Redes neuronales artificiales a partir de la función de supervivencia de Kaplan-Meier

Luzardo B., Marianela
Chediak, Georges J.
Borges P., Rafael

Luzardo B., Marianela
Licenciada en Estadística, Magíster en Estadística Aplicada
Profesora Agregada de la Facultad de Ciencias Económicas y Sociales, Universidad de Los Andes.

Chediak, Georges J.
Ingeniero de Sistemas, Magíster en Estadística Aplicada
Profesor Asistente de la Facultad de Ingeniería, Universidad de Los Andes.

Borges P., Rafael
Licenciado en Ciencias Estadísticas
Magíster en Estadística Aplicada
Profesor Agregado de la Facultad de Ciencias Económicas y Sociales, Universidad de Los Andes.

Recibido: 28-11-07
Revisado: 05-06-08
Aceptado: 27-06-08

Hoy en día, las aplicaciones estadísticas computacionales incluyen módulos con técnicas avanzadas para el desarrollo de modelos que permiten simular el comportamiento de variables claves en la organización. El análisis de confiabilidad, o análisis de supervivencia, se define como un conjunto de técnicas que se encargan de analizar el tiempo transcurrido desde el origen bien definido hasta la ocurrencia de un evento de interés previamente establecido; a su vez, una red neuronal artificial (RNA) puede ser definida como un modelo matemático cuya construcción se lleva a cabo mediante un proceso que imita el funcionamiento de las redes neuronales biológicas, y puede ser usada para modelar fenómenos que involucran alguna respuesta que depende de algún conjunto de factores. Esta investigación aborda el análisis de supervivencia con técnicas de inteligencia artificial con la finalidad de estimar a partir de una RNA la función de supervivencia de Kaplan-Meier. Los resultados demuestran que los modelos de redes neuronales artificiales permiten el manejo de datos de supervivencia sin necesidad de imponer supuestos de partida en los mismos. Así, queda evidenciado el potencial de las RNA para evaluar la información parcial proveniente de un conjunto de datos censurados de supervivencia.
Palabras clave: Confiabilidad, Kaplan-Meier, redes neuronales artificiales.

RESUMEN

Nowadays, the statistical applications include modules with advanced technologies for the models' development that allow the simulation of the behaviour of key variables in the organization. The reliability analysis, or survival analysis, is defined as a set of techniques that analyze the elapsed time from the well defined origin up to the occurrence of a previously established event of interest; in turn, an artificial neural network can be defined as a mathematical model whose construction is carried out by means of a process that imitates the functioning of the biological neural networks, and can be used to shape phenomena that involve some response that depends on a combination of factors. This research approaches the survival analysis using artificial intelligence technologies for the purpose of estimating, since a neural network, the survival Kaplan-Meier function. The results demonstrate that the models of artificial neural networks allow the managing of survival data without needing to impose departure assumptions in the mentioned models. Thus, it is evident the potential of the RNA to evaluate the partial information from a censured data set of survival.
Key words: Reliability, Kaplan-Meier, artificial neural networks.

ABSTRACT

1. Introducción

En cualquier organización, la sistematización se utiliza con la finalidad de optimizar los procedimientos inherentes a su funcionamiento. Cada procedimiento involucra características particulares que conllevan a planificar, dirigir, supervisar y evaluar.

Con el avance tecnológico y social, la elaboración de sistemas para el desempeño de una organización es cada vez más compleja. Deben incluirse herramientas que permitan evaluar situaciones reales, para evitarle a la organización riesgos que conlleven a pérdidas económicas que desencadenen crisis.

Hoy en día, las aplicaciones estadísticas computacionales incluyen módulos con técnicas avanzadas para el desarrollo de modelos que permiten simular el comportamiento de variables claves en la organización.

Esta investigación aborda el análisis de supervivencia con técnicas de inteligencia artificial con la finalidad de estimar, a partir de una red neuronal, la función de supervivencia de Kaplan-Meier.

1.1. Análisis de confiabilidad

El análisis de confiabilidad, también conocido como análisis de supervivencia, se define como un conjunto de técnicas que se encargan de analizar el tiempo transcurrido desde el origen bien definido hasta la ocurrencia de un evento de interés que ha sido previamente establecido (Daponte y Sherman, 1991; Faraggi y Simon, 1995; Meeker y Escobar, 1998).

Para realizar un análisis de confiabilidad se necesitan por lo menos los valores de dos variables para cada uno de los componentes (individuos) bajo estudio:

a) Una variable que defina el tiempo transcurrido, desde que el componente empieza a ser parte del estudio (origen), el cual puede ser distinto para cada elemento, hasta que ocurre la falla o algún evento de interés definido previamente. Si al momento de finalizar el estudio, el componente no ha presentado el cambio de estado, se toma como fecha final la del cierre del estudio.

b) Una variable que indique si el componente presentó o no la falla o el evento de interés. A esta variable se le suele conocer como condición de censura.

Una característica fundamental de los datos de supervivencia se refiere al hecho de que un componente aún formando parte de la muestra, no presente el evento de interés, haciendo de esta manera que la información sea incompleta o censurada (Kaplan y Meier, 1958; Meeker y Escobar, 1998). Esto puede deberse a:

El componente falla por una causa distinta al evento de interés.

En la fecha de culminación del estudio el componente no registró el cambio de estado.

El componente sale del estudio por algún motivo.

En el gráfico 1 se muestra el esquema de los cambios que se estudian en el análisis de confiabilidad, censurándose aquellos componentes en los que no se produce el cambio de interés, los que por alguna causa desaparecen del estudio y los que fallan por otra razón distinta a la que se está estudiando. A este tipo de censura se denomina censura por la derecha y es sobre el cual se basa el estudio.

En este orden de ideas, se define la función de supervivencia o confiabilidad, como la

probabilidad de que a un componente no le ocurra el evento de interés (es decir, sobreviva) al menos hasta el tiempo (t) (Faraggi y Simon, 1995; Molinero, 2004; Ravdin y Clark, 1992; Ravdin y De Laurentis, 1993) cuya definición formal es:

> Sea T una variable aleatoria no negativa, la cual mide el tiempo de ocurrencia de un evento, cuyas funciones de densidad y de distribución son y $f(t)$ y $F_T(t)$, respectivamente, luego la función de supervivencia viene dada por:

$$S(t) = 1 - F(t) = P(T > t) \qquad (1)$$

Luego, esta función en un tiempo (t), expresa la probabilidad de que un sujeto no haya realizado el cambio hasta ese instante.

La función de supervivencia puede ser estimada por distintos métodos, siendo el más utilizado el propuesto por Kaplan y Meier (1958).

Para el caso en que los datos puedan presentar censura por la derecha, este estimador puede calcularse mediante:

$$\hat{S}_{KM}(t) = \prod_{t_i \le t} \frac{r(t_i) - d(t_i)}{r(t_i)} \qquad (2)$$

Donde $r(t_i)$ y $d(t_i)$ son el número de individuos en riesgo y el número de muertes (o de ocurrencia del evento de interés) en el tiempo (t_i), respectivamente.

1.2. Redes neuronales artificiales (RNA)

Numerosos avances han sido alcanzados en el campo de los sistemas inteligentes, algunos de ellos, inspirados por las redes neuronales biológicas. Investigadores de múltiples disciplinas científicas están diseñando RNAs para resolver una gran variedad de problemas en el reconocimiento de patrones, predicción, clasificación de datos, estimación, optimización, memorias asociativas, y control de procesos entre otros (Jain, Mao y Mohiuddin, 1996).

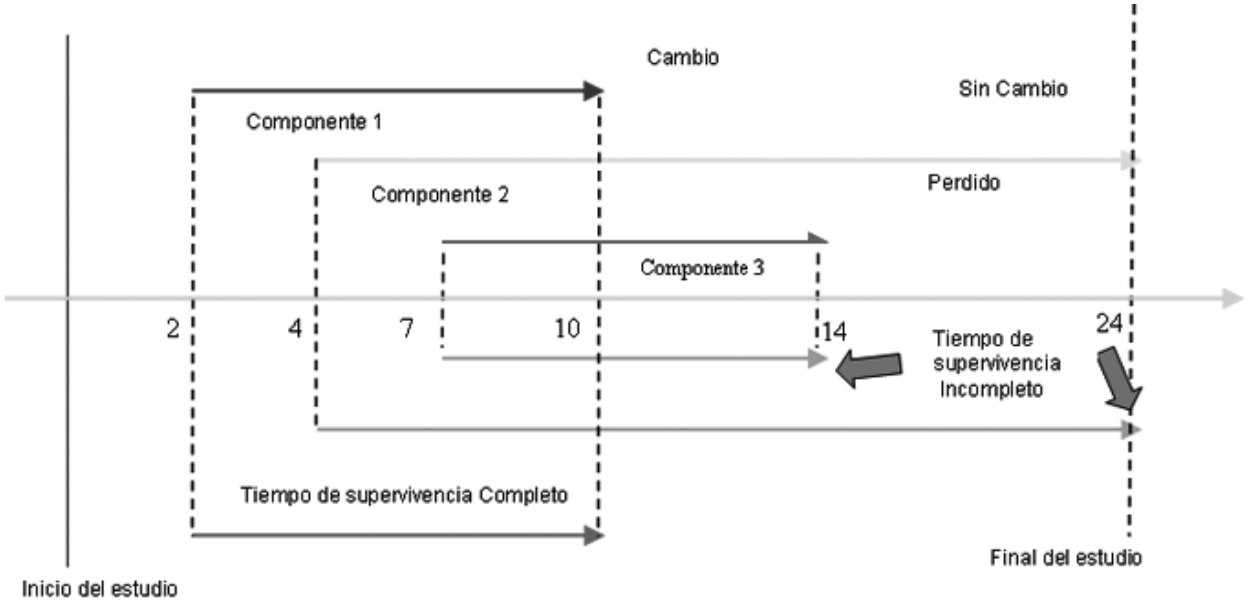

Gráfico 1. Esquema de datos censurados por la derecha

Una red neuronal artificial puede ser definida de múltiples maneras, una de ellas establece que una RNA es un modelo matemático cuya construcción se lleva a cabo mediante un proceso que, en parte, imita el funcionamiento de las redes neuronales biológicas (Gupta, Jin y Homma, 2003). Este modelo matemático consta de una o más variables de entrada (o variables independientes), y de una o más variables de salida (o variables dependientes), por tanto, una RNA puede ser usada para modelar fenómenos en los cuales existe alguna respuesta variable(s) explicada(s) que depende de algún conjunto de factores variable(s) explicativa(s). Esto implica que para modelar un fenómeno mediante una RNA, debe existir una relación matemática (lineal o no lineal) entre la(s) variable(s) explicativa(s) y la(s) variable(s) explicada(s), pues mientras mayor sea esta relación, mejor será el desempeño de la RNA (Gupta, Jin y Homma, 2003). En el caso bajo estudio, la salida de la RNA corresponde a la estimación de la función de supervivencia o de confiabilidad, mientras que las entradas corresponden a las variables que determinan dicha estimación, estas son el tiempo transcurrido hasta lograr el evento de interés (falla) y la variable característica que indica si esta falla se produjo.

Desde el punto de vista computacional, una RNA consiste en un conjunto de neuronas artificiales interconectadas entre sí. Además, las neuronas se encuentran, generalmente, agrupadas en capas o niveles. La manera como se pueden dar las conexiones entre las diversas neuronas de la red, es lo que comúnmente se denomina la topología de la red. Existen diversas topologías, sin embargo, sólo se hará referencia a aquella en la cual cada neurona de una capa dada, está conectada con todas las neuronas de la siguiente capa, tal como está representada en el gráfico 2 (Gupta, Jin y Homma, 2003):

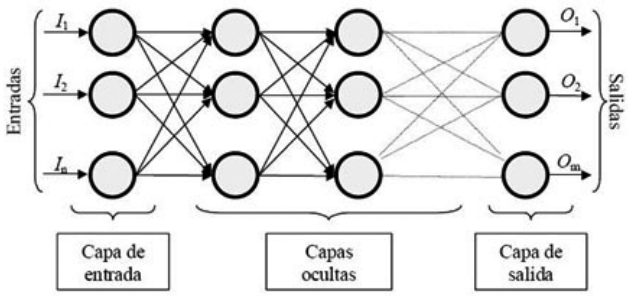

Gráfico 2. Red perceptrónica multicapa

Este tipo de RNAs se conocen como redes perceptrónicas multicapa (RPM), y fue el tipo de red utilizado en esta investigación. Tal como se puede apreciar en el gráfico 2, las neuronas (nodos) están agrupadas en capas verticales. Una RNA puede tener 2 o más capas, aunque generalmente se utilizan al menos 3, y cada capa puede tener cualquier cantidad de neuronas, dependiendo del problema que se esté tratando. La primera capa (de izquierda a derecha) es la capa de entrada, la última es la capa de salida, y las restantes son las capas ocultas. A su vez, cada nodo de la RNA está formado por los siguientes elementos (Gupta, Jin, y Homma, 2003):

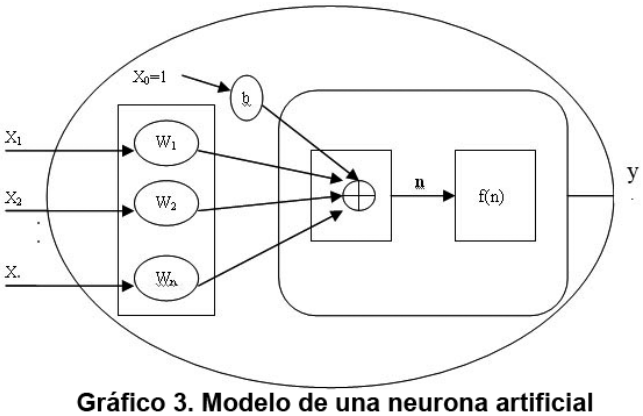

Gráfico 3. Modelo de una neurona artificial

Donde Xi es la i-ésima entrada de la neurona, Wi es el peso sináptico de la i-ésima entrada, y b es el sesgo de la neurona (una constante). La salida de la neurona es y = f(n), siendo f una función lineal o no lineal denominada función de transferencia, y n la suma ponderada de las entradas y el sesgo, tal

como se muestra a continuación:

$$n = \sum_{i=1}^{n} W_i X_i + b \qquad (3)$$

Por tanto, puede observarse que una neurona artificial es un modelo matemático, en donde la salida no es más que una transformación matemática (o función) de las entradas. A su vez, esto implica que la RNA es también un modelo matemático, el cual viene dado por el modelo matemático de las neuronas de la capa de salida.

Luego de plantear la arquitectura de la RNA (entradas, número de capas, número de neuronas para cada capa y la función de transferencia de cada capa), se procede con el entrenamiento de la misma. El entrenamiento de una RNA es un proceso mediante el cual se ajustan los pesos sinápticos y los sesgos de las neuronas, con el propósito de que la RNA pueda llevar a cabo con mayor exactitud la tarea para la cual fue desarrollada. En esta investigación se utilizó al algoritmo de retropropagación del error con momento (Gupta, Jin y Homma, 2003).

2. Metodología

La empresa Industria Venezolana de Aluminio, C.A. (CVG-Venalum), fue creada el 29 de agosto de 1973, con el objeto de producir aluminio primario con fines de exportación. Está ubicada sobre la margen del río Orinoco, en la ciudad de Puerto Ordaz, estado Bolívar, al sur de Venezuela y constituye la mayor planta reductora de aluminio primario en Latinoamérica con una capacidad instalada de cuatrocientas treinta mil toneladas por año.

CVG-Venalum cuenta con cinco líneas de producción de aluminio, cuatro que utilizan tecnología Reynolds P-19 y una quinta que usa tecnología Hydro-Aluminium. El aluminio producido viene presentado en lingotes, cilindros para extrusión y aluminio líquido. Las celdas que utilizan tecnología Reynolds P-19, y sobre la cual se basa el estudio, se identifican porque el sistema de alimentación de alúmina, está compuesto por cuatro alimentadores con su respectivo rompecostra que operan independientemente. Cada celda usa 18 ánodos con una vida útil de 22 días cada uno de ellos y una capacidad útil de producción mensual de 36 toneladas de aluminio por celda. La temperatura de operación de la celda es 960oC, la adición de fluoruro de aluminio es manual y el voltaje de operaciones 162 KA. La frecuencia de trasegado es cada 24 horas y la subida de puente es realizada cada 15 días (CVG-Venalum, 1998).

El presente es un estudio longitudinal que abarca un período de 5 años y 6 meses, para las dos líneas de producción que forman cada uno de los Complejos I y II de la Industria productora de Aluminio primario CVG-Venalum de Venezuela (Altuve, 2005). El inicio del estudio es el 1 de enero de 1998 y la fecha final fue el 30 de junio de 2004. El objetivo es comparar el tiempo de vida de las celdas electrolíticas de los dos complejos que presenten como evento de desincorporación de la celda la perforación en el cátodo (uno de los elementos fundamentales de las celdas electrolíticas).

Se procederá en primer lugar a estimar la probabilidad de supervivencia de Kaplan-Meier para cada uno de los complejos. Para esto se utilizará el programa R versión 2.4.1. Una vez obtenida esta función se procederá a entrenar la red neuronal con el paquete Statistical Neural Network V-4.0E, donde se usó una proporción 80-20 para los conjuntos de entrenamiento y validación. La matriz de datos consta de 1.708 observaciones para el complejo I y 1.843 para el complejo II.

3. Resultados

3.1. Usando el estimador de Kaplan Meier

Para las líneas pertenecientes al Complejo I se tiene que la estimación de la función de supervivencia, se obtiene para las 1.708 celdas de reducción electrolítica, con una mediana de la supervivencia de 532 días, es decir, que al menos la mitad de las celdas en producción lograron sobrevivir hasta el día 2.334 de seguimiento sin haber sido desincorporada por perforación en el cátodo (Cuadro 1).

En el gráfico 4 se puede observar cómo en los primeros 1.900 días (aprox.) de la vida de las celdas, ninguna presentó falla por perforación del cátodo mostrando un patrón paralelo al eje horizontal, decreciendo de manera casi lineal la función de supervivencia a partir de este valor, lo cual indica que las desincorporaciones de las celdas por esta causa tienen un comportamiento uniforme en el tiempo hasta los 3.000 días.

Cuadro 1
Valores resumen de la estimación de la función de supervivencia debida a la perforación del cátodo de la celda para el Complejo I

n	eventos	mediana	LCI(0.95)	LCS(0.95)
1708	532	2334	2291	2382
Donde:				
n es el numero de individuos				
eventos: número total de desincorporaciones				
Mediana: sobrevida mediana de las celdas				

Al observar los resultados del Complejo II, se puede apreciar que el número de celdas que presentaron la falla de interés fue de aproximadamente el 80%. La mediana de la sobrevida de las celdas es para este complejo de 1.647 días, lo que indica que la mitad de las celdas permanece en funcionamiento sin salir del mismo por presentar perforación en el cátodo de éstas por 1.647 días.

En el gráfico 5 se aprecia el comportamiento de la función de supervivencia para este complejo, observándose en el mismo una

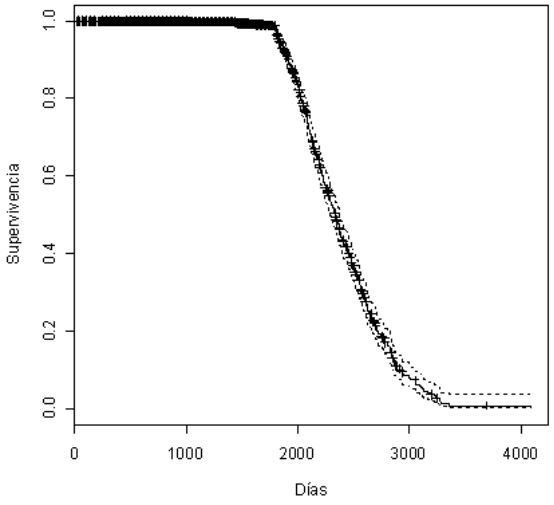

Gráfico 4. Estimador de Kaplan y Meier S(t) Complejo I

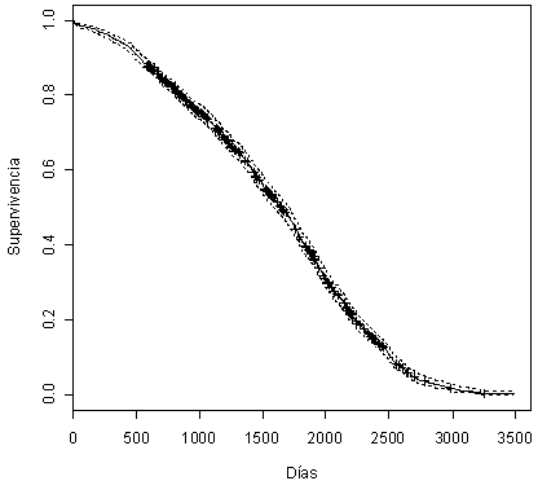

Gráfico 5. Estimador de Kaplan y Meier S(t) Complejo II

Cuadro 2
Valores resumen en la estimación de la función de supervivencia debida a la perforación del cátodo de la celda para el Complejo II

n	eventos	mediana	LCI(0.95)	LCS(0.95)
1843	1483	1647	1592	1710
Donde:				
n es el numero de individuos				
eventos: número total de desincorporaciones				
Mediana: sobrevida mediana de las celdas				

disminución acentuada los primeros 500 días de seguimiento, y una disminución menor para el resto del período en estudio.

3.2. Usando RNAs

Al entrenar las redes neuronales para los complejos I y II, se eliminaron todos los primeros patrones ya que no presentaron el cambio de estado en estudio, es decir, se encuentran censurados y esto además de no afectar la función de supervivencia puede perturbar el entrenamiento de la red.

Por otro lado, el algoritmo de entrenamiento utilizado fue el backpropagation con momento, con 1.000 ciclos y el tipo de error empleado fue el Error Cuadrático Medio (ECM), con un nivel deseado de 10-5. Las arquitecturas de la RNAs evaluadas se basaron en una capa oculta, con funciones de activación sigmoidal en todos sus nodos. La

Cuadro 3
RNAs Evaluadas

Nº de Neuronas en la Capa Oculta	ECM
5	0,0097
7	0,0082
9	0,0076
13	0.0074
15	5.2408e-004 *
17	0.0057
20	0,0040
30	0,0075

cantidad de neuronas en la capa oculta se varió de una RNA a otra, tomándose la cantidad óptima a aquella que generó el menor ECM.

A continuación se presenta una cuadro resumen de las diferentes estructuras utilizadas para entrenar las redes del Complejo.

En el cuadro 3 se puede apreciar que la mejor arquitectura es la que contiene 15 neuronas en la capa oculta, arrojando un ECM de 5.2408*10-4.

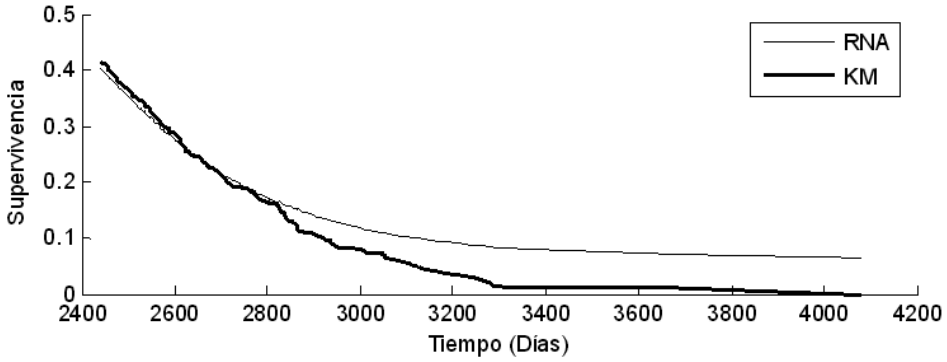

Gráfico 6. Comparación de las curvas de supervivencia

Cuadro 4
Neuronas en la capa escondida

Nº de neuronas en la capa escondida	Error
10	9.5638e-005
5	1.5234e-004
8	1.2574e-004
12	9.1567e-005*
15	1.9519e-004

Ahora bien, gráficamente se puede observar que entre los 2.400 y 2.800 días la RNA copia perfectamente la función de supervivencia de KM mientras que a partir de aquí la RNA sobreestima la función:

Por otro lado, al analizar el comportamiento de las diferentes estructuras utilizadas para el Complejo II, se obtuvieron los siguientes resultados:

Siendo la mejor estructura la red con 12 neuronas en la capa escondida ya que es la que menor error presenta.

En el gráfico 7, se puede observar que la estimación del estimador obtenido mediante el modelado de la RNA es muy similar al estimar la función de supervivencia de KM para este complejo.

4. Conclusiones

En este estudio se muestra la potencialidad que tienen los modelos basados en redes neuronales en el manejo de datos de supervivencia, teniendo una ventaja adicional: no es necesario imponer supuestos de partida en los mismos.

Se puede conocer que la mediana de vida de las celdas electrolíticas para los complejos I y II de la empresa CVG, Venalum es de 2.334 y 1.647 respectivamente debido a la perforación del cátodo; luego la empresa debe tomar las previsiones del mismo para poder de esta manera hacer los correctivos necesarios.

5. Bibliografía

Altuve, L. (2005). Estimación de la vida útil de las celdas de reducción electrolítica. CVG-VENALUM (enero 1.998 - junio 2.004). Trabajo Especial de Grado, Escuela de Estadística, Universidad de Los Andes, Mérida, Venezuela.

CVG-Venalum (1998). Manual para Ingenieros.

Daponte, J.S. y Sherman, P. (1991). Classification of ultrasonic image texture by statistical discriminant analysis of neural network. En: Computerized Medical Imaging and Graphics, 15, 3-9.

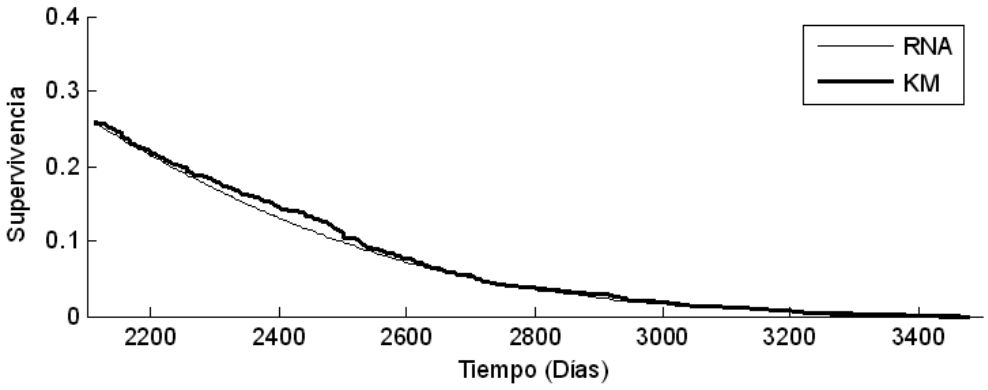

Gráfico 7. Comparación de las curvas de supervivencia

Faraggi, D. y Simon, R. (1995). A Neural Network Model for Survival Data. En: Statistics in Medicine, 14, 73-82.

Gupta, M.; Jin, L. y Homma, N. (2003). Static and dynamic neural networks. John Wiley and Sons.

Jain, A.; Mao, J. y Mohiuddin, K., (1996). Artificial neural networks: A tutorial. IEEE.

Kaplan, E. L. y Meier, P. (1958). Nonparametric estimation from incomplete observations. En: Journal of the American Statistical Association, 53: 457-481.

Meeker, W. Q. y Escobar, L. A (1998). Statistical methods for reliability data. Nueva York: John Wiley & Sons, Inc.

Molinero, L. (2004). Verificación de los modelos de supervivencia de Cox. Disponible en: www.seh_lelha.org/stat1.htm

Ravdin, P. y Clark, G. (1992). A practical application of neural network analysis for predicting outcome of individual breast cancer patients. En: Breast Cancer Res Treat, 22, 285-293.

Ravdin, P. y De Laurentis M. (1993). A Technique for using neural Network Analysis to Perform Survival Analysis of Censored Data. En: Cancer Letters, 77, pp. 127-138