

# **Clasificación de microarrays de genes del cerebro para la detección de tumores, utilizando algoritmos no convencionales.**

**Osvaldo Miguel González Prieto<sup>1</sup>, Katia Andrea Ayala Diaz<sup>2</sup>**

Facultad Politécnica - UNE.

Ciudad del Este - Paraguay

<sup>1</sup>osvalcde@gmail.com, <sup>2</sup>ktiaayala@gmail.com

## **Resumen**

Este trabajo tiene por objetivo comparar los resultados de algoritmos de clasificación en minería de datos, sobre un conjunto de genes pertenecientes al cerebro, con el fin de encontrar el mejor clasificador entre genes enfermos y no enfermos. La metodología utilizada fue la Data Mining CRISP-DM (Cross Industry Standard Process for Data Mining) [CRISP-DM00] que está definida en términos de un modelo jerárquico de procesos, consiste de un conjunto de tareas descritas en cuatro niveles de abstracción (desde lo general a lo específico): Fases, Tareas Genéricas, Tareas Especializadas e Instancias de procesos. Los datos de prueba consistieron en 7070 genes para 69 muestras en el archivo de entrenamiento y 23 muestras en el archivo de prueba, todos pertenecientes a un tipo de gen. Para realizar la exploración y análisis de los datos, se ha seleccionado como apoyo la herramienta WEKA, con la cual se obtuvieron ciertos resultados estadísticos que permitieron comprender el comportamiento de los genes en todas las muestras. Se ha probado con un determinado número de algoritmos incluidos en la herramienta WEKA y a partir de los resultados obtenidos se determinó cuál es la mejor clasificación.

Los algoritmos utilizados son los siguientes: NaiveBayes, J48, IBK para K=1,2,3,4 y Multi-ClassClassifier. El algoritmo de clasificación con mejores resultados, fue el Naive-Bayes. En este trabajo se ha podido demostrar la gran utilidad que tiene la minería de datos, algo que se ha podido ver con un caso real mediante la aplicación de diferentes algoritmos.

**Descriptores:** **microarreglo, minería de datos**

## **Abstract**

This work compares the results found in a data mining classifying, applied to a set of certain type of genes from human brain. The goal was to find the best classifier capable of distinguishing between healthy and ill genes. The methodology used was Data Mining CRISP-DM (Cross Industry Standard Process for Data Mining) [CRISP-DM00] which is defined in terms of a hierarchical model of processes as a set of tasks described in four levels of abstraction (from generic to specific): Phases, Generic tasks, Specialized Tasks and process Instances. For this test, 7070 genes were used: 69 samples in the training file, and 23 samples in the test file; all genes from the same type. The tool WEKA was chosen for exploring and classifying, which allowed for a genes behavior comprehension in all samples. A certain number of algorithms were used which are contained in the tool WEKA, looking for the best classification. The algorithms applied were: NaiveBayes, J48, IBK for K=1,2,3,4 and Multi-ClassClassifier; from these, NaiveBayes has rendered the best classification. This work has shown the great utility data mining has, which could be proved with a real case through different algorithms.

**Keywords:** **microarray, data mining**

## 1. Introducción.

En los últimos años ha habido una explosión en la velocidad de adquisición de datos biomédicos y biotecnológicos. Los avances en las tecnologías de genética molecular, como ser los microarreglos (*microarrays*) de ADN han permitido obtener una visión global de la célula. A través de esta metodología es que se puede observar y medir la expresión simultánea de miles de genes. Los microarreglos han abierto la posibilidad de crear conjuntos de datos con información molecular para representar distintos sistemas biológicos o de interés clínico. En la actualidad los *microarrays* se están aplicando en una gran diversidad de aplicaciones biomédicas, tales como, cáncer [4], terapia génica, hipertensión arterial, toxicidad ambiental, reconocimiento de nuevos fármacos, etc.

Las enormes cantidades de datos biológicos y crecientes demandas de la investigación biológica moderna exigen cada vez más la sofisticación y computación potente de las tecnologías de la información (TI). Más concretamente, la utilización óptima de estos instrumentos exige conocer en qué puntos se encuentran los datos al transcurrir la investigación biológica. Este trabajo consiste en la exploración de datos, basado en la metodología CRISP-DM [8, 1]. Para tal efecto se utiliza una base de datos con informaciones de *microarrays* de genes de una zona del cerebro.

Este trabajo esta enfocado al estudio de genes extraídos del cerebro, con una muestra de 7070 registros, obtenidos de un repositorio público.

### Objetivos

#### Objetivo general

Comparar los resultados de algoritmos de clasificación en una Minería de Datos, aplicados sobre un conjunto de genes del cerebro, con el fin de encontrar el mejor clasificador entre genes enfermos y no enfermos.

#### Objetivos Específicos.

- Seleccionar una metodología de minería de datos.
- Estudiar datos genéricos del cerebro (comprensión del negocio).
- Analizar la generación de *microarrays*, sus características y tipos (comprensión de los datos).
- Preparación de los datos.
- Aplicar diferentes algoritmos de clasificación.
- Evaluar resultados.

## 2. Materiales y Métodos.

### 2.1 Metodología de exploración de datos.

La metodología propuesta es una de las más utilizadas (Figura 1), lo cual alienta su utilización. La implementación de minería de datos no es una tarea trivial, CRISP-DM la divide en seis etapas estructuradas y relacionadas entre sí, simplificando la compleja tarea de su desarrollo.

¿Cuales son las principales metodologías de Explotación de datos?

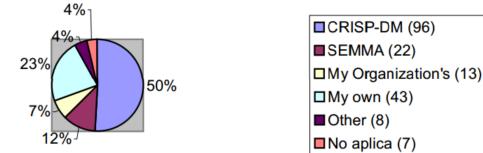


Figura 1. Resultado de la encuesta realizada en <http://www.kdnuggets.com>, año 2010.

#### Metodología CRISP-DM.

La metodología de minería de datos CRISP-DM (Cross Industry Standard Process for Data Mining) [CRISP-DM00] que está definida en términos de un modelo jerárquico de procesos, consiste de un conjunto de tareas descritas en 4 niveles de abstracción (desde lo general hacia lo específico): Fases, Tareas Genéricas, Tareas Especializadas e Instancias de procesos (Fig. 2).

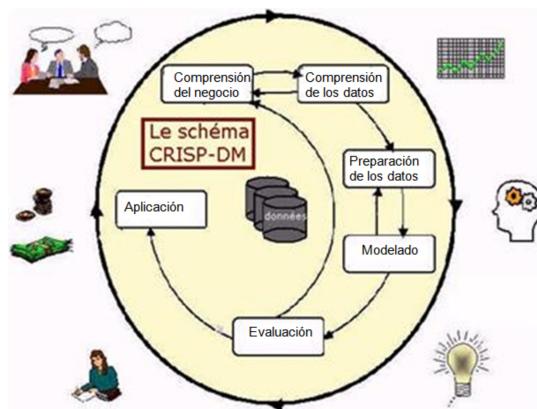


Figura 2. Estructura del ciclo de vida de un proyecto de minería de datos.

### 2.2 Recolección de datos iniciales:

Todos los días, y casi desapercibidamente, se genera gran cantidad de datos informatizados, por ejemplo cuando se realiza una compra, cuando se marca el ingreso y el egreso en el trabajo.

El conjunto de datos analizados en el trabajo, se ha obtenido del siguiente servidor: [http://pegaso.ls.fi.upm.es/~omarban/final\\_project\\_data.zip](http://pegaso.ls.fi.upm.es/~omarban/final_project_data.zip). El mismo pertenece a la minería de datos de Piatesky: “**Predecir clases**

de enfermedades genéticas mediante datos de microarray”[7]. Esta técnica se compone de tres tipos de archivos diferentes: Por un lado se tienen los datos de entrenamiento (archivo pp5i\_train.gr.csv) que permiten construir el modelo el cual, posteriormente es utilizado para clasificar los datos; por otro lado se tienen los datos de prueba (pp5i\_test.gr.csv) y por último un archivo (pp5i\_train\_class.txt) que contiene las etiquetas de las clases de los genes.

Tanto el archivo de entrenamiento como el archivo de prueba, contienen datos con 7.070 genes para 69 muestras en el archivo de entrenamiento y 23 muestras en el archivo de prueba, todos pertenecientes a un tipo de gen, que son como sigue: EDD, APP, MED, MGL, RHB. Los datos que serán utilizados para el proyecto se obtuvieron con una simple descarga, pero para una mejor comprensión es necesario conocer y entender cómo se generan y de donde se extraen. Al efecto, se ejemplifica en un proceso simplificado de cuatro etapas, como se muestra a continuación:

### 2.3 Comprensión de los datos.

#### Proceso simplificado de extracción de genes.

*Paso 1:* Proceso de extracción de los genes (Fig. 3).

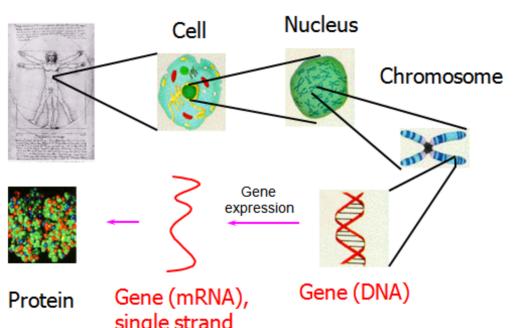


Figura 3. Proceso de extracción de los genes[5].

*Paso 2:* Escaneo de los genes (Fig. 4).

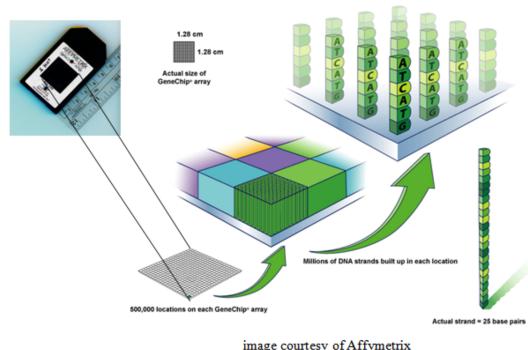


Figura 4. Proceso de extracción microarrays, escaneo de genes (affymetrix)[5].

*Paso 3 (Fig. 5):*

- Etiquetar segmentos de mRNA utilizando químicos fluorescentes.
- Dividirlos en sondas complementarias.
- Medir la fluorescencia con láser.

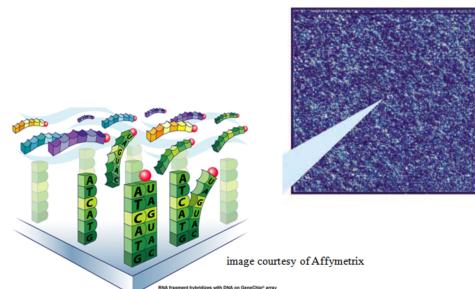


Figura 5. Etiquetar segmentos de mRNA[5].

*Paso 4: Extracción de los valores para cada tipo de gen encontrado (Fig. 6).*

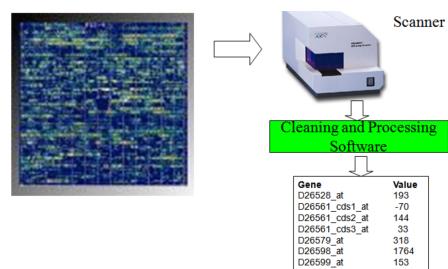


Figura 6. Extracción de la información[5].

### 2.4 Descripción de los datos.

El objetivo de la descripción de los datos incluye el formato de los mismos, su cantidad, los identificadores de los campos, y cualquier otro rasgo superficial que ha sido descubierto. En esta muestra se introdujeron deliberadamente datos de células normales y células enfermas, como ejemplo en la Figura 6, se aplica el mismo proceso de obtención de microarrays de las células normales y con tumor. En la figura 7 se ilustra el proceso de preparación de muestras celulares.

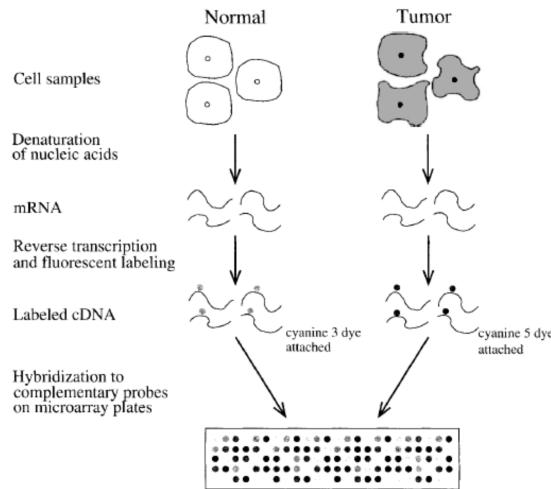


Figura 7. Preparación de muestras: el proceso de las muestras de células de los microarrays.

#### 2.4.1 Archivo de clases.

Contiene las clases separadas para cada muestra, correspondiente al orden de las muestras del archivo de entrenamiento. Existen 5 clases, etiquetadas de la siguiente manera (tabla 1):

Tabla 1. Relación entre tipo de gen y cantidad.

Clase	Cantidad
MED	39
MGL	7
RHB	7
EPD	10
JPA	6

Este conjunto de datos consiste en 7070 genes obtenidos utilizando “Affymetrix gene chip”. Contiene cinco clases (MED, MGL, RHB, EPD, JPA) y 69 muestras de las cuales 39 son MED, 7 MGL, 7 RHB, 10 EPD y 6 JPA. Según [LING08] estos genes se utilizan para el estudio de los tumores cerebrales, en inglés *brain tumor*.

#### 2.4.2 Exploración de datos.

En esta fase, se elabora un informe que brinda resultados que permiten obtener mayores detalles acerca de las características que tienen los datos que son utilizados para realizar la práctica. Cabe destacar, que el conocimiento que se obtiene en esta fase de la metodología aplicada, afecta inmediatamente las acciones que se desarrollan en la siguiente fase (Preparación de los Datos). Para realizar la exploración de los datos, se ha seleccionado como apoyo la herramienta WEKA, con la cual se obtienen ciertos resultados estadísticos que permiten comprender el comportamiento de los datos (7070 genes) en todas las muestras (69). En la Fig. 8 se puede observar en el cuadro de

*Current relation* la cantidad de instancias y atributos, en el cuadro de *Selected attribute* el valor mínimo, máximo, media y desviación; más abajo se puede ver el gráfico de acumulación de gen por valor de cada uno.

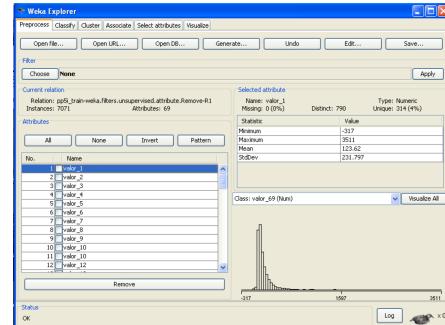


Figura 8. Ventana de Exploración de WEKA

#### 2.4.3 Verificación de la calidad de los datos.

La verificación de la calidad de los datos consiste en la descripción de posibles errores que pudieran afectar la fiabilidad de los resultados. La muestra de datos fue analizada, con la ayuda de la herramienta WEKA, en busca de elementos faltantes; donde no se detectaron valores vacíos. También se pudo constatar que todos los datos cumplen la restricción de tipo de dato por cada muestra, es decir, todos los valores son de tipo numérico.

#### 2.4.4 Selección de técnica de modelado.

Finalmente, y una vez realizado todo el preprocesamiento de los datos proporcionados, se pasó al procesamiento de los mismos para clasificar correctamente los genes. Lo que se busca es el algoritmo que dé los mejores resultados de clasificación. Para ello se prueba un determinado número de algoritmos incluidos en la herramienta WEKA y a partir de los resultados ofrecidos se determina cual es la mejor clasificación.

### 2.5 Algoritmos utilizados.

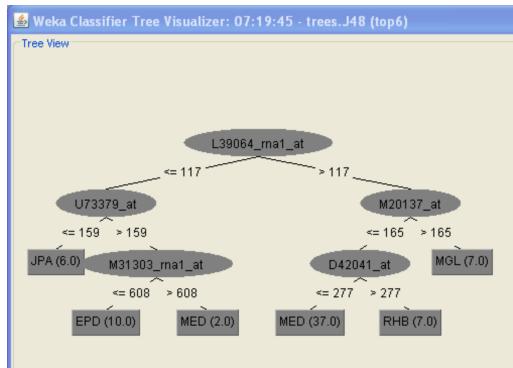
Los algoritmos utilizados son los siguientes NaiveBayes, J48, IBK para K=1,2,3,4 y Multi-ClassClassifier. A continuación se presenta una pequeña descripción de cada algoritmo.

#### Breve descripción de cada algoritmo.

- *NaiveBayes*: Es un método de aprendizaje que reduce su calidad ante la presencia de atributos no relevantes[2]. Que pertenece al conjunto de métodos Bayes, que se trata de una técnica de clasificación descriptiva y predictiva basada en la teoría de la probabilidad del análisis de T. Bayes[3], que data

de 1763. Esta teoría supone un tamaño de la muestra asintóticamente infinito e independencia estadística entre variables independientes, refiriéndose en este caso a los atributos, no a la clase. Con estas condiciones, se puede calcular las distribuciones de probabilidad de cada clase para establecer la relación entre los atributos (variables independientes) y la clase (variable dependiente).

- **Algoritmo J48:** Este algoritmo es un clásico de aprendizaje de árbol de decisión, basado en el algoritmo C4.5[3]. Este método forma parte del grupo de *trees*, que son métodos que aprenden mediante la generación de árboles de decisión[3]. En la figura 9 se tiene el árbol resultante en cuyos nudos aparecen los genes más significativos y como posibles resultados los cinco tipos de muestras. La característica fundamental de este algoritmo es que incorpora una poda del árbol de clasificación una vez que éste ha sido inducido, es decir, una vez construido el árbol de decisión, se podan aquellas ramas del árbol con menor capacidad predictiva[2].



- **Algoritmo IBk:** Este algoritmo está basado en instancias, por ello consiste únicamente en almacenar los datos presentados. Cuando una nueva instancia es encontrada, un conjunto de instancias similares relacionadas es devuelto desde la memoria y usado para clasificar la instancia consultada[2]. Se trata, por tanto, de un algoritmo del método *Lazy Learning*. Este método de aprendizaje se basa en que los módulos de clasificación mantienen en memoria una selección de ejemplos sin crear ningún tipo de abstracción en forma de reglas o de árboles de decisión (de ahí su nombre, *lazy*, que significa perezoso). Cada vez que una nueva instancia es encontrada, se calcula su relación con los ejemplos

previamente guardados con el propósito de asignar un valor de la función objetivo para la nueva instancia[6]. La idea básica sobre la que se fundamenta este algoritmo es que un nuevo caso se ha de clasificar en la clase más frecuente a la que pertenecen sus K vecinos más cercanos. De ahí que sea también conocido como método K-NN: K Nearest Neighbours[6]. La aplicación del método se hace en tres etapas con el valor de K = 1,2,3 correspondiendo un valor para cada etapa. Como en los demás métodos la configuración de parámetros estándares ha sido mantenida.

### 3. Resultados.

#### 3.1 Resultados de cada algoritmo.

##### 3.1.1 Algoritmo Naive Bayes.

Utilizando este algoritmo se obtuvo el mejor resultado, es decir, tras entrenar la máquina con los *top 6*, se obtuvieron los siguientes resultados: un 98,5507 % de instancias bien clasificadas, y como resultado un 1,4493 % incorrectos. Esto es un resultado óptimo, a pesar de que varios de los grupos y algoritmos igualaron estos resultados, prevaleció la variable de error Absoluto Medio (EAM) para identificar el mejor. Luego de esta comparación el conjunto de *top 6*, aplicando el algoritmo de NaiveBayes con 0,0093 de EAM resultó como mejor combinación (Fig. 10).

En la siguiente figura (Fig. 10) se presentan los resultados de instancias bien clasificadas, instancias incorrectamente clasificadas y el error medio absoluto, para *top 6*, utilizando estas variables se realizó la clasificación de mejores conjuntos. Vale destacar que para el conjunto de *top 6* se obtuvieron los valores más bajos en error medio absoluto. Utilizando el algoritmo de clasificación de Naive-Bayes se obtuvo el mejor resultado.

==== Summary ===		
Correctly Classified Instances	68	98.5507 %
Incorrectly Classified Instances	1	1.4493 %
Kappa statistic	0.9768	
Mean absolute error	0.0093	
Root mean squared error	0.0819	
Relative absolute error	3.5954 %	
Root relative squared error	22.9754 %	
Total Number of Instances	69	

Figura 10. Resultados obtenidos a través del algoritmo NaiveBayes, con el conjunto de *top 6*.

##### 3.1.2 Algoritmo J48.

El mejor resultado de este algoritmo se presenta con el conjunto de *top 30*, en él se muestra un 89.8551 % de instancias bien clasificadas. En promedio entre los diferentes conjuntos de genes *top*, este algoritmo es el que presenta los valores más bajos para instancias bien clasificadas (Fig. 11).

== Summary ==		
Correctly Classified Instances	62	89.8551 %
Incorrectly Classified Instances	7	10.1449 %
Kappa statistic	0.8426	
Mean absolute error	0.0454	
Root mean squared error	0.2002	
Relative absolute error	17.5922 %	
Root relative squared error	56.1719 %	
Total Number of Instances	69	
Ignored Class Unknown Instances	72	

Figura 11. Resultados obtenidos a través del algoritmo J48, con el conjunto de top 30.

### 3.1.3 Algoritmo de IBK ( $k=1,2,3,4$ ).

Como se comentara anteriormente este algoritmo está basado en instancias, fue probado con diferentes profundidades desde 1 hasta 4, para la profundidad 1 con pequeñas muestras se presentó una excepción. Realizando una comparación entre los resultados obtenidos con este algoritmo, los mejores resultados, es decir, las mejores cantidades de muestras bien clasificados se presentan para profundidades del tipo 1 y 4, exactamente los extremos, esto podría ocurrir porque para  $k=1$  el nivel de comparación o profundidad es mínimo, por esto aparecen promedios de errores muy bajos (Fig. 12).

== Summary ==		
Correctly Classified Instances	67	97.1014 %
Incorrectly Classified Instances	2	2.8986 %
Kappa statistic	0.955	
Mean absolute error	0.0116	
Root mean squared error	0.1077	
Relative absolute error	4.4901 %	
Root relative squared error	30.1957 %	
Total Number of Instances	69	

Figura 12. Uno de los mejores resultados obtenidos aplicando IBK ( $k=1$ ) y con el conjunto top=6.

## 3.2 Evaluación de Resultados.

### Análisis del mejor conjunto y algoritmo resultante.

En este apartado se aplican algunas modificaciones en las variables del algoritmo de clasificación con mejores resultados, en este caso el NaiveBayes, como es recomendado en [4]. Con el objetivo de analizar las variaciones que podrían sufrir sus resultados. A continuación los resultados iniciales de este algoritmo (Fig. 13).

== Summary ==		
Correctly Classified Instances	68	98.5507 %
Incorrectly Classified Instances	1	1.4493 %
Kappa statistic	0.9768	
Mean absolute error	0.0093	
Root mean squared error	0.0819	
Relative absolute error	3.5954 %	
Root relative squared error	22.9754 %	
Total Number of Instances	69	

== Detailed Accuracy By Class ==						
Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
1	0.033	0.975	1	0.987	0.999	MED
1	0	1	1	1	1	MGL
1	0	1	1	1	1	RHB
0.9	0	1	0.9	0.947	0.998	EFD
1	0	1	1	1	1	JPA
Weighted Avg.	0.986	0.019	0.986	0.986	0.985	0.999

== Confusion Matrix ==					
a	b	c	d	e	<-- classified as
39	0	0	0	0	a = MED
0	7	0	0	0	b = MGL
0	0	7	0	0	c = RHB
1	0	0	9	0	d = EFD
0	0	0	0	6	e = JPA

Figura 13. Resultado arrojado para el conjunto de top 6, aplicando el algoritmo de NaiveBayes.

Vale destacar que mejorar los resultados obtenidos con este conjunto de genes y el algoritmo de NaiveBayes es muy difícil, si esto ocurriese se llegaría a la perfección con 100 % de las instancias bien clasificadas, este escenario no se ha presentado en ningún otro artículo o informe relacionado a proyectos de minería de datos. Después de ver todos los algoritmos utilizados en la presente práctica, a continuación se muestran los resultados obtenidos para cada uno de los algoritmos a partir de los datos pre-procesados (tabla 2 y Fig. 14).

Tabla 2. Resumen de resultados para los diferentes algoritmos.

	Clasificados Correctamente	Clasificados Incorrectamente	Mean Absolute Error
NaiveBayes	98.5507	1.4493	0.0093
J48	75.3623	24.6377	0.099
IB1	97.1014	2.8986	0.0116
IBk (k=2)	95.6522	4.3478	0.0291
IBk (k=3)	95.6522	4.3478	0.0291
IBk (k=4)	95.6522	4.3478	0.029
MultiClassClassifier	95.6522	4.3478	0.285

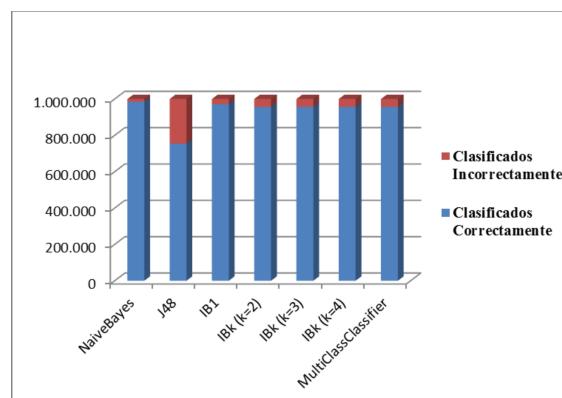


Figura 14. Resumen de resultados para los diferentes algoritmos.

#### 4. Comentarios finales.

En esta práctica se ha podido demostrar la gran utilidad que tiene la minería de datos, algo que se ha podido notar con un caso real y mediante la aplicación de diferentes algoritmos para el tratamiento de los datos desprendidos del problema. Se deben mencionar algunas dificultades que se presentaron en esta etapa, relacionadas principalmente a la interpretación de los resultados, se considera que esta falencia podría solucionarse con el conocimiento de un experto en genética, de esta manera sería posible dar un mayor significado a los resultados estadísticos que arrojan los algoritmos.

La siguiente tarea es difundir los resultados, a través de reuniones con profesionales del área que han encargado el proyecto. También debe ser estudiado un plan de implementación conjuntamente con el equipo de expertos, para lograr una estrategia eficiente de utilización de los resultados obtenidos.

#### Referencias bibliográficas

- [1] CRISP-DM. [En línea] <http://www.crisp-dm.org/CRISPWP-0800.pdf> [Marzo, 2011].
- [2] R. Bouckaert, W. Frank. “Manual WEKA 3.6.0”. The University of Waikato. (2011).
- [3] J. Orallo,[En línea] <http://users.dsic.upv.es/~jorallo/master/seminari.part.I.pdf> [Mayo, 2011].
- [4] kdnuggets. [En línea] [http://www.kdnuggets.com/data\\_mining\\_course/](http://www.kdnuggets.com/data_mining_course/) assignments/final-project.html [Febrero, 2011].
- [5] Knowledge Base, [En línea] [http://www.socialresearchmethods.net/kb/stat\\_t.php](http://www.socialresearchmethods.net/kb/stat_t.php) [Marzo, 2011].
- [6] J. Febles, A.l González. Aplicación de la minería de datos en la bioinformática. [En línea] [http://bvs.sld.cu/revistas/aci/vol10\\_2\\_02/aci03202.htm](http://bvs.sld.cu/revistas/aci/vol10_2_02/aci03202.htm) [Marzo, 2011].
- [7] Kidshealth. [En línea] [http://kidshealth.org/teen/en\\_espanol/cuerpo/genes\\_genetic\\_disorders\\_esp.html](http://kidshealth.org/teen/en_espanol/cuerpo/genes_genetic_disorders_esp.html) [Marzo, 2011].
- [8] Chapman, P., Clinton, J., Keber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R..1999. CRISPDM 1.0 Step by step BIguide. “CRISPWP-0800.pdf”, [En línea] [www.crispdm.org](http://www.crispdm.org) [Mayo, 2011].

#### Bibliografía complementaria

- D. Larose, Data Mining Methods and Models, Departmen of Mathematical Sciences Central Connecticut State University, (2010).
- Universidad de Waikato. [En línea] <http://www.cs.waikato.ac.nz/ml/weka> [Abril, 2011].
- L. Molina, Data mining: torturando a los datos hasta que confiesen, Universitat Oberta de Catalunya, 2002.
- Christine Lehman, “Calculate the T-Value”.[En línea] [http://www.ehow.com/how\\_5092736\\_calculate-tvalue.html](http://www.ehow.com/how_5092736_calculate-tvalue.html) [Marzo, 2011].