Análisis de métodos de clasificación para el diagnóstico de fertilidad

Lydia Jazmín Hernández Figueroa

Instituto Tecnológico de Celaya

11030655@itcelaya.edu.mx

Axel Serna Manríquez

Instituto Tecnológico de Celaya

11030783@ itcelaya.edu.mx

Jesús Alonso Gómez Melesio

Instituto Tecnológico de Celaya

11030759@itcelaya.edu.mx

Josefina Guadalupe Hurtado Mendoza

Instituto Tecnológico de Celaya

11030820@itcelaya.edu.mx

Norma Verónica Ramírez Pérez

Instituto Tecnológico de Celaya

norma.ramirez@itcelaya.edu.mx

Resumen

En la actualidad muchos de los procesos cotidianos requieren de grandes bases de datos para el manejo y análisis de información importante. La Inteligencia Artificial ofrece diferentes herramientas para la obtención de resultados sobre ellos. Por otra parte, una de estas herramientas relacionadas con esta área es la minería de datos que a través de métodos matemáticos, se extraen deducciones significativas a partir del aprendizaje automático. En este estudio se utilizó el software WEKA, herramienta contenedora de una gran colección de algoritmos de clasificación, la cual permitió hacer

uso de: Kstar, Random Forest, Regresión Lógica Bayesiana y Nnge proporcionando información relevante para hacer una comparación sobre la efectividad en la clasificación a la base de datos de Fertilidad (Fertility) extraída del repositorio UCI.

Palabras clave. *Inteligencia Artificial, Métodos de clasificación, Minería de datos, WEKA.*

Abstract

Nowadays so many of the daily processes require large databases for the managing and analysis of important information. Artificial Intelligence offers different tools for finding results on them. Concerning this, one of these tools related to this area is data mining through mathematical methods, significant deductions are taken from machine learning. The WEKA software, a tool container collection classification algorithms used in this study, which allowed us to use: Kstar, Random Forest, Regression and Bayesian Logic Nnge providing relevant information to make a comparison of the effectiveness of the classification database fertility extracted from the UCI repository.

Keywords: *Artificial Intelligence, Classification methods, Data mining**,** WEKA*

1. Introducción

Con la llegada de la tecnología hoy en día en cualquier entorno existe la necesidad del manejo de grandes cantidades de datos para la optimización y/o solución de muchos procesos de distintas áreas como: la medicina, el mercado de valores, la robótica, la educación entre otras áreas. En este tipo de áreas el análisis de sus datos puede llegar a ser muy tardado o hasta imposible si fuera estudiado por personas analistas, por ello es necesario que sean procesados mediante un programa que ayude a hacer una clasificación de la información, con ello se pretende realizar procesos con mayor

eficacia. Afortunadamente ahora se cuentan con diferentes recursos para la clasificación, por ejemplo: la minería de datos.

Este recurso se ha vuelto una de las principales herramientas para la extracción de datos, transformando la información en estructuras comprensibles y así poder usarlas posteriormente para su clasificación y aprendizaje automático.

"La minería de datos, es usada para descubrir conocimiento útil a partir de grandes cantidades de datos. Además, el descubrimiento del conocimiento es considerado un proceso que consta de varias etapas, tales como: Comprensión del dominio, preparación del conjunto de datos, descubrimiento de patrones, análisis de patrones descubiertos y utilización de resultados, permitiendo así negocios más inteligentes desde el punto de vista estratégicos y tácticos." (Abulkari y Job, 2003).

Existe una gran cantidad de métodos que pueden ser utilizados para la clasificación de los hechos que se desean estudiar, métodos que son basados en algoritmos matemáticos y estadísticos como Regresión Lógica Bayesiana, Nnge, entre otros.

En este estudio se pretende observar estos algoritmos para estudiar los resultados que arrojan con respecto al diagnóstico de fertilidad como por ejemplo: el tiempo de ejecución, cuántos de los casos logró clasificar correctamente, qué porcentaje de error significativo obtuvo, cual es el grado de concordancia en las mediciones y con ello poder determinar qué método es de mayor efectividad y por qué resulta ser el mejor método para la clasificación de los casos obtenidos dentro de la base de datos fertility tratada anteriormente por David Gil y José Luis Gírela de la universidad de Alicante.

2. Métodos

Entre los casos de infertilidad en parejas el 50% de estos, es debido a la infertilidad masculina. Para las parejas la posibilidad de ser infértiles reduce su autoestima y llega a ser una condición difícil y estresante tanto para los clínicos como para los pacientes.

Las causas de la infertilidad masculina pueden llegar a ser por infecciones, consecuencias de cirugías, enfermedades pulmonares crónicas, traumatismos, genética, factores ambientales, o simplemente disfunción sexual.

Cuando se pretende analizar clínicamente a un paciente es aconsejable saber las causas que orillan a la consulta, y a partir de estas obtener un resultado que permita decidir si dar un tratamiento, o simplemente detectar una posible causa de origen.

Para cumplir con el objetivo propuesto se trabajó con la base de datos Fertility diagnosis extraída del repositorio UCI, esta base de datos nos muestra los resultados de 100 voluntarios que proporcionaron una muestra de semen, que fue analizada de acuerdo al criterio WHO 2010. Es importante mencionar que la concentración de esperma es relativa a las condiciones socio-demográficas, factores ambientales, estado de salud y hábitos diarios. Se toman en cuenta los siguientes 9 atributos:

Tabla 1. Descripción de atributos

Atributo	Descripción
Temporada en la cual el análisis fue hecho.	Invierno (-1) Primavera(-0.33) Verano (0.33) Otoño(1)
Edad en el momento de análisis	Se muestran los años en centésimas Por ejemplo: 18 años(0.18)
Enfermedad infantil	Son enfermedades como varicela, sarampión, paperas, etc. Solo se indica si las tuvo o no 1. Si (0) 2. No (1)
Traumatismos	De la misma manera solo se indica si los hubo o no 1. Si (0) 2. No (1)
Intervención quirúrgica	Solo se indica si las tuvo o no 1. Si (0) 2. No (1)
Fiebre alta en el último año	1. Hace menos de 3 meses (-1) 2. Hace más de 3 meses (0) 3. No(1)

Tabla 1. Descripción de atributos (continuación)

Atributo	Descripción
Frecuencia de consumo de alcohol	Valores del 0 al 1. 1. Varias veces al día 2. Todos los días 3. Varias veces a la semana 4. Una a la semana 5. Casi nunca o nunca
Habito fumador	Nunca (-1) Ocasional (0) Diario (1)
Número de horas sedentarias	Valores del 0 al 1

Analizando los datos de cada voluntario se diagnosticaba como: Normal o Alterado.

WEKA.

Para probar y comparar una serie de algoritmos de clasificación se usó una herramienta, desarrollada en la Universidad de Waikato, Nueva Zelanda, conocida como WEKA.

WEKA es una colección de algoritmos de máquinas de aprendizaje para la minería de datos, pero también pueden ser aplicados directamente a los datos o ser llamados desde un código java. WEKA contiene herramientas para el pre-procesamiento, clasificación, regresión, clustering, asociación de reglas y visualización, también es buena herramienta para desarrollar nuevas máquinas de esquemas de aprendizaje.

Fig. 1 Pantalla inicial del software de WEKA

Este sistema está escrito en Java. Ha sido probada en Linux, Windows y Macintosh. Java permite proveer una interfaz uniforme para diversos algoritmos de aprendizaje, acompañado de métodos de pre y post procesamiento. Y evaluando los resultados del aprendizaje en cualquier conjunto de datos. (Witten, y otros, 2000)

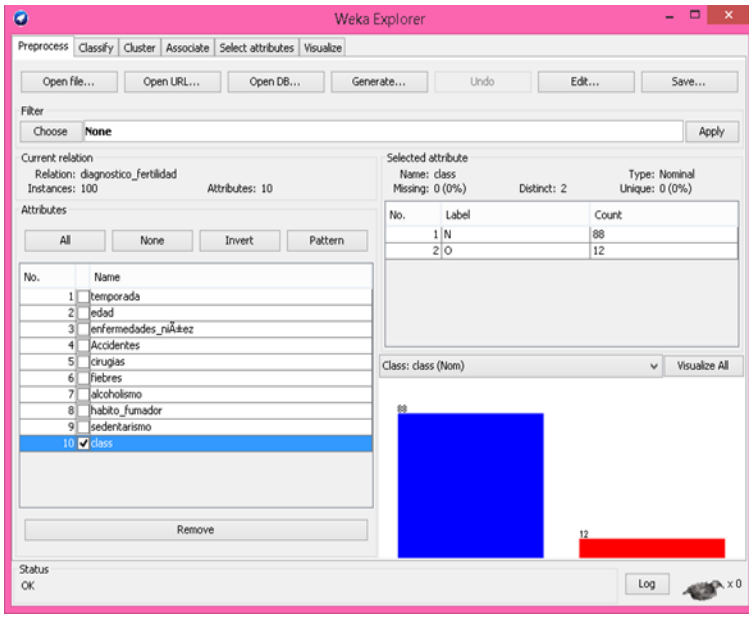

Fig. 2 Pantalla con base de datos seleccionada

Con Weka se aplicaron métodos de aprendizaje a la bases de datos fertility, y se analizaron las salidas para extraer información sobre los datos. Otra forma es aplicar

varios algoritmos de aprendizaje y comparar su ejecución para escoger uno para la predicción. Estos métodos de aprendizaje son llamados Clasificadores.

La base de datos fue tratada utilizando cuatro métodos de clasificación: NNge, Random forest, Kstar y Regresión Lógica Bayesiana.

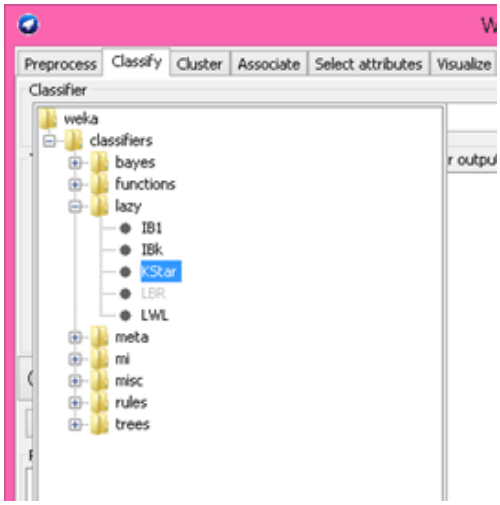

Fig. 3 Pantalla selectora de método de clasificación

Se ha propuesto dos etapas en la cuestión de la aplicación de los algoritmos de clasificación:

- Etapa de entrenamiento: la cual consiste en el proceso de aprendizaje que permita desarrollar correctamente una tarea. Durante este proceso se va refinando iterativamente la solución hasta alcanzar un nivel de operación suficientemente bueno. Este proceso se puede dividir en tres grupos:
    - Aprendizaje supervisado.
    - Aprendizaje no supervisado.
    - Aprendizaje por refuerzo.
- Etapa de operación: es el resultado de la etapa de entrenamiento, ya que finalizada esta, la red puede ser utilizada para realizar las tareas para las que fue

entrenada. La gran ventaja de este modelo es que se aprende la relación que existe entre todos los datos.

Random forest

Sus principales ventajas son:

- Es uno de los algoritmos de aprendizaje más certeros que hay disponible. Para un set de datos lo suficientemente grande produce un clasificador muy certero.
- Corre eficientemente en bases de datos grandes.
- Puede manejar cientos de variables de entrada sin excluir ninguna.
- Da estimados de qué variables son importantes en la clasificación.
- Tiene un método eficaz para estimar datos perdidos y mantener la exactitud cuando una gran proporción de los datos está perdida.
- Computa los prototipos que dan información sobre la relación entre las variables y la clasificación.
- Computa las proximidades entre los pares de casos que pueden usarse en los grupos, localizando valores atípicos, o (ascendiendo) dando vistas interesantes de los datos.
- Ofrece un método experimental para detectar las interacciones de las variables.

Un bosque aleatorio o random forest es un clasificador que consiste en una colección de árbol estructurado, sus clasificadores {h (x, k), k = 1,...} donde el {k} son independientes e idénticamente vectores aleatorios distribuidos y cada árbol arroja un voto unidad para la clase más popular en la entrada x. (Leo Breiman, 2001).

En los algoritmos basados en Random forest, un límite superior se pueden derivar para el error de generalización en términos de dos parámetros que son medidas de la precisión de los clasificadores individuales y de la dependencia entre ellos. La interacción entre estos dos da la base para la comprensión del funcionamiento de los bosques al azar. (Análisis de Amit y Geman 1997).

Nnge

Es un híbrido entre los algoritmos basados en instancias y los de inducción de reglas. Aprende incrementalmente, primero clasificando y luego generalizando cada nuevo ejemplo. La generalización consiste en fusionar la nueva instancia con el ejemplar de la misma clase más próximo. Si el ejemplar más próximo era un ejemplo aislado se crea un híper rectángulo que los contiene a ambos. De lo contrario, si el ejemplar más próximo era un híper rectángulo, este crece para abarcar el nuevo ejemplo. Los híper rectángulos se representan mediante reglas. Para determinar el vecino más cercano se utiliza una función de distancia Euclidiana modificada capaz de manejar híper rectángulos y atributos simbólicos (Sánchez Tarragó, 2007).

Kstar

K * es un clasificador basado en instancia, que es la clase de una instancia de prueba se basa en la clase de esas instancias de capacitación similares a la misma, según lo determinado por una función de similitud. Se diferencia de otros estudiantes basados en instancia en que utiliza una función de la distancia basada en la entropía. Es un método en el que se aplica una medida de similitud distinta de la euclidiana, que en la práctica pondera la influencia de los vecinos en función de su proximidad al patrón que se requiere clasificar.

Las características de este algoritmo son las siguientes:

- Permite que la clase sea simbólica o numérica.
- Admite atributos numéricos y simbólicos por cada instancia.

(G. Cleary & E. Trigg, 1995)

Regresión Lógica Bayesiana

El objetivo de la Regresión Logística es encontrar el mejor ajuste del modelo con el menor número de parámetros y describir la relación entre la variable respuesta y un conjunto de variables (covariables) explicatorias independientes.

Como problema central de este paradigma está el hecho de proporcionar una metodología que permita asimilar la información que se tiene con el objetivo de mejorar el conocimiento del mundo real. La metodología del paradigma bayesiano consta de:

- Proceso de aprendizaje, el cual constituye la base de todo problema de inferencia sobre el valor del parámetro, y se reduce básicamente a determinar su distribución posterior o final.
- Distribución predictiva, la cual es utilizada para poder describir la información que se posee sobre posibles valores de las observaciones.
- Comportamiento asintótico, análisis que se realiza y es más preciso en la medida en que se disponga de una mayor cantidad de datos.

La Regresión Logística es uno de los métodos estadísticos más expresivos y versátiles disponibles para el análisis de datos. Muchos especialistas de diferentes ramas han trabajado con ella para predecir o pronosticar una variable respuesta binaria o dicotómica, donde las variables independientes pueden ser de cualquier naturaleza, convirtiéndola en un método estándar para el análisis de regresión cuando los datos son binarios (Hosmer and Lemeshow, 1989, Silva, 1994).

La base del paradigma bayesiano es encontrar la distribución posterior del o de los parámetros o cantidades de interés en el modelo, donde resulta interesante el hecho de que los parámetros son tratados como variables aleatorias, o sea son una descripción de la incertidumbre del modelo cuantificada a través de la probabilidad (Bernardo, 2003).

El índice Kappa o estadístico Kappa.

El índice kappa (κ) se usa para evaluar la concordancia o reproducibilidad de instrumentos de medida cuyo resultado es categórico (2 o más categorías). El índice kappa (κ) representa la proporción de acuerdos observados más allá del azar respecto del máximo acuerdo posible más allá del azar. En la interpretación del índice kappa (κ) hay que tener en cuenta que el índice depende del acuerdo observado, pero también de

la prevalencia del carácter estudiado y de la simetría de los totales marginales (V. Abraira, 2000).

Este estadístico sólo puede tomar valores entre -1 y +1. Mientras más cercano a +1, mayor será el grado de concordancia, por el contrario, mientras más cercano a -1 el grado de discordancia aumenta. Cuando el valor de k es igual a 0 refleja que la concordancia observada es precisamente la que se espera a causa únicamente del azar.

El índice kappa Supongamos que dos observadores distintos clasifican independientemente una muestra de n ítems en un mismo conjunto de C categorías nominales (López y Fernández, 2001).

3. Resultados

Al realizar la clasificación del diagnóstico de fertilidad en la muestra de 100 personas se observó que el 12% de estas son infértiles, dominando sobre los resultados de la muestra que el 88% son fértiles.

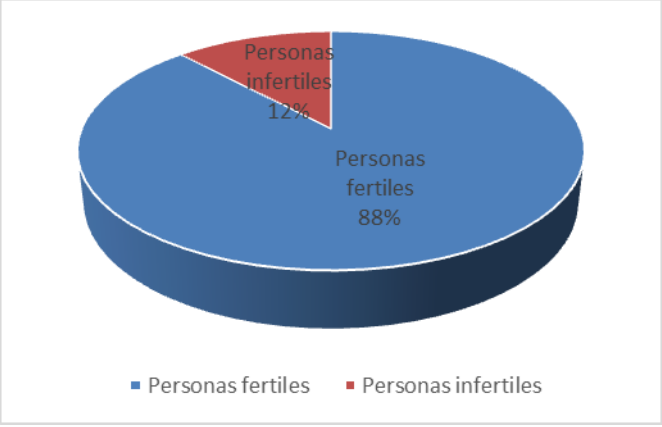

Fig. 4. Resultados de fertilidad

Después de haber sido tratada la base de datos con los métodos de clasificación se pudo observar que tres de los cuatro métodos utilizados arrojan una clasificación casi perfecta con un porcentaje de 99% de las instancias correctamente clasificadas.

Comparando inicialmente el tiempo de ejecución se aprecia que el método KStar no alcanza ni siquiera 1 segundo de ejecución, lo que lo vuelve el método con mayor velocidad de clasificación. Los métodos de NNge y de Regresión Lógica Bayesiana, se encuentran igualados respecto al tiempo de ejecución, pero al analizar las instancias correctamente clasificadas, se aprecia que la Regresión Lógica Bayesiana clasifica un menor número de instancias de la manera correcta contra el resto de los métodos, por lo cual se puede asegurar que este no resulta ser óptimo en relación a los resultados obtenidos por los demás métodos como se puede apreciar en la tabla 2.

Tabla 2. Resultados de los métodos (1)

Método	Instancias correctamente clasificadas		Instancias incorrectamente clasificadas		Estadístico de Kappa	Tiempo de ejecución
Kstar	99	99%	1	1%	0.9509	0 s
Random forest	99	99%	1	1%	0.9509	0.16 s
NNge	99	99%	1	1%	0.9543	0.03 s
Bayesian Logistic Regression	88	88%	12	12%	0	0.03 s

Ahora bien revisando el estadístico kappa, los métodos Kstar y Random forest arrojan el mismo valor, para poder determinar cuál es el mejor entre estos métodos es necesario comparar sus resultados respecto a error significativo absoluto, en donde se manifiesta que el método Random forest, proporciona un alto margen de error y por consecuente su clasificación implica un nivel de borrosidad más alto, siendo entonces Kstar el

método que tiene una mayor rango de superioridad y se puede ver reflejado en la tabla 3.

Tabla 3. Resultados de los métodos (2)

Método	Error significativo absoluto	Error cuadrático significativo	Error relativo absoluto	Error cuadrático relativo
Kstar	0.0131	0.0715	6.03%	21.99%
Random forest	0.0785	0.138	36.22%	42.45%
NNge	0.01	0.1	4.61%	30.76%
Bayesian Logistic Regression	0.12	0.3464	55.33%	106.57%

Habiendo ya aislado los métodos con menor eficacia solo queda comparar los que presentaron un mayor predominio sobre los demás. Se tiene en cuenta que el método KStar demuestra una ventaja en tiempo de ejecución sobre NNgE, sin embargo el atributo con más peso en la comparación es el estadístico de kappa y es en esto donde NNge se aproxima más al valor unitario que, como se dijo con anterioridad, es el máximo valor que se espera en el índice kappa, lo que permite indicar que NNge es el mejor método de clasificación para esta base de datos.

Apreciando la figura 5, se muestra un gráfico comparativo de las estancias correctamente clasificadas de cada uno de los métodos y se puede apreciar que existe una diferencia muy notoria entre la regresión lógica bayesiana y los demás métodos.

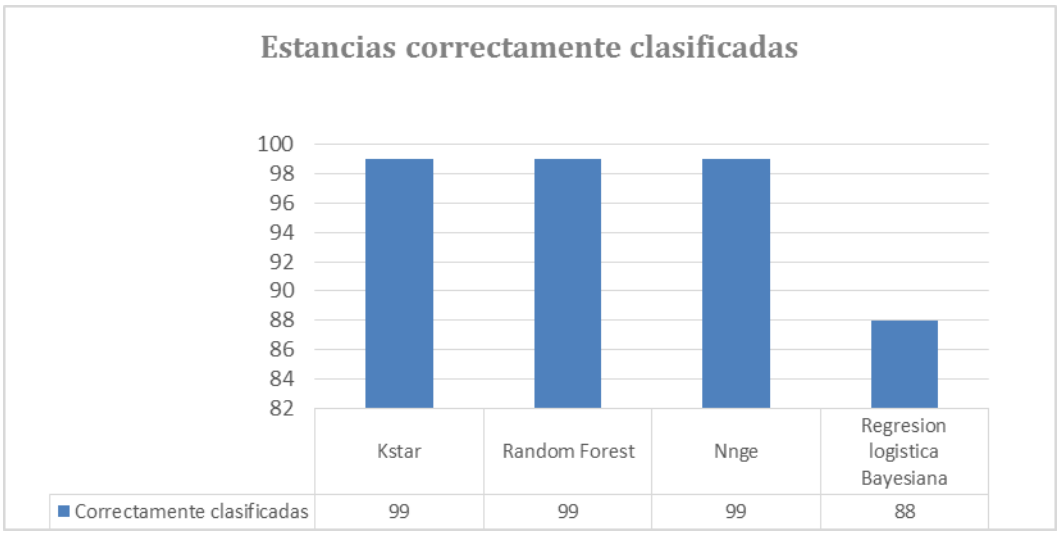

Fig. 5. Cantidad de estancias correctamente clasificadas.

Después de descartar el método de regresión logística bayesiana, se analizaron los datos significativos de los métodos restantes para determinar cuáles proporcionaban mejores resultados.

Como se muestra en la siguiente figura se destaca que el método Random Forest es el que arroja el mayor error cuadrático significativo, por lo que se puede decir que el mejor método se encuentra entre Kstar y NNge.

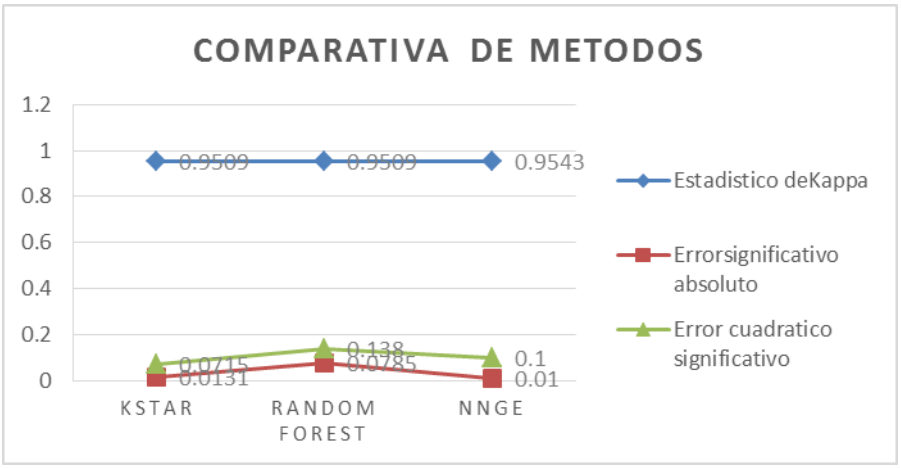

Fig. 6. Tabla de datos de métodos con mayor porcentaje de éxito.

En las dos tablas posteriores se separaron los datos de los dos métodos que podrían ser los que mejor porcentaje de resultado arrojarán, para poder analizarlos y compararlos entre ellos.

Tabla 4. Métodos Kstar y Nnge (1).

Método	Instancias correctamente clasificadas		Instancias incorrectamente clasificadas		Estadístico de Kappa
Kstar	99	99%	1	1%	0.9509
NNge	99	99%	1	1%	0.9543

Tabla 5. Métodos Kstar y Nnge (2).

Método	Error significativo absoluto	Error cuadrático significativo	Error relativo absoluto	Error cuadrático relativo	Tiempo de ejecución en segundos
Kstar	0.0131	0.0715	6.03%	21.99%	0
NNge	0.01	0.1	4.61%	30.76%	0.03

Para determinar cuál de estos métodos es el que resulta más efectivo se realizó otro gráfico donde se pudiera apreciar con más claridad la ventaja de uno sobre otro.

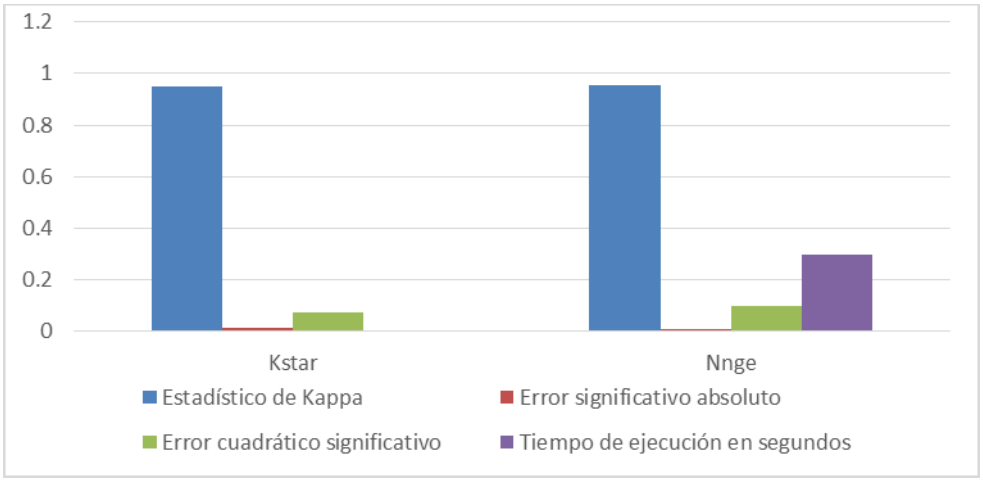

Fig. 7. Datos de Métodos Kstar y NNge.

Finalmente en el gráfico comparativo entre los métodos Kstar y NNge claramente se aprecia la ventaja que tiene el método Kstar en cuanto a tiempo de ejecución, sin embargo debe notarse que el estadístico Kappa es mayor en el método NNge. Entonces como se ha dicho anteriormente, se debe tener claro que mientras más se acerque nuestro índice Kappa a la unidad, nos indica que los elementos de la base de datos tienen mayor grado de concordancia entre sí.

4. Discusión

Primeramente al incursionar por el software de Weka se observó que es un programa de fácil y sencillo uso, muy recomendable para tratar bases de datos de todo tipo inclusive las que son muy extensas. Gracias a que la base de datos de fertilidad contaba solo con 100 instancias la velocidad de clasificación fue muy rápida de calcular en todos los métodos que fueron seleccionados.

La principal dificultad que se presenta al realizar este tipo de proyecto de clasificación de bases de datos es el tratado de las mismas, ya que entre más instancias y atributos contenga será más laborioso de tratar.

Respecto a los algoritmos utilizados se concluye que de acuerdo al desempeño dictaminado por el porcentaje de clasificación correcta, los algoritmos Nnge, Random Forest y Kstar tienen un porcentaje casi exacto con un 99%, lo que los vuelve muy viables para su uso. De acuerdo al tiempo de ejecución el algoritmo con menor tiempo de ejecución es el Kstar con 0 segundos.

Concluyendo finalmente con toda la información presentada se concretó que el algoritmo más efectivo es el Nnge, ya que a pesar de que no es el más veloz su estadístico de Kappa (0.9543) es más significativo que todos los demás algoritmos probados, añadiendo que también es el que presenta menor porcentaje de errores, lo que hace notar que aprendió correctamente de la base de datos con un mínimo margen de error.

Bibliografía

[1] Caruana, R., Karampatziakis, N., & Yessenalina, A. (2008). An Empirical Evaluation of Supervised Learning in High Dimensions. 25th International Conference on Machine Learning (ICML) (pág. 8). Ithaca, NY: Deparment of Computer Science, Cornell University.

[2] Breiman L. & Cluter A. (20 de Enero del 2004). University of California, Berkeley: Department of Statistics. Obtenido de https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.html

[3] Abraira, Dr. V. (Mayo de 2001). El índice kappa. Notas estadísticas, 27(5), 249.

[4] Breiman, L. (2001). Random Forests. Machine Learning (45), 5-32.

[5] Lorena Pradenas Rojas, Carlos Parra. (Septiembre a Diciembre del 2012). Uso de minería de datos para determinar la disponibilidad de una red IPv4 en una cadena de terminales distribuidos. Estudio de caso en una empresa de juegos de azar. PODes, 4, 270.

[6] Alejandro D. Teppa-Garrán, Anselmo Palacios-Torres. (2004) Evaluación actual de la fertilidad masculina. Investigación Clínica. 45(4), 16.