 (para el algoritmo CPS1) la combinación con la que se obtiene el menor error de mala clasificación es cuando se utiliza el método de selección de variables RFE y el clasificador de Regresión Logística Nominal.

Para el algoritmo CPS2 (Tabla 4.24) con los métodos de selección de variables RFE y Relief, empleando el clasificador de Regresión Logística se obtienen los mejores resultados.

En resumen, la combinación que brinda mejores resultados se da cuando se utiliza el primer algoritmo, con el método de selección RFE, y el clasificador de Regresión Logística.

5 CONCLUSIONES Y RECOMENDACIONES

5.1 Conclusiones

- Mediante el presente trabajo se pretende implementar una metodología que muestre una forma correcta de estimar el error de mala clasificación en datos provenientes de expresión genética. Este método combina la selección y extracción de variables e involucra la división de la muestra en una parte de entrenamiento y otra de prueba, y sólo en esta última estimar el error de mala clasificación.

- Se presentaron dos algoritmos denominados CPS1 y CPS2. Con el primer algoritmo se obtuvo menores errores de mala clasificación en la mayoría de conjuntos de datos utilizados. Así mismo debido a la estructura que presenta el algoritmo CPS1 es computacionalmente más rápido con respecto al algoritmo CPS2.

- En general, el hecho de seleccionar variables hace que el tiempo de procesamiento y el error de mala clasificación se reduzca.

- Los métodos de selección de variables de envoltura (excepto RFE), seleccionan pocas variables, y su tiempo de procesamiento es muy elevado.

- El clasificador Rpart para datos provenientes de expresión genética no es generalmente muy bueno en conjunto de datos que presentan pocas observaciones y muchas clases.

- La combinación que minimiza el error de mala clasificación se da cuando se utiliza el algoritmo CPS1 con el método de selección de variables RFE y el clasificador de regresión logística.

5.2 Recomendaciones

- El interés por encontrar un método que trate de reducir lo máximo posible el error de mala clasificación, no nos debe hacer olvidar que, antes de utilizar alguna prueba estadística se necesita obligatoriamente la verificación de algunos supuestos (sobre todo en pruebas de tipo paramétrica correspondiente a algunos métodos de selección de variables o en el uso de clasificadores). Si estos supuestos no se cumplen, los resultados de la prueba no deberían ser considerados como válidos. Por lo tanto, se recomienda que no se debe perder un fundamento importante en estadística, que indica que las conclusiones son el reflejo del análisis previo de la información.

- La mayoría de conjuntos de datos que se utilizaron en la presente tesis, ya se encontraban preprocesadas. Parte del preprocesamiento implica la estandarización por filas (individuos), por lo que se recomienda verificar esto en los conjuntos de datos

antes de realizar algún análisis.

- Se recomienda utilizar otros métodos de selección de variables y otros tipos de clasificadores para estimar el error de mala clasificación con Componentes Principales Supervisados.

- Nosotros fijamos en $p_1 = 100$ el número de variables seleccionadas; sin embargo este valor puede ser modificado, con la finalidad de reducir el error de mala clasificación.

- Si bien es cierto que se presentó que los métodos de envoltura (SFS y SFFS) seleccionan pocas variables, sería recomendable probar si las variables seleccionadas son buenos genes marcadores. Es decir utilizar solo esos genes para estimar el error de mala clasificación.

REFERENCIAS

[1] Acuña, E. (2005). Curso de Data Mining. Universidad de Puerto Rico – Recinto Universitario de Mayagüez.

[2] Acuña, E. (2002). Statistical Methods for Microarray data. Universidad de Puerto Rico – Recinto Universitario de Mayagüez.

[3] Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moorre, T., Hudson, J., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Wesenburger, D. D., Armitage, J. O., Warnke, R., Levy, R., Wilson, W., Grever, M. R., Byrd, J. C., Botstein, D., Brown, P. O., Staudt, L. M. (2000). Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. Nature 403, 503-511.

[4] Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., Levine, A. (1999). Broad patters of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proceedings of the National Academy of Sciences 96, 6745-6750.

[5] Ambroise, C., McLachlan, G. (2002). Selection bias in gene extraction on the basis of microarray gene-expression data. PNAS vol. 99, 6562-6566.

[6] Bair, E., Hastie, T., Debashis, P., y Tibshirani, R. (2004). Prediction by supervised principal components. Technical Report. Department of Statistics, Stanford University.

[7] Bair, E. y Tibshirani, R. (2004). Semi-Supervised Methods to Predict Patient Survival from Gene Expression Data. PLoS Biology, Volume 2 Issue 4. Pag. 511-522.

[8] Ben-Dor, A., Bruhn, L., Friedman, N., Nachman, I., Schummer, M., Yakhini, Z. (2000). Tissue classification with gene expression profiles. Journal of Computational Biology 7, 559-584.

[9] Boulesteix, A.L. (2004). Dimension Reduction and Classification with High-Dimensional Microarray Data. Dissertation an der Fakultät für Mathematik, Informatik und Statistik der Ludwig-Maximilian-Universität München.

[10] Braga-Neto, U., Dougherty, E. R. (2004). Is cross-validation valid for small-sample microarray classification?. Bioinformatics 20, 2465-2472.

[11] Coaquira, F. (2002). Selección de Variables para Clasificación Supervisada. Tesis para optar al título de Magíster en Matemática – Especialidad Estadística. Universidad de Puerto Rico – Recinto Universitario de Mayagüez.

[12] Dettling, M. (2004). BagBoosting for Tumor Classification with Gene Expression Data. Seminar für Statistik ETH Zürich CH-8092 Switzerland.

[13] Dettling, M. y Bühlmann, P., (2002). Supervised clustering of genes. Seminar für Statistik, Eidgenössische Technische Hochschule (ETH) Zürich, 8092 Zürich, Switzerland.

[14] Draghici, S. (2003). Data Analysis Tools for DNA Microarrays. Chapman & Hall / CRC Mathematical Biology and Medicine Series.

[15] Dudoit, S., Fridlyand, J., Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. Journal of the American Statistical Association 97, 77-87.

[16] Fort, G. y Lambert-Lacroix S. (2005). Classification using Partial Least Squares with penalized logistic regression. Bioinformatics 21, 1104-1111.

[17] Furey, Y. S., Cristianini, N., Duffy, N,, Bednarski, D. W., Schummer, M , Haussler, D. (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 16, 906-914.

[18] Garrett-Mayer, E. y Parmigiani, G. (2004). Clustering and Classification Methods for Gene Expression Data Analysis. Johns Hopkins University, Department of Biostatistics Working Papers.

[19] Golub, T., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J., Caligiuri, M. A., Bloomfield, C. D., Lander, E. S., (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286, 531-537.

[20] Guyon, I. y Elisseeff, A. (2003). An introduction to Variable and Feature Selection. Journal of Machine Learning Research 3 1157-1182.

[21] Guyon, I., Weston, J., Barhill, S., Vapnik, W. (2001). Gene Selection for Cancer Classification using Support Vector Machines. Barnhill Bioinformatics, Savannah, Georgia, USA.

[22] Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Raffeld, M., Yakhini, Z., Ben-Dor, A., Dougherty, E., Kononen, J., Bubendorf, L., Fehrle, W., Pittaluga, S., Gruvberger, S., Loman, N., Johannsson, O., Olsson, H., Wilfond, B., sauter, G., Kallioniemi, O. P., Borg, A., Trent, J., (2001). Gene-expression profiles in hereditary breast cancer. New England Journal of medicine 344, 539-548.

[23] Johnson, D. (2000). Métodos Multivariados aplicados al Análisis de Datos. Internacional Thompson Editores S.A. de C.V.

[24] Kahn, J., Wei, J. S., Ringner, M., Saal, L. H., Ladanyi., Westermann, F., Berthold,

F., Schwab, M., Antonescu, C. R., Peterson, C., Meltzer, P. S. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature Medicine 7, 673-679.

[25] Kohavi, R y John, G. (1997). Wrappers for feature selection. Artificial Intelligence, 97 (1-2): 273-324.

[26] Notterman, D. A., Alon, U., Sierk, A. J., Levine, A. J., (2001). Trancriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays. Cancer Research 61, 3124-3130.

[27] Park, P.; Pagano, M. y Bonetti M. (2001). A Nonparametric scoring algorithm for identifying informative genes from microarray data. Department of Biostatistics, Harvard School of Public Health.

[28] Pomeroy, S., Tamayo, P., Gaasenbeek, M., Sturla, L., Angelo, M., McLaughlin, M., Kim, J., Goumnerova, L., Black, P., Lau, C. et al. (2002). Prediction of Central Nervous System Embryonal Tumor Outcome based on Gene Expression. Nature, 415, 436-442.

[29] Ross, D, Scherf, U., Eisen, M. Perou, C. Spellman, P., Iyer, V., Jeffrey, S., de Rijn, M. Waltham, M. Pergamenschikov, A., Lee, J., Lashkari, D., Shalon, D., Myers, T., Weinstein, J., Botstein, D., Brown, P., (2000). Systematic variation in gene expression patterns in human cancer cell lines. Nature Genetics 24, 227-234.

[30] Selvin, S. (1998). Modern Applied Biostatistical Methods Using S-Plus. University of California, Berkeley. Monographs in Epidemiology and Biostatistics Volume 28.

[31] Simon, R., Korn, E., McShane, L. Radmacher, M., Wright, G., Zhao, Y. (2003). Design and Analysis of DNA Microarray Investigations Springer-Verlag, New York.

[32] Singh, D., Febbo, P. G.,, Ross, K., Jackson, D. G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. A., D'Amico, A. V., Richie, J. P., Lander, E. S., Loda, M., Kantoff, P. W., Golub, T.R., Sellers, W. R., (2002). Gene expression correlates of clinical prostate cancer behavior. Cancer Cell 1, 203-209.

[33] Smith, L. (2002). A tutorial on Principal Components Analysis. http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf

[34] Tibshirani R. y Efron, B. (2002). Pre-validación e inferencia en microarreglos. Statistical Applications in Genetics and Molecular Biology. Volume 1, Issue 1, article 1. Pag 1-18.

[35] Tusell, F. (2004). Análisis Multivariante. Universidad del País Vasco. http://etpx22.bs.ehu.es/~etptupaf/pub/papiros/multi/multi.pdf.

[36] Vega, J. (2004). Generalizaciones de Mínimos Cuadrados Parciales con aplicación en clasificación supervisada. Tesis para optar al titulo de Doctor en Ciencias e Ingeniería de la Información y Computación. Universidad de Puerto Rico – Recinto Universitario de Mayagüez.

[37] Wang, Y., Makedon, F., Ford, J. Pearlman, J. (2004). HykGene: An Hybrid Approach for Selecting Marker Genes for Phenotype Classification using Microarray Gene Expression Data. Bioinformatics Vol 01 Pages 1-10.

[38] Weston, J., Mukherjee, S., Chapelle, O., Pontil, M., Poggio, T. y Vapnik, V. (2000). Feature selection for SVMs. In NIPS 13.

[39] Yeung, K.Y. y Ruzzo, W.L. (2001). An empirical study of Component Principal Analysis for clustering gene expression data. Bioinformatics, Vol. 17 no. 9, pp. 763-774.

[40] Yu, L., Liu, H. (2004). Redundancy Based Feature Selection for Microarray Data.

Research Track Poster. Pag. 737-742.

[41] Zheng Lihong y He Xiangjian (2005). Classification Techniques in Pattern Recognition. Faculty of IT, University of Technology, Sydney, Australia.

[42] Zhu, Ji y Hastie, Trevor (2004). Classification of Gene Microarrays by Penalized Logistic Regression.

APÉNDICE A. PROGRAMAS AUXILIARES

APÉNDICE A1 FUNCIÓN: SELECCIÓN DE VARIABLES PARA CPS1

Descripción: Realiza la selección de variables mediante las pruebas: Kruskal-Wallis, RFE y Relief. La función necesita como argumentos: la matriz de datos, el tipo de prueba que se realizará y el número de variables que serán seleccionadas. Esta función muestra como resultados: los índices de las variables que han sido seleccionadas y la matriz de data reducida, es decir la matriz *X* solo conteniendo las variables seleccionadas. Para desarrollar esta librería se hizo uso de las funciones: kruskal.test de la librería stats; reliefcont de la librería dprep (elaborada por Acuña y Rodríguez (2005)) y rfe.fit de la librería rfe (Ambroise y McLachlan (2005)).

```
varselection=function(data,metodo=c("kw","rfe","relief"),numvarsel=100)
{
 library(rfe)
 library(dprep)
 n=dim(data)[1]
 p=dim(data)[2]-1
 pvalue=matrix(0,p,1)
 predic = data[,1:p]
 clas = data[,(p+1)]

 if (metodo == "kw")
 {
  for (j in 1:p)
   {
    pvalue[j]=kruskal.test(data[,j]~data[,(p+1)])$p.value
   }

 orden=sort(pvalue,index.return = TRUE)$ix
 var.selec = orden[1:numvarsel]
 pred.red  = predic[,var.selec]
 data.red = cbind(pred.red,clas)
 }
```

```
if (metodo == "rfe")
{
var.selec1 = rfe.fit(predic,as.factor(clas),minf=numvarsel,speed="high")$Flist
var.selec = var.selec1[1:numvarsel]
pred.red = predic[,var.selec]
data.red = cbind(pred.red,clas)
}

if (metodo == "relief")
{
var.selec = reliefcont(data,n,0)
var.selec = var.selec[1:numvarsel]
pred.red = predic[,var.selec]
data.red = cbind(pred.red,clas)
}

return(list(var.selec = var.selec, data.red = data.red))

}
```

APÉNDICE A2 FUNCIÓN: SELECCIÓN DE VARIABLES PARA CPS2

Descripción: Realiza la selección de variables mediante las pruebas de Kruskal-Wallis, RFE y Relief. La función necesita como argumentos, la matriz de data de entrenamiento, matriz de data de prueba, el tipo de prueba que se realizará y el número de variables que serán seleccionadas. Esta función muestra como resultados: los índices de las variables seleccionadas. Para elaborar esta función se hizo uso de las mismas funciones requeridas en la función anterior

```
varselection=function(datae,datap,metodo=c("kw","rfe","relief"),numvarsel=10)
{
library(rfe)
library(dprep)
n=dim(datae)[1]
p=dim(datae)[2]-1
pvalue=matrix(0,p,1)
predice = datae[,1:p]
```

```
predicp = datap[,1:p]
clase = datae[,(p+1)]
clasp = datap[,(p+1)]

if (metodo == "kw")
{
 for (j in 1:p)
  {
   pvalue[j]=kruskal.test(datae[,j]~datae[,(p+1)])$p.value
  }
 orden=sort(pvalue,index.return = TRUE)$ix
 var.selec = orden[1:numvarsel]
 predered  = predice[,var.selec]
 predpred  = predicp[,var.selec]
 datae.red = cbind(predered,clase)
 datap.red = cbind(predpred,clasp)
}

if (metodo == "rfe")
{
var.selec=rfe.fit(predice,as.factor(clase),minf=numvarsel,speed="high")$Flist
var.selec=var.selec[1:numvarsel]
predered = predice[,var.selec]
predpred = predicp[,var.selec]
datae.red = cbind(predered,clase)
datap.red = cbind(predpred,clasp)
}

 if (metodo == "relief")
{
var.selec = reliefm(datae,n,0)
var.selec = var.selec[1:numvarsel]
predered = predice[,var.selec]
predpred = predicp[,var.selec]
datae.red = cbind(predered,clase)
datap.red = cbind(predpred,clasp)
}

 return(list(var.selec = var.selec))
}
```

APÉNDICE A3 FUNCIÓN: ERROR DE MALA CLASIFICACION EN LA MUESTRA DE PRUEBA

Descripción: Es utilizada como función auxiliar para los programas cps1 y cps2. Realiza la estimación del error de mala clasificación en la muestra de prueba mediante el uso de tres clasificadores: rpart, knn y logística. Como argumentos se requiere la data de prueba, el tipo de clasificador y en el caso que este sea el clasificador knn se requiere el número de vecinos más cercanos (k). En este programa se utilizó la función rpart de la librería del mismo nombre elaborada por Terry M. Therneau y Beth Atkinson; la función knn de la librería class, que fue elaborada por Venables y Ripley y la función multinom de la librería nnet elaborada también por Venables y Ripley.

```
errorap=function(data,clasificador=c("rpart","knn","logistica"),k=3)
{
 library(rpart)
 library(class)
 library(nnet)

 p=dim(data)[2]-1
 nombres=colnames(data)
 f1=as.formula(paste(nombres[p+1],".",sep="~"))
 pred=data[,1:p]
 class=data[,(p+1)]

 if(clasificador == "rpart")
 {
  data=data.frame(data)
  arbol=rpart(f1,data=data,method="class")
  ajus=predict(arbol)
  pred=max.col(ajus)
  error=mean(pred!=class)
 }

 if(clasificador == "knn")
 {
 pred = knn(as.matrix(data[, 1:p]), as.matrix(data[,1:p]), data[,(p+1)], k=k,prob=T)
 error=mean(pred!=class)
 }
```

```
if(clasificador == "logistica")
{
data=data.frame(data)
pred = multinom(f1, data = data, MaxNWts = 7500)
pred1 = predict(pred, data)
error = mean(pred1 != as.numeric(data[, (p+1)]))
}

return(list(error = error))

}
```

APÉNDICE B PROGRAMAS PRINCIPALES

APÉNDICE B1 FUNCIÓN CPS1

Descripción: Realiza la estimación del error de mala clasificación mediante el algoritmo CPS1. Esta función necesita como argumentos: el conjunto de datos, el número de componentes principales supervisados que serán utilizados, el clasificador (en el caso de KNN el valor de los vecinos más cercanos *k*), el número de partes en la cual será dividida la muestra de prueba para obtener el *m* óptimo, y el número de muestras aleatorias distintas que se desean obtener (repeticiones). El programa brinda como resultado la variabilidad explicada por los componentes utilizados, el error medio de mala clasificación así como su desviación estándar. Cabe mencionar que dentro de este programa se usa la función *stratsrs* perteneciente a la librería pps, la cual fué elaborada por Jack Gambino (2004); las funciones crossval y cv10log de la librería dprep; la función knn de la librería class; la función rpart de la librería rpart y la función multinom de la librería nnet.

```
cps1=function(data,n.componentes=2,clasificador=c("rpart","knn","logistica"),k=3,n.parts=2,rep=10)
{
 library(pps)
 library(dprep)
 library(class)
 library(rpart)
 library(nnet)

 n=dim(data)[1]
 p=dim(data)[2]-1
```

```
strat=as.vector(data[,(p+1)])
frec=as.matrix(table(data[,(p+1)]))
lonfre=dim(frec)[1]
nh=rep(0,lonfre)
for(i in 1:lonfre)
 {
  nh[i]=round(2/3*frec[i,1])
 }
ne=sum(nh)
np=n-ne
indices=seq(1:n)
indicese=matrix(0,ne,rep)
indicesp=matrix(0,np,rep)
for (j in 1:rep)
 {
  indicese[,j]=sort(stratsrs(strat,nh))
  indicesp[,j]=indices[-indicese[,j]]
 }
 error=matrix(0,2,rep)
 for (j in 1:rep)
{
 datae = data[indicese[,j],]
 datap = data[indicesp[,j],]
  x.rede = datae[,1:p]
 x.redp = datap[,1:p]
 meane  = apply(x.rede, 2, mean)
 sde    = apply(x.rede,2,sd)
 x.redec = scale(x.rede,center=TRUE, scale=TRUE)
 x.svd = svd(x.redec)
 eiges = x.svd$d
 eigesc = x.svd$d[1:n.componentes]
 toteiges = sum(eiges)
 vari=sum(eiges[1:n.componentes])/toteiges
```

```
componentes=prcomp(x.rede,retx = TRUE, center = TRUE, scale = TRUE)
cargas=componentes$rotation[,1:n.componentes]
puntuacionese=componentes$x[,1:n.componentes]
puntuacionese=cbind(puntuacionese,datae[,(p+1)])

 if (clasificador == "rpart")
 {
  error[1,j] = crossval1(puntuacionese, nparts=n.parts, method="rpart", repet=1)
 }
 if (clasificador == "knn")
 {
  error[1,j] = crossval1(puntuacionese, nparts=n.parts, method="knn", kvec=k, repet=1)
 }
 if (clasificador == "logistica")
 {
  error[1,j] = logistica(puntuacionese, nparts=n.parts, repet=1)
 }
xcenp=scale(x.redp,center = meane,scale=sde)
puntuacionesp = xcenp%*%cargas
transp=cbind(puntuacionesp,datap[,(p+1)])
nombres=paste("CPS",1:n.componentes,sep="")
nombres=c(nombres,"CLASE")
colnames(transp) = nombres

 if (clasificador == "rpart")
 {
  error[2,j] = errorap(transp, clasificador="rpart")$error
 }
 if (clasificador == "knn")
 {
  error[2,j] = errorap(transp, clasificador="knn", k=k)$error
 }
 if (clasificador == "logistica")
```

```
  {
   error[2,j] = errorap(transp, clasificador="logistica")$error
  }

 }
  emcmean = apply(error,1,mean)
  emcsd = apply(error,1,sd)

  return(list( vari = vari, error = error, emcmean = emcmean, emcsd = emcsd))
}
```

APÉNDICE B1 FUNCIÓN CPS2

Descripción: Realiza la estimación del error de mala clasificación mediante el algoritmo CPS2. Esta función requiere como argumentos: el conjunto de datos, el número de componentes principales supervisados que serán utilizados, el método de selección de variables a realizarse, el número de variables que serán seleccionadas, el tipo de clasificador, el valor de *k* (el número de vecinos más cercanos, en el caso de utilizar el clasificador KNN) , el número de partes en que será dividida la muestra de entrenamiento al utilizar validación cruzada para estimar el error de mala clasificación y el número de repeticiones. La función da como resultados la variabilidad explicada por los componentes utilizados el error medio de mala clasificación de la muestra de prueba y la muestra de entrenamiento así como sus respectivas desviaciones estándar. Para elaborar esta función se utilizaron las mismas funciones de la función cps1.

```
cps2=function(data,n.componentes=2,selec=c("kw","rfe","relief"),numvarsel=10,clasificador=c("rpart
","knn","logistica"),k=3,n.parts=10,rep=50)

{
 library(pps)
 library(dprep)
 library(class)
 library(rpart)
 library(nnet)

 n=dim(data)[1]
 p=dim(data)[2]-1
 clas=data[,(p+1)]

 strat=as.vector(clas)
 frec=as.matrix(table(clas))
 lonfre=dim(frec)[1]
 nh=rep(0,lonfre)
 for(i in 1:lonfre)
  {
   nh[i]=round(2/3*frec[i,1])
  }
 ne=sum(nh)
 np=n-ne
 indices=seq(1:n)
 indicese=matrix(0,ne,rep)
 indicesp=matrix(0,np,rep)
 for (j in 1:rep)
  {
   indicese[,j]=sort(stratsrs(strat,nh))
   indicesp[,j]=indices[-indicese[,j]]
  }
 error=matrix(0,2,rep)
```

```
for (j in 1:rep)
{
 orden=sort(clas,index.return = TRUE)$ix
 datao=as.matrix(data[c(orden),])
 datared=varselection(datao[indicese[,j],],datao[indicesp[,j],],metodo=selec,numvarsel=numvarsel)
 dataerpred = datao[indicese[,j],datared$var.selec]
 dataeclas  = datao[indicese[,j],(p+1)]
 dataer  = cbind(dataerpred,dataeclas)
 dataprpred = datao[indicesp[,j],datared$var.selec]
 datapclas  = datao[indicesp[,j],(p+1)]
 datapr  = cbind(dataprpred,datapclas)
  x.rede=dataer[,1:numvarsel]
 x.redp=datapr[,1:numvarsel]
 meane=apply(x.rede, 2, mean)
 sde    = apply(x.rede,2,sd)

 x.redec = scale(x.rede,center=TRUE, scale=TRUE)
 x.svd = svd(x.redec)
 eiges = x.svd$d
 eigesc = x.svd$d[1:n.componentes]
 toteiges = sum(eiges)
 vari=sum(eiges[1:n.componentes])/toteiges
  componentes=prcomp(x.rede,retx = TRUE, center = TRUE, scale = TRUE)
 cargas=componentes$rotation[,1:n.componentes]
 puntuacionese=componentes$x[,1:n.componentes]
 puntuacionese=cbind(puntuacionese,dataer[,(numvarsel+1)])

 if (clasificador == "rpart")
  {
   error[1,j] = crossval1(puntuacionese, nparts=n.parts, method="rpart", repet=1)
  }

  if (clasificador == "knn")
```

```
 {
  error[1,j] = crossval1(puntuacionese, nparts=n.parts, method="knn", kvec=k, repet=1)
 }
 if (clasificador == "logistica")
 {
  error[1,j] = logistica(puntuacionese, nparts=n.parts, repet=1)
 }

xcenp=scale(x.redp,center = meane,scale=sde)
puntuacionesp = xcenp%*%cargas
transp=cbind(puntuacionesp,datapr[,(numvarsel+1)])
nombres=paste("CPS",1:n.componentes,sep="")
nombres=c(nombres,"CLASE")
colnames(transp) = nombres

 if (clasificador == "rpart")
 {
  error[2,j] = errorap(transp, clasificador="rpart")$error
 }
 if (clasificador == "knn")
 {
  error[2,j] = errorap(transp, clasificador="knn", k=k)$error
 }
 if (clasificador == "logistica")
 {
  error[2,j] = errorap(transp, clasificador="logistica")$error
 }
 }
 emcmean = apply(error,1,mean)
 emcsd = apply(error,1,sd)
    return(list(toteiges = toteiges, vari = vari, error = error, emcmean = emcmean, emcsd = emcsd))
}
```

APÉNDICE C ALGUNOS GRÁFICOS ADICIONALES

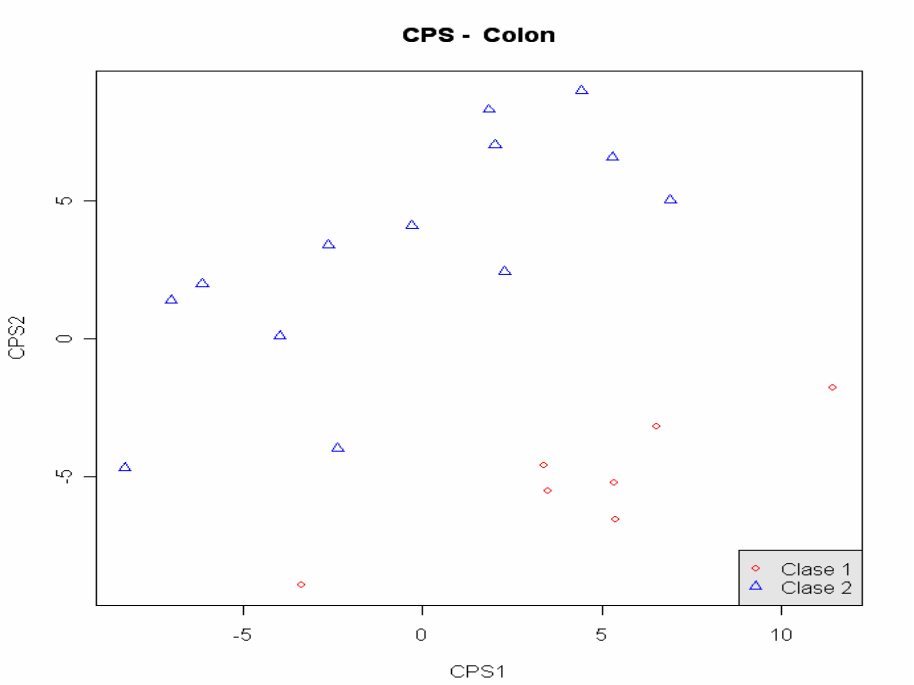


Figura 4.19 CPS variables seleccionadas con RFE – Colon – CPS1 – 2D

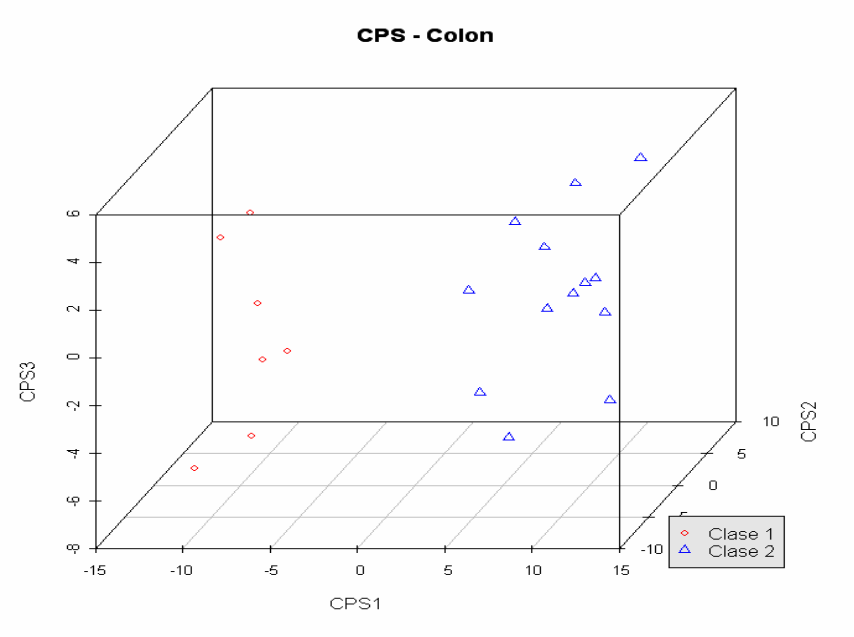


Figura 4.20 CPS variables seleccionadas con RFE - Colon – CPS1 – 3D

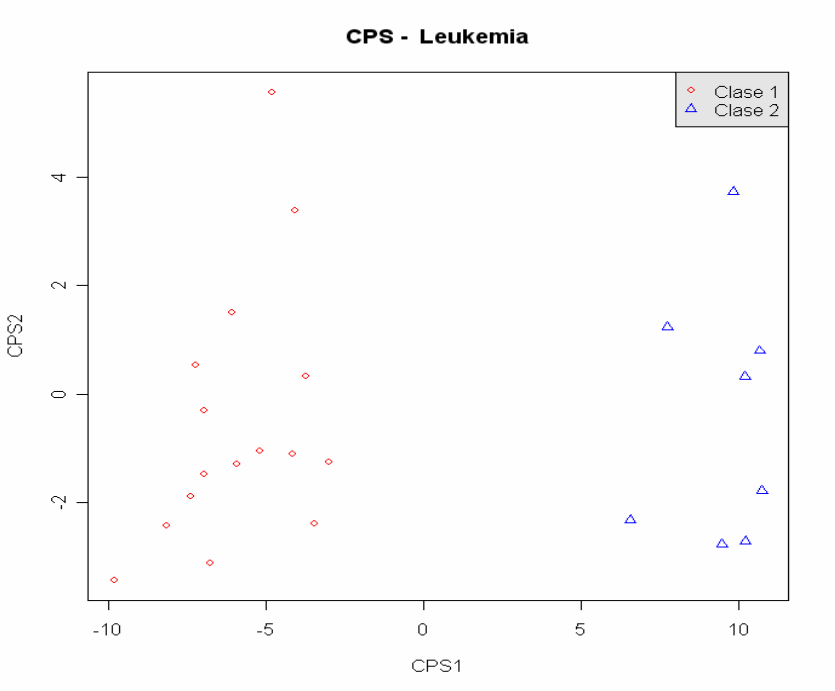


Figura 4.21 CPS variables seleccionadas con RFE – Leucemia – CPS1 – 2D

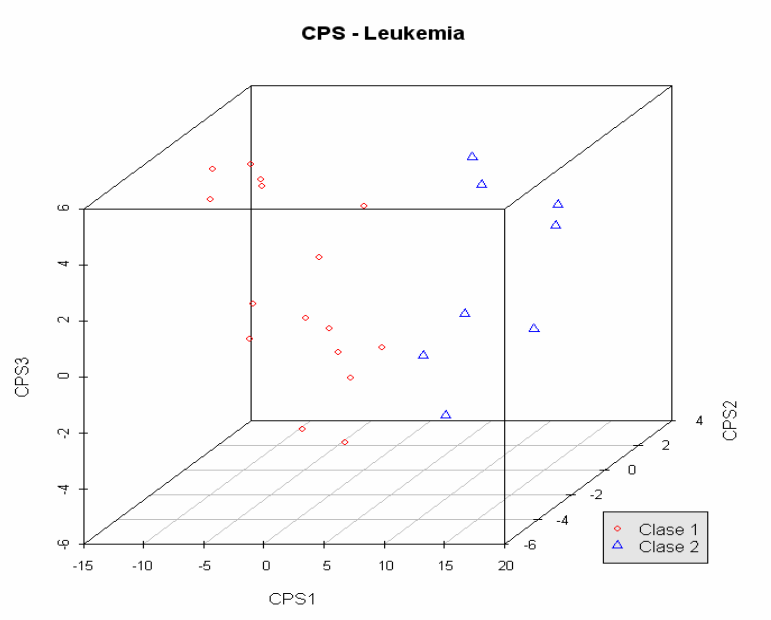


Figura 4.22 CPS variables seleccionadas con RFE – Leucemia – CPS1 – 3D

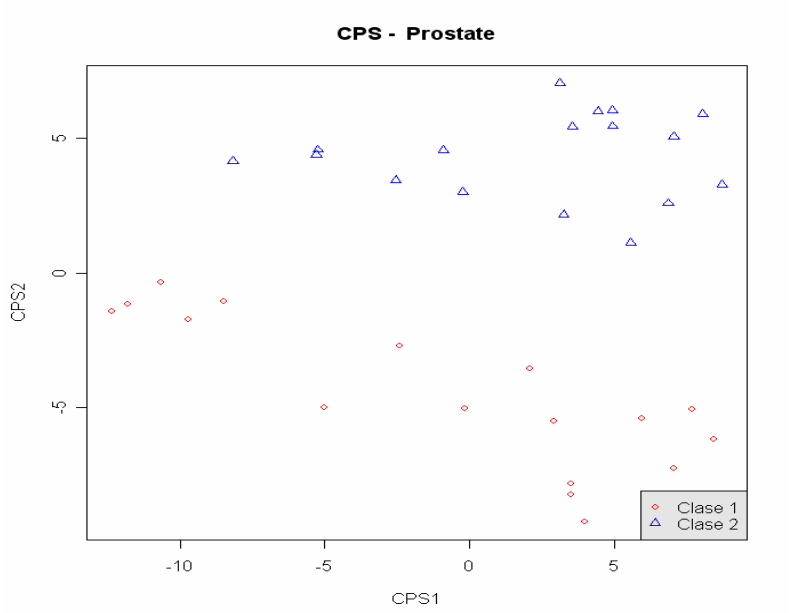


Figura 4.23 CPS variables seleccionadas con RFE – Prostate – CPS1 – 2D

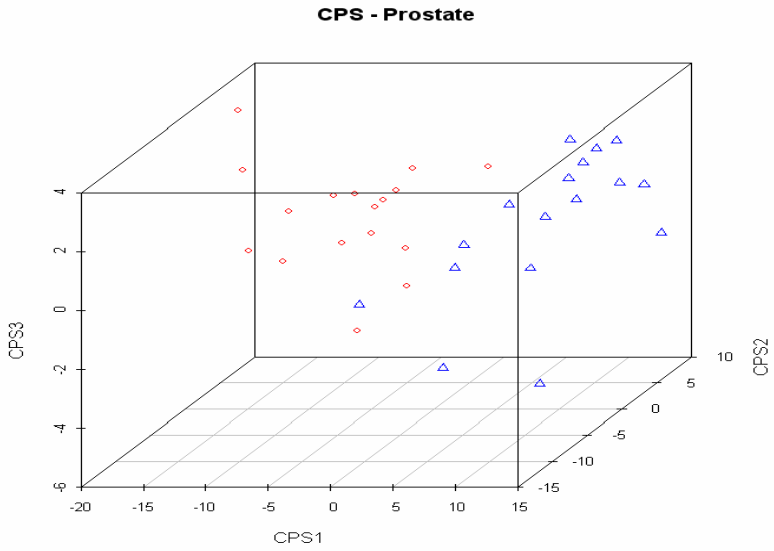


Figura 4.24 CPS variables seleccionadas con RFE – Prostate – CPS1 – 3D

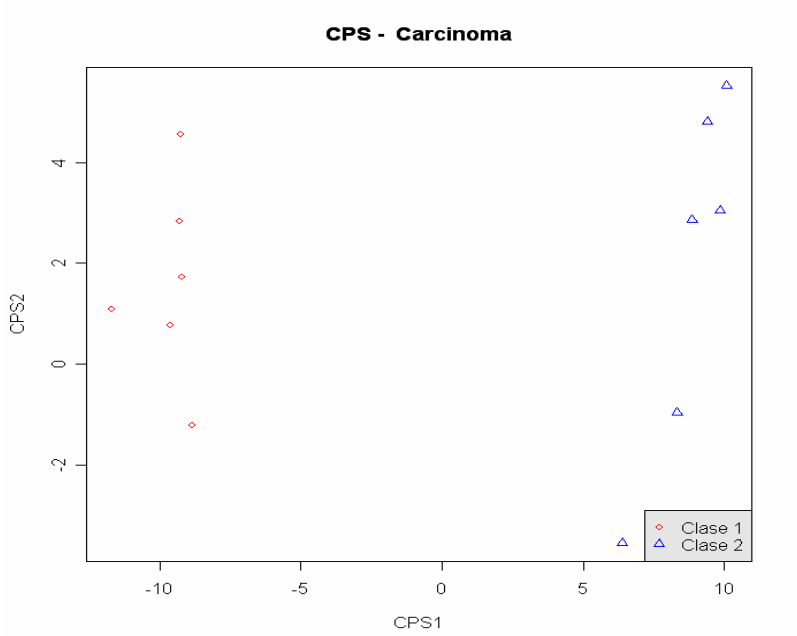


Figura 4.25 CPS variables seleccionadas con RFE – Carcinoma – CPS1 – 2D

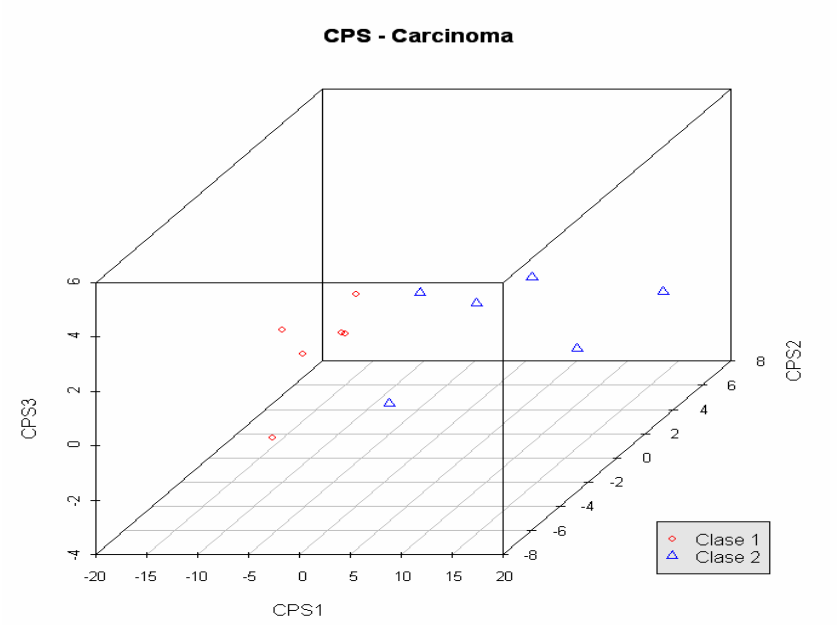


Figura 4.26 CPS variables seleccionadas con RFE – Carcinoma – CPS1 – 3D

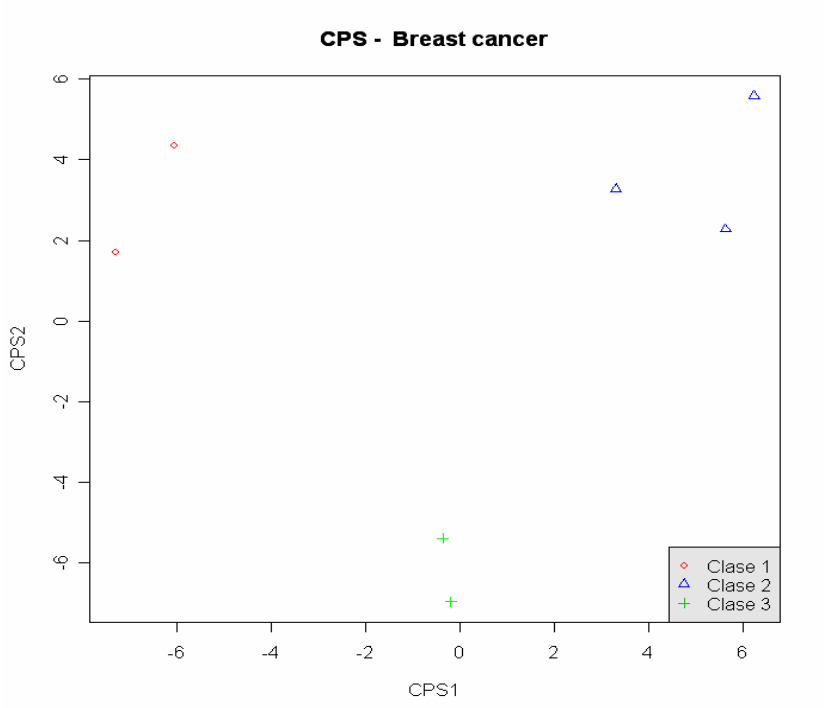


Figura 4.27 CPS variables seleccionadas con RFE – BRCA – CPS1 – 2D

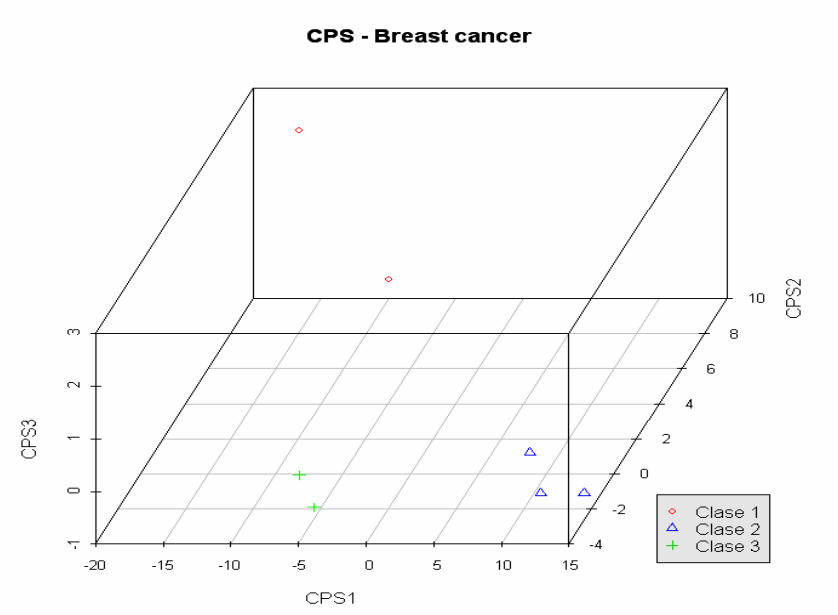


Figura 4.28 CPS variables seleccionadas con RFE – BRCA – CPS1 – 3D

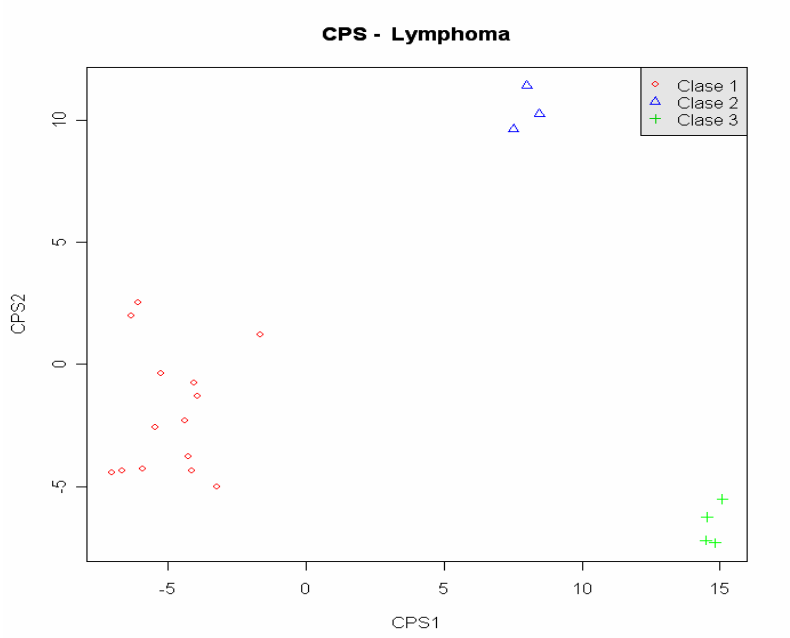


Figura 4.29 CPS variables seleccionadas con RFE – Lymphoma – CPS1 – 2D

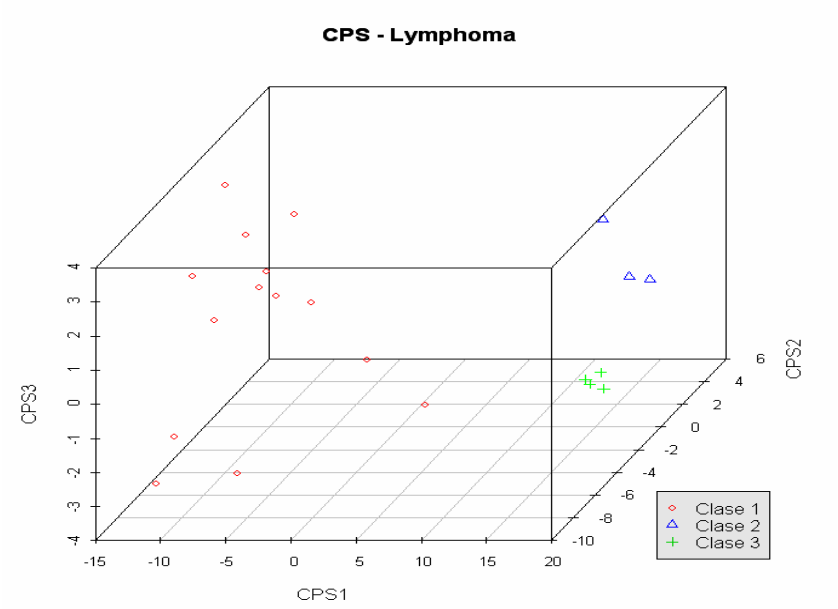


Figura 4.30 CPS variables seleccionadas con RFE – Lymphoma – CPS1 – 3D

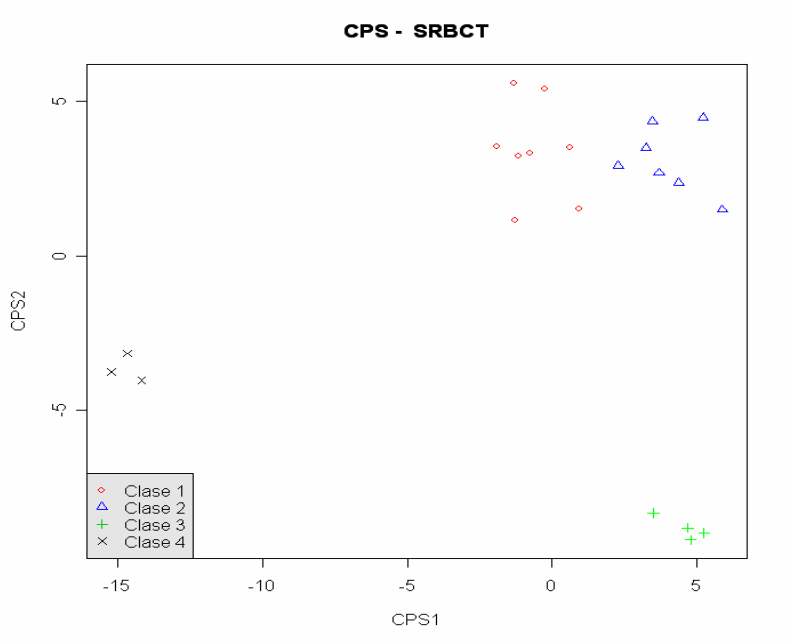


Figura 4.31 CPS variables seleccionadas con RFE – SRBCT – CPS1 – 2D

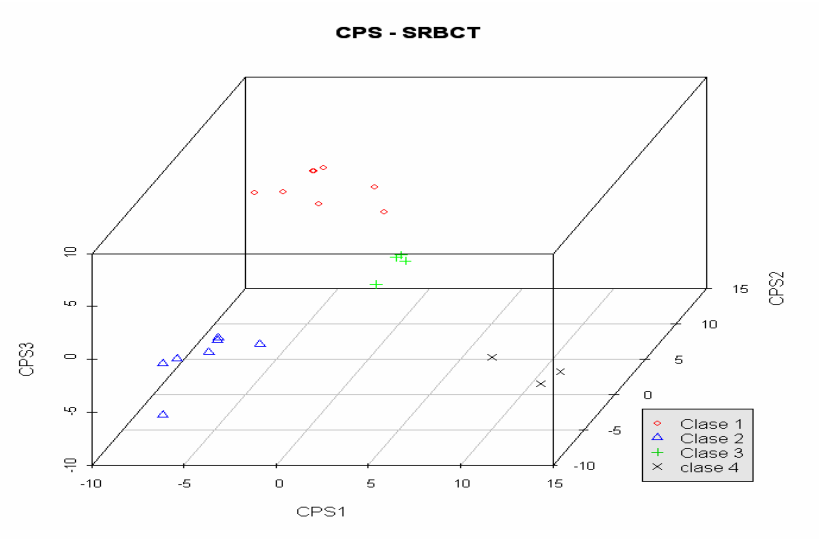


Figura 4.32 CPS variables seleccionadas con RFE – SRBCT – CPS1 – 3D

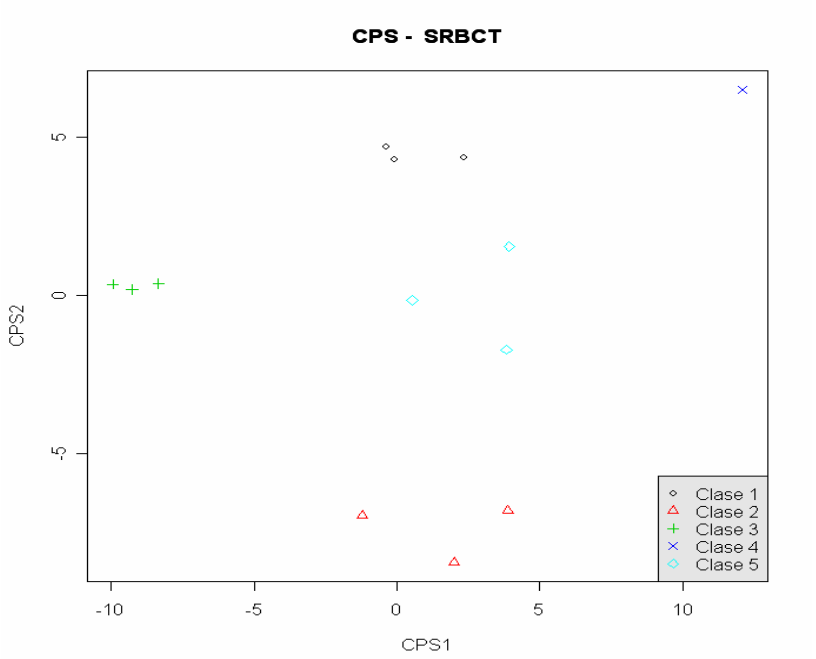


Figura 4.33 CPS variables seleccionadas con RFE – Brain – CPS1 – 2D

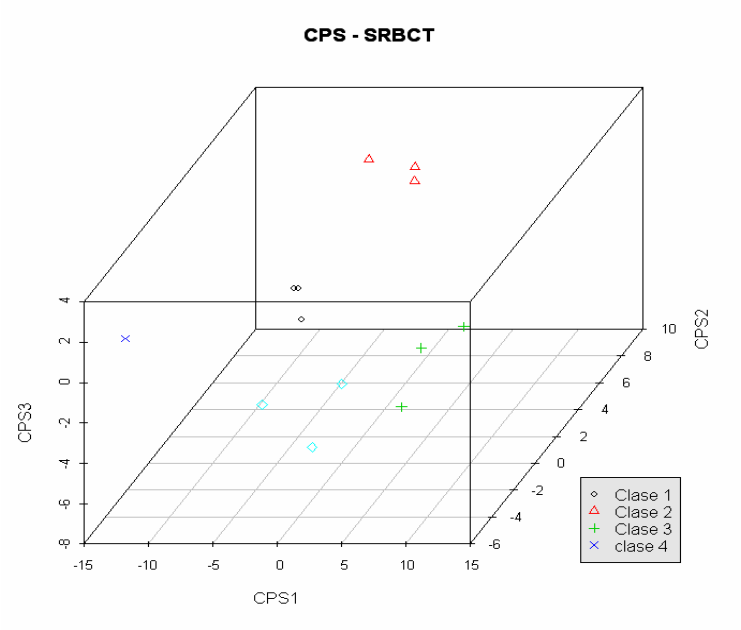


Figura 4.34 CPS variables seleccionadas con RFE – Brain – CPS1 – 3D

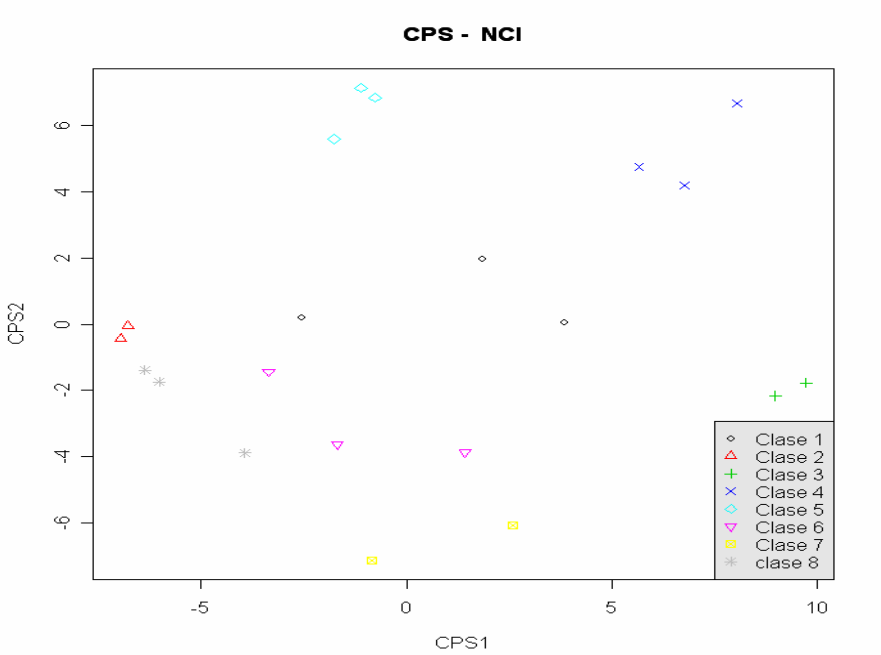


Figura 4.35 CPS variables seleccionadas con RFE – NCI – CPS1 – 2D

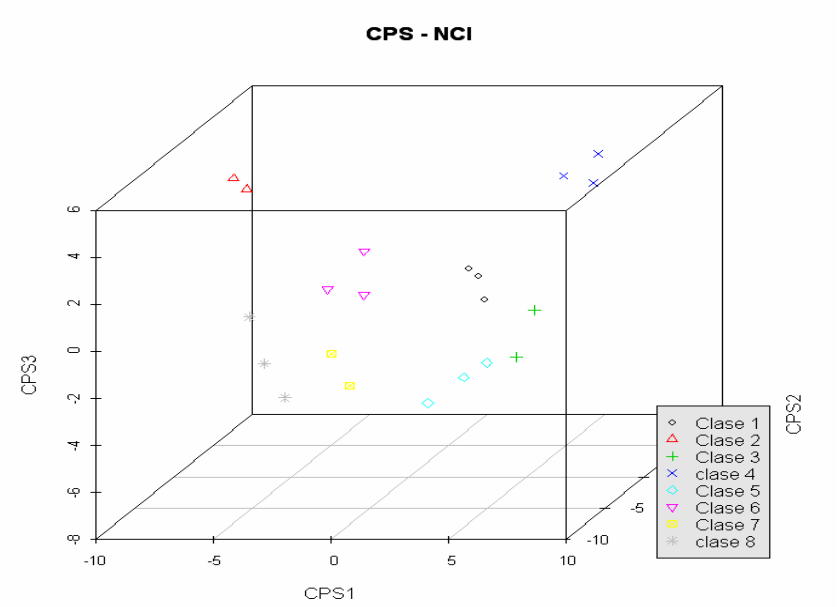


Figura 4.36 CPS variables seleccionadas con RFE – NCI – CPS1 – 3D