Métodos de clasificación: Análisis de fertilidad

Pedro Javier Corrales Gasca
Instituto Tecnológico de Celaya
peter-13-jcg@hotmail.com

Jorge Armando Razo Granados
Instituto Tecnológico de Celaya
jarmandorazo@hotmail.com

Miguel Ángel Violante Hernández
Instituto Tecnológico de Celaya
anyel_31@live.com.mx

Iván de Jesús Rodríguez Hernández
Instituto Tecnológico de Celaya *ivan-jrh@hotmail.com*

Rodrigo Ramírez Soto
Instituto Tecnológico de Celaya
rodrigo.ramirez02@gmail.com

Resumen

El presente artículo muestra los resultados obtenidos en un estudio de fertilidad realizado a 100 voluntarios del sexo masculino que otorgaron una muestra de semen para su estudio, esta información fue tomada de una Base de datos multivariada del repositorio UCI, posteriormente se realizó el análisis de esta información con el software WEKA, en el cual se emplearon diferentes algoritmos de clasificación como lo son: BayesianLogisticRegression, MultilayerPerceptron, ClassificationViaRegression, KStar, REPTree, con la finalidad de determinar cuál de ellos resulta ser más efectivo de acuerdo a criterios como los tiempos de clasificación y la efectividad de los mismos.

Palabra(s) Clave(s): BayesianLogisticRegression, ClassificationViaRegression, KStar MultilayerPerceptron, REPTree.

Abstract

The present article shows the results obtained in a study of fertility realized to 100 volunteers of the masculine sex who granted a semen sample for its study, the analysis of the obtained information was realized by the software WEKA in which there were used different algorithms of classification as it they are: BayesianLogisticRession, MultilayerPerceptron, ClassificationViaRegression, KStar, REPTree, for the purpose of determines which of them turns out to be more effective in accordance with criteria like the times of classification and the effectiveness of the same ones.

Keyboards: BayesianLogisticRegression, ClassificationViaRegression, KStar MultilayerPerceptron, REPTree.

1. Introducción

A través de la historia y las civilizaciones el ser humano ha expresado un gran interés en conocer las causas de la infertilidad. Para ello se ha valido de diversos medios que lo lleven a entender un poco más sobre aquello que lo origina. Actualmente y gracias a la tecnología se ha hecho cada vez más sencillo obtener todo tipo de información en cualquier momento, sin embargo esto a su vez ha originado una gran responsabilidad en cuanto a la aplicación de buenas prácticas que ayuden a la manipulación de información y a la extracción de conocimiento de estos registros, que permita elegir entre diversas alternativas la más adecuada para cada situación.

En este artículo se presentan los resultados obtenidos de diversas pruebas de clasificación realizadas a una base de datos multivariada denominada “Análisis de fertilidad” extraída del repositorio UCI [1], buscando encontrar el algoritmo de clasificación con mayor acierto, así como reducir el tiempo de ejecución ante cualquier proceso relacionado con la minería de datos, siendo estos los objetivos fundamentales de la clasificación.

Langley y Simon (1998) opinaban en su artículo que “los objetivos de las máquinas de aprendizaje eran proporcionar un incremento del nivel de autonomía en los procesos de ingeniería del conocimiento con técnicas automáticas que mejorasen los porcentajes de acierto y la eficiencia, por medio del descubrimiento y el uso de reglas sobre los datos.

El último test de una máquina de aprendizaje es su habilidad para producir sistemas que sean usados regularmente en la industria, en la educación y en cualquier otro campo".

Así mismo, se ha comprobado que la estratificación ayuda a disminuir considerablemente el tiempo de evaluación de un conjunto de muestras de datos, mejorando también su escalabilidad, siendo que esta solamente es una estrategia de ordenación de los datos de entrada, aplicada en los métodos de clasificación. Por otra parte en cuanto a los algoritmos de aprendizaje automático se refiere, se deberán elegir adecuadamente los atributos con la finalidad de generar las decisiones, ya que si se introducen datos irrelevantes, el algoritmo se confundirá y su desempeño se verá afectado.

Por todo lo expuesto anteriormente, surge la necesidad de conocer el funcionamiento de cada algoritmo de clasificación, ya que no existe ninguno que funcione adecuadamente para cualquier conjunto de datos, además debe considerarse la realización de un ajuste de parámetros previos a su clasificación. Por último una buena práctica a implementar durante cada análisis de datos, es combinar los algoritmos de clasificación con la intención de lograr mejores resultados más fiables y exactos.

2. Métodos y conceptos

Fertilidad

La fertilidad constituye hoy en día uno de los aspectos más importantes de nuestra vida. El hecho cotidiano de concebir un niño contrasta con el hecho de que la reproducción humana es una de las de menor eficacia, según los expertos, las posibilidades de quedar embarazada en una mujer normal son sólo del 30% cada mes. Este porcentaje es mayor cuando las mujeres son más jóvenes y se reduce con la edad.

La fertilidad humana es la capacidad de producir o sustentar una descendencia numerosa, resultado de diversos factores de índole biológica (edad, estado de salud, funcionamiento del sistema endocrino), cultural (prescripción sobre sexo y matrimonio, división del trabajo, tipo y ritmo de ocupación), que provocan las abruptas variaciones de una situación a otra.

La fertilidad cambia conforme pasa el tiempo. El ser humano se vuelve fértil en la adolescencia, después de su paso por la pubertad. En el caso de las mujeres, el inicio de su vida reproductiva está marcado por el inicio de la ovulación y la menstruación.

Normalmente después de la menopausia las mujeres dejan de ser fértiles. En general, las posibilidades de reproducción disminuyen a medida que se envejece y normalmente la fertilidad finaliza entre cinco y diez años antes de la menopausia.

Minería de datos

[2] La minería de datos abarca todo un conjunto de técnicas enfocadas en la extracción de conocimiento implícito en las bases de datos. Las bases de la minería de datos se encuentran en la inteligencia artificial y en el análisis estadístico. Mediante los modelos extraídos utilizando técnicas de minería de datos se aborda la solución a problemas de predicción, clasificación y segmentación. Un proceso típico de minería de datos consta de los siguientes pasos generales:

- **Selección del conjunto de datos:** Se refiere tanto a la selección de variables dependientes, variables objetivo, y al muestreo de los registros disponibles.
- **Análisis de las propiedades de los datos:** Se refiere en especial a la representación de los datos por medio de graficas como: histogramas, diagramas de dispersión, presencia de valores atípicos y ausencia de datos (valores nulos).
- **Transformación del conjunto de datos de entrada:** Se refiere al conjunto de operaciones realizadas para preparar los datos de análisis, con el objetivo de adaptarlos para aplicar la técnica de minería de datos que mejor se adapte al problema.
- **Seleccionar y aplicar la técnica de minería de datos:** Dependerá de la naturaleza del problema a resolver. Para poder implementar la técnica seleccionada, se debe proceder a elegir algún software que facilite el trabajo de aprendizaje automático.
- **Evaluar los resultados:** Se contrasta con un conjunto de datos (datos de entrenamiento) previamente reservados para validar la generalidad del modelo.

Algoritmos de clasificación

[3] Los algoritmos dedicados al problema de la clasificación supervisada operan usualmente sobre la información suministrada por un conjunto de muestras, patrones, ejemplos o prototipos de entrenamiento que son asumidos como representantes de las

clases, y los mismos poseen una etiqueta de clase correcta. A este conjunto de prototipos correctamente etiquetados se le llama conjunto de entrenamiento (TS, training set), y es el conocimiento empleado para la clasificación de nuevas muestras.

Estos algoritmos tienen como objetivo determinar cuál es la clase, de las que ya se tiene conocimiento, a la que debe pertenecer una nueva muestra, teniendo en cuenta la información que se puede extraer del conjunto de entrenamiento.

Clasificadores comparativos

La clasificación se lleva a cabo de forma diferente según el método utilizado, sin embargo todos parten de lo que es un mismo conjunto de datos, terminando en lo que es la predicción correspondiente al porcentaje de acierto.

Clasificadores utilizados

Bayesiano

Es una proposición planteada por el filósofo inglés Thomas Bayes , que expresa la probabilidad condicional de un evento aleatorio. [4] La clasificación mediante el algoritmo Bayesiano ofrece la solución óptima al problema, incluyendo la característica fuzzy de la probabilidad de pertenencia de cada muestra a todas las clases. Para la evaluación de esta regla se requiere un conocimiento a priori de la probabilidad y de la densidad de las clases.

KStar

[5] Este algoritmo determina cuáles son las instancias más parecidas, puede utilizar la entropía, o contenido de información de las instancias, como medida de distancia entre ellas. Son destacables las siguientes características:

- Admite atributos numéricos y simbólicos, así como pesos por cada instancia.
- Permite que la clase sea simbólica o numérica.
- En el caso de que se trate de una clase numérica se empleará la siguiente ecuación para predecir el valor de un ejemplo de test:

$$v(a) = \frac{\sum_{i=1}^{n} P * (b|a) * v(b)}{\sum_{i=1}^{n} P * (b|a)}$$

Donde v(i) es el valor (numérico) de la clase para el ejemplo i, n el número de ejemplos de entrenamiento, y P*(i|j) la probabilidad de transformación del ejemplo j en el ejemplo i. Proporciona cuatro modos de actuación frente a pérdidas en los atributos en ejemplos de entrenamiento. Para el cálculo de los parámetros x0 y s permiten basarse en el parámetro b o en el cálculo de la entropía.

Multilayer Perceptron

[6] Los perceptrones multicapa son redes compuestas por multitud de unidades llamada neuronas e interconectadas entre sí. Las neuronas no son más que elementos que proporcionan una salida en función de sus entradas, a las que le aplican una función predeterminada para la obtención de dicha salida.

Las funciones que aplica una neurona a sus entradas para obtener la salida suele ser sencilla y simple, pero una red neuronal, debido a la asociación de distintas de estas neuronas, puede adoptar el comportamiento de una función extremadamente compleja, haciéndose más compleja conforme añadimos capas de neuronas o neuronas a dichas capas.

REPTree

[7] Funcionamiento en dos fases (datos de aprendizaje y datos de poda). Primero se crea un conjunto de reglas que se sobreajusta a los datos usados para el aprendizaje, después se poda el conjunto de reglas usando ejemplares que no participaron en el aprendizaje.

WEKA

[8] WEKA se distribuye como software de libre distribución desarrollado en Java. Está constituido por una serie de paquetes de código abierto con diferentes técnicas de preprocesado, clasificación, agrupamiento, asociación, y visualización, así como facilidades para su aplicación y análisis de prestaciones cuando son aplicadas a los datos de entrada seleccionados. [2] Las principales herramientas de Weka son:

- **Simple CLI:** la interfaz "Command-Line Interfaz" es simplemente una ventana de comandos java para ejecutar las clases de WEKA.
- **Explorer:** es la opción que permite llevar a cabo la ejecución de los algoritmos de análisis implementados sobre los ficheros de entrada, una ejecución independiente por cada prueba. Esta es la opción sobre la que se centra la totalidad de esta guía.
- **Experimenter:** esta opción permite definir experimentos más complejos, con objeto de ejecutar uno o varios algoritmos sobre uno o varios conjuntos de datos de entrada, y comparar estadísticamente los resultados.
- **KnowledgeFlow:** esta opción permite llevar a cabo las mismas operaciones del "Explorer", con una configuración totalmente gráfica, inspirada en herramientas de tipo "data-flow" para seleccionar componentes y conectarlos en un proyecto de minería de datos, desde que se cargan los datos, se aplican algoritmos de tratamiento y análisis, hasta el tipo de evaluación deseada.

3. Metodología

En este trabajo se utilizó la herramienta Weka para realizar el análisis de los datos. Se tomó como muestra de estudio la base de datos “Análisis de fertilidad” del repositorio UCI, esta base de datos contiene información recabada de 100 voluntarios que otorgaron una prueba de semen que se analizó de acuerdo a los criterios de WHO 2010. La concentración de esperma está relacionada con datos sociodemográficos, factores ambientales, estado de salud y hábitos de vida. Como parte de la metodología se aplicaron dos diferentes tipos de test a cada algoritmo, siendo estos el test de Crossvalidation que realiza una validación cruzada en el que dado un número n se dividen los datos en n partes y, por cada parte, se construye el clasificador con las n−1 partes restantes y se realiza la prueba, y el test Use training set en el que Weka entrena el método con todos los datos disponibles y luego lo aplicara otra vez sobre los mismos, a partir de esto se definirá cuál de los diferentes test y algoritmos arrojan los mejores resultados en cuestión de clasificación y tiempo. La información de los atributos de esta base de datos se detalla a continuación:

a) Temporadas en que los análisis se realizaron. En tabla 1 se muestra la clasificación de los datos respecto a la temporada del año en que se realizaron los análisis.

Tabla 1 Periodo de análisis.

Invierno	Primavera	Verano	Otoño
-1	-0.33	0.33	1

b) Edad al momento del análisis. La tabla 2 representa el rango de edad de las personas que se sometieron al análisis, siendo este rango de los 18 a los 36 años de edad.

Tabla 2 Edad al momento del análisis.

18-36	0-1

c) Enfermedades de la infancia. En tabla 3 se presenta la clasificación de los datos respecto a las posibles enfermedades que pudieron haber presentado los voluntarios al análisis en su niñez.

Tabla 3 Enfermedades de la infancia.

	Viruela	**Sarampión**	**Paperas**	**Polio**
Si (0)				
No (1)				

d) Accidentes con lesiones físicas. La tabla 4 contiene información respecto a posibles lesiones físicas que el voluntario presente al momento del análisis.

Tabla 4 Accidentes con lesiones físicas.

	Heridas o afecciones gravedad de	**Cirugías**
Si (0)		
No (1)		

e) Fiebre alta en el último año. La tabla 5 clasifica los datos respecto a si ha presentado fiebre el último año y la frecuencia del mismo padecimiento.

Tabla 5 Fiebre alta en el último año.

Hace menos de 3 meses	Hace más de 3 meses	No
-1	0	1

f) Frecuencia de consumo de alcohol. En la siguiente tabla se presenta la clasificación de la información respecto a la frecuencia en que el voluntario consume alcohol.

Tabla 3.6. Frecuencia de consumo de alcohol.

Varias veces al día	Diario	Varias veces a la semana	Una vez a la semana	Rara vez o nunca

g) Tabaquismo. La tabla 7 contiene la información sobre la frecuencia en que el voluntario consume tabaco.

Tabla 7 Tabaquismo.

Nunca	Ocasionalmente	Diario

h) Número de horas que permanece sentado. La tabla 8 contiene información sobre el tiempo que el voluntario sometido al análisis pasa sentado.

Tabla 8 Número de horas que permanece sentado.

1-16	(0,1)

i) Salida. Por último, la tabla 9 contiene los resultados obtenidos al final del análisis realizado.

Tabla 9 Salida.

Diagnostico Normal	Alterado

4. Resultados

En las tablas 4.1 y 4.5 presentadas a continuación se detalla el porcentaje de instancias clasificadas correctamente, así como el de las instancias mal clasificadas. El estadístico

kappa expresa la medición de la coincidencia de la predicción con la clase real, siendo que el valor 1.0 indica que ha habido coincidencia absoluta. Las siguientes columnas expresan el resultado del nivel de error generado del modelo al haber aplicado el algoritmo correspondiente, finalmente la última columna refleja el tiempo que se ha tardado cada algoritmo en clasificar la información.

Como podemos ver en la tabla 10 se aplicó el test Cross-validation dando como resultado que el algoritmo de mejor desempeño es MultilayerPerceptron, al haber obtenido el 90 % de instancias bien clasificadas y un tiempo de procesamiento de 0.39 segundos, lo que lo posiciona como el mejor de todos los algoritmos empleados en este test. El motivo por el que se ha determinado a este algoritmo como el más óptimo es debido a que representa el porcentaje de error cuadrático (figura 1) más bajo, siendo este el objetivo principal de las redes neuronales minimizar este error.

Tabla 10 Clasificación de algoritmos con el test Cross-validation (100 instancias).

Algoritmo	Instancias bien clasificadas. (%)	Instancias mal clasificadas. (%)	Kappa statistic	Error absoluto Medio	Raíz del error cuadrático medio	Error absoluto relativo. (%)	Raíz del error cuadrático relativo	Tiempo
BayesianLogisticRegression	88	12	0	0.12	0.364	55.08	106.38	0.02
MultilayerPerceptron	90	10	0.4898	0.1287	0.3123	59.07	95.9	0.39
ClassificationViaRegression	86	14	-0.0355	0.2203	0.3466	101.14	106.42	0.17
KStar	84	16	0.1837	0.1758	0.3582	80.72	109.99	0
REPTree	84	16	-0.0638	0.2267	0.3581	104.06	109.96	0

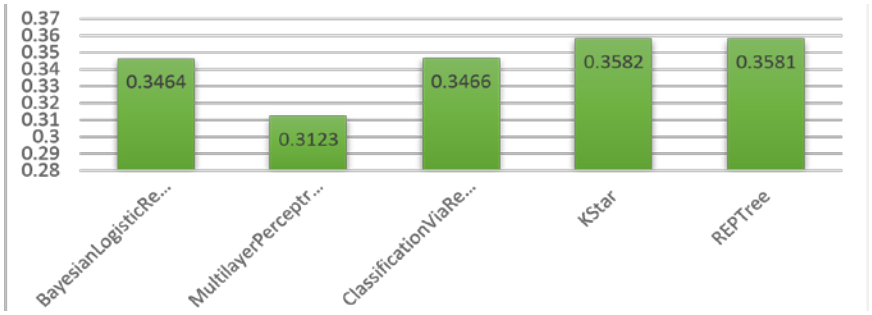

Figura 1 Error cuadrático Medio test Cross-validation (100 instancias).

En la figura 2 se puede observar el porcentaje de instancias clasificadas correctamente siendo el algoritmo MultilayerPerceptron el algoritmo con más instancias clasificadas correctamente.

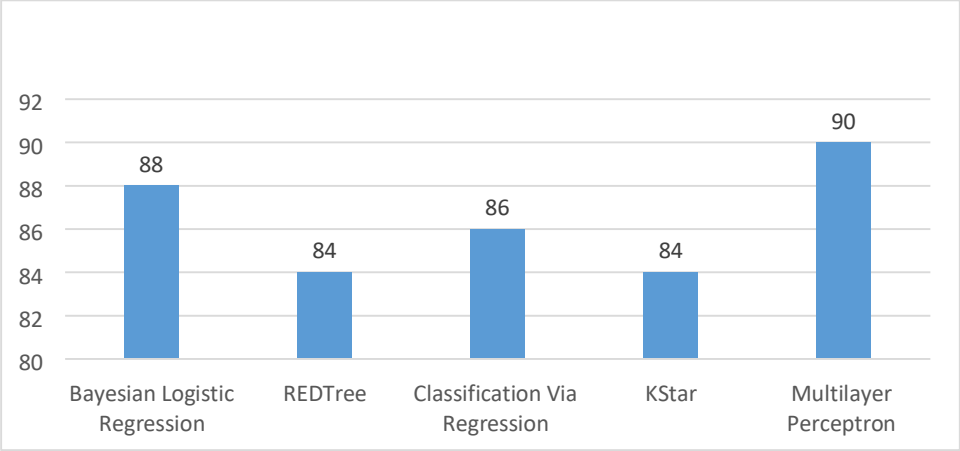

Figura 2 Instancias clasificadas correctamente.

En la figura 3 se puede observar el porcentaje de instancias clasificadas incorrectamente siendo el algoritmo Multilayer Perceptron el algoritmo con menos instancias clasificadas incorrectamente.

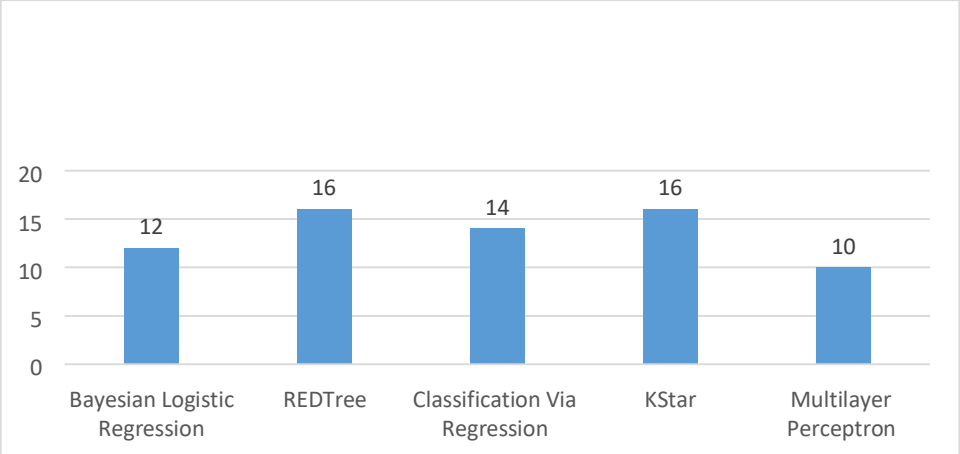

Figura 3 Instancias clasificadas incorrectamente.

Como podemos ver en la tabla 11 se aplicó el test Use training set dando como resultado que el algoritmo de mejor desempeño es KStar, al haber obtenido el 99 % de instancias bien clasificadas y un tiempo de procesamiento de 0 segundos, lo que lo posiciona como el mejor de todos los algoritmos empleados en este test.

Tabla 11 Clasificación de algoritmos con el test Use training set (100 instancias).

Algoritmo	Instancias bien clasificadas. (%)	Instancias mal clasificadas. (%)	Kappa statistic	Error absoluto Medio	Raíz del error cuadrático medio	Error absoluto relativo. (%)	Raíz del error cuadrático relativo	Tiempo
BayesianLogisticRegression	88	12	0	0.12	0.3464	55.33	106.57	0.02
MultilayerPerceptron	96	4	0.7788	0.0494	0.2014	22.77	61.94	0.48
ClassificationViaRegression	88	12	0	0.1712	0.2832	78.93	87.13	0.31
KStar	99	1	0.9509	0.0131	0.0715	6.02	21.99	0
REPTree	88	12	0	0.2112	0.325	97.38	99.97	0

Sin embargo podemos analizar también el desempeño obtenido por los algoritmos MultilayerPerceptron y Reptree que en cuestión de instancias bien clasificadas el primero de ellos tiene el 96% de instancias bien clasificadas pero su tiempo de procesamiento es ligeramente superior al obtenido por Reptree. El motivo por el que se ha determinado a este algoritmo como el más óptimo es debido a que representa el porcentaje de error cuadrático (figura 4) más bajo, siendo este el objetivo principal de las redes neuronales minimizar este error.

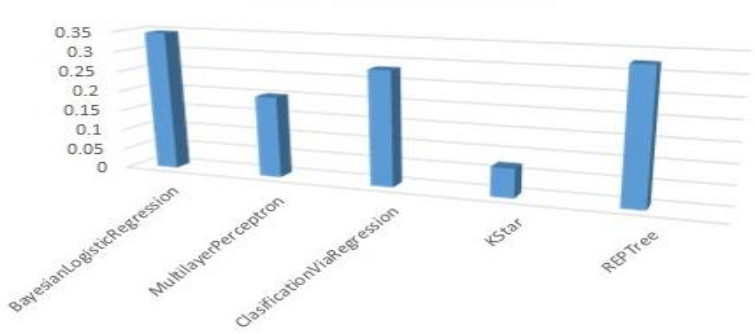

Figura 4 Error cuadrático Medio test Use training set (100 instancias).

En la figura 5 se puede observar el porcentaje de instancias clasificadas correctamente siendo el algoritmo MultilayerPerceptron el algoritmo con más instancias clasificadas correctamente.

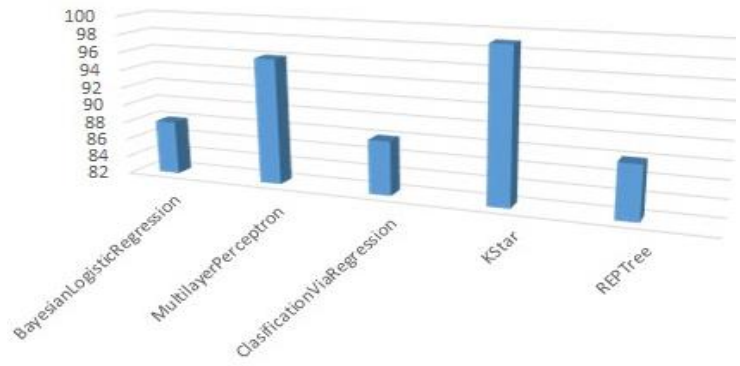

Figura 5 Instancias clasificadas correctamente.

En la figura 6 se puede observar el porcentaje de instancias clasificadas incorrectamente siendo el algoritmo MultilayerPerceptron el algoritmo con más instancias clasificadas incorrectamente.

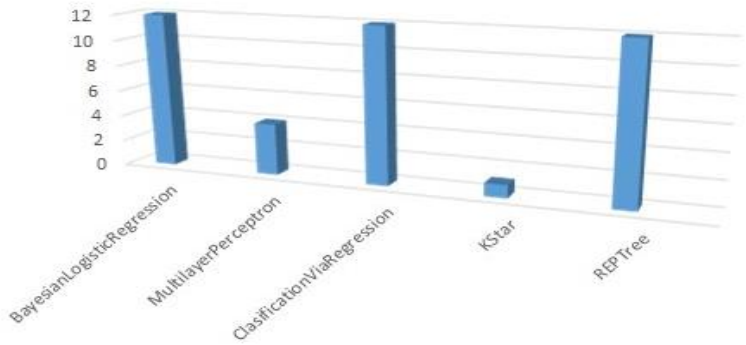

Figura 6 Instancias clasificadas incorrectamente.

4. Discusión

Se puede definir que el test que mayor confianza representa respecto a su desempeño es Use Training set al elevar considerablemente las instancias bien clasificadas y al hacer coincidir de mejor forma la predicción con la clase real, por otro lado el algoritmo con mejor desempeño ha sido KStar al tener un 99% de instancias bien clasificadas, así mismo se ha logrado determinar que la elección del algoritmo más eficiente bajo el criterio de mejor clasificación no se ve afectado por el tiempo de procesamiento de la información, ya que el número de instancias determinadas por el conjunto de datos no influye considerablemente en el aumento de los tiempos de procesamiento y la

efectividad de cada análisis realizado depende principalmente de una muestra bien definida y significativa de los datos a analizar, además de una correcta estratificación de los datos, así como de la elección correcta del test que realizara la prueba, por último el desempeño del algoritmo de clasificación depende de las reglas específicas para lo que ha sido programado y que de acuerdo a las características de la muestra a tratar el desempeño del mismo puede variar. Por ello es importante conocer el funcionamiento de cada uno de ellos y para lo que han sido diseñados, para que en base a ello se pueda elegir el más apropiado para cada conjunto de datos. En lo que a nosotros respecta consideramos que debe buscarse la realización de un algoritmo cuyo ajuste inicial de sus parámetros genere porcentajes de aciertos tan elevados como sea posible y que se ajuste a la mayor cantidad de muestras posibles, con la finalidad de eliminar el debate continuo sobre que clasificador elegir.

5. Bibliografía

[1] David Gil, José Luis Girela, Joaquín De Juan, M. José Gómez-Torres, y Magnus Johnsson. La predicción de la calidad seminal con inteligencia artificial métodos. Sistemas Expertos con Aplicaciones

[2] López Molina José Manuel, Herrero José García,"Técnicas de Análisis de Datos.Aplicaciones prácticas utilizando Microsoft Excel y Weka", 2006

[3] EcuRed Conocimiento con todos y para todos http://www.ecured.cu /index.php/Algoritmos_de_clasificaci%C3%B3n_supervisada.

[4] RemcoBouckaert, http://classes.engr.oregonstate.edu/eecs/winter2003/cs534/ weka/weka-3-3-4/doc/weka.classifiers.bayes.BayesNet.html.

[5] Dánel Sánchez Tarragó. Departamento de Informática del Instituto Superior de Ciencias Médicas de Villa Clara. www.informatica2007.sld.cu/.../2006-11 15.5808751092/download- [5]José Manuel Molina López, Jesús García Herrero http://scalab.uc3m.es/~docweb/ad/transparencias/apuntesAnalisisDatos.pdf

[6] Data Mining, Apuntes de la asignatura. 2º Ingeniería Informática. Universidad de Jaén http://wwwdi.ujaen.es/asignaturas/dm/tema8.pdf

[7] Data Mining, Apuntes de la asignatura. 2º Ingeniería Informática. Universidad de Jaén, http://wwwdi.ujaen.es/asignaturas/dm/tema8.pdf [8] Capitulo 1. Técnica de Análisis de Datos en Weka.