REVISTA VIRTUAL
Universidad Católica del Norte
CATÓLICA DEL NORTE Fundación Universitaria
Pioneros en educación virtual
http://revistavirtual.ucn.edu.co
No. 49 | Septiembre - Diciembre, 2016
ISSN-0124-5821 (en línea)

Cómo citar el artículo

Bedoya, O. M., López Trujillo, M. & Marulanda Echeverry, C. E. (2016). Minería de datos en egresados de la Universidad de Caldas. *Revista Virtual Universidad Católica del Norte, 49,* 110-124. Recuperado de http://revistavirtual.ucn.edu.co/index.php/RevistaUCN/article/view/800/1320

# Minería de datos en egresados de la Universidad de Caldas*

**Oscar Mauricio Bedoya**

Ingeniero de sistemas

Candidato a magíster en ingeniería computacional

Profesor, Departamento de Sistemas e Informática, Universidad de Caldas

oscar.bedoya@ucaldas.edu.co

**Marcelo López Trujillo**

Ingeniero de Sistemas

Magister en educación

Doctor en ingeniería informática, sociedad de la información y del conocimiento

Profesor, Departamento de Sistemas e Informática, Universidad de Caldas

Profesor, Departamento de Administración, Universidad Nacional de Colombia, Sede Manizales.

mlopez@ucaldas.edu.co y mlopeztr@unal.edu.co

* Investigación: proyecto de investigación titulado *Desarrollo de un modelo de inteligencia de negocios organizacional, estudio de caso*. Universidad de Caldas, Grupo GITIR, Departamento de Sistemas e informática, Universidad de Caldas, Calle 65 n.° 26-10, Manizales (Colombia).

**Carlos Eduardo Marulanda Echeverry**

Ingeniero Industrial

Especialista en diseño y manufactura asistida por computador

Magister en administración

Candidato a doctor en ingeniería, industria y organizaciones

Profesor, Departamento de Sistemas e Informática, Universidad de Caldas

Profesor, Departamento de Administración, Universidad Nacional de Colombia, Sede Manizales.

carlose@ucaldas.edu.co y cemarulandae@unal.edu.co

**Recibido:** 20 de noviembre de 2015.
**Evaluado:** 27 de julio de 2016.
**Aprobado:** 8 de agosto de 2016.

**Tipo de artículo:** investigación científica y tecnológica.

## Resumen

Este artículo presenta los resultados del uso de técnicas de clasificación en minería de datos de los factores asociados a la percepción que el recién egresado de la Universidad de Caldas tiene de la utilidad de los conocimientos y destrezas adquiridos a lo largo de sus estudios, que forman parte vital en su rol laboral. Para su desarrollo se utilizaron enfoques investigativos como el exploratorio y el descriptivo, conjuntamente con cuatro técnicas de minería de datos de clasificación y un repositorio de datos con información del entorno social, personal y familiar, académico, laboral y de percepción de utilidad de habilidades frente al reto profesional. El rasgo del egresado determina que las habilidades y destrezas adquiridas durante sus estudios en la Universidad de Caldas son muy útiles; se espera que constituyan un aporte significativo en la búsqueda de mecanismos que mejoren el nivel de satisfacción de los egresados con su formación y la pertinencia de los planes de estudio.

## Palabras clave

Egresados, Minería de datos, Técnicas de clasificación.

## Data Mining in Graduates of the University of Caldas

This article presents the results of the use of classification techniques in data mining of the factors associated to the perception of the newly graduates of the University of Caldas, Colombia, regarding the usefulness of the knowledge and skills acquired in their careers which constitutes an important part in their work role. Exploratory and descriptive research approaches were used along with four data mining techniques of classification and a data repository with information of their social, personal, familiar, academic and work context and data about usefulness perception regarding the professional challenges. The characteristic of the newly graduates shows that the skills and competencies acquired during their education in the University of Caldas are very useful; we hope that these features will constitute a significant contribution in the search of mechanisms that improve the satisfaction level of the graduates with their education and the appropriateness of their curriculums.

## Keywords

Graduates, Data mining, Classification techniques.

## Exploration de données chez les diplômés de l'Université de Caldas

## Résumé

Cet article présente les résultats de l'usage de techniques de classification dans exploration de données avec les facteurs liés à la perception qui ont les nouveaux diplômés de l'Université de Caldas, Colombie, au sujet de l'utilité de les connaissances acquis pendant leurs études, celui qui est une partie très importante dans leur rôle au travail. Pour réaliser cette recherche on a utilisé des approches comme l'exploratoire et le descriptif, avec quatre techniques d'exploration de données avec information du contexte social, personnel et familial, académique, de travail et de perception de l'utilité

de compétences par rapport aux défis professionnels. Le trait des diplômés détermine que les compétences et habiletés acquis pendant leurs études dans l'Université de Caldas sont très utiles ; on espère qu'ils constituent une contribution significative dans la recherche de moyens qui améliorent le niveau de satisfaction des diplômés avec leur éducation et la pertinence des curriculums.

**Mots-clés**
Diplômés, Exploration de données, Techniques de classification.

## Introducción

Según Pérez-Palacios, Caballero, Caro, Rodríguez y Antequera (2014), la minería de datos es una parte importante de un proceso más amplio conocido como descubrimiento de conocimiento en bases de datos (KDD por sus iniciales en inglés). El objetivo principal de la minería de datos consiste en extraer información oculta de un conjunto de datos. Esto puede ser alcanzado por el análisis automático o semiautomático de gran cantidad de datos, lo que permite la extracción de patrones desconocidos. Estos patrones pueden ser grupos de registros de datos (análisis clúster), inusuales registros (detección de anomalías) y dependencias entre datos (asociación de reglas). Por lo tanto, los patrones pueden ser vistos como un resumen de los datos de entrada, y se pueden utilizar para su posterior análisis.



Tsai (2013) complementa lo anterior en tanto explica que la minería de datos es un campo interdisciplinario que combina la inteligencia artificial, la gestión de bases de datos, visualización de datos, aprendizaje automático, algoritmos matemáticos y estadísticos. Esta tecnología ofrece diferentes metodologías para la toma de decisiones, resolución de problemas, el análisis, la planificación, el diagnóstico, la detección, la integración, la prevención, el aprendizaje y la innovación. En esta misma línea, Natek y Zwilling (2014) explican los pasos para el análisis de la MD: primero, crear los datos conjuntos; segundo, definir la herramienta de MD a utilizar; tercero, evaluar las técnicas de MD a utilizar; y cuarto, analizar los datos por cada modelo y elegir el mejor.

Ahora bien, la Universidad cumple un rol dentro de la sociedad relacionado con la formación de jóvenes profesionales desde lo teórico, técnico, competencias y habilidades de una ciencia específica. En Latinoamérica, muchas universidades parecen convencidas de la competitividad de sus egresados en los campos académico y científico, sin tener investigaciones sólidas que brinden mecanismos para hacer seguimiento preciso a sus egresados.

En este sentido, y en el ejercicio de su misión, la Universidad de Caldas busca establecer un enfoque por medio del cual la información de sus egresados del último lustro pueda ser aprovechada a través de técnicas de minería de datos, generando un mayor conocimiento en pro de la competencia y productividad. Este artículo presenta un primer acercamiento de los factores resultantes con técnicas de

clasificación de minería de datos que determina la predicción de la utilidad de los conocimientos de los graduandos de dicha universidad.

La percepción de la utilidad de los conocimientos adquiridos es un claro indicador de la culminación exitosa en el plan de estudios de algún alumno, e inclusive puede contribuir en la imagen que un determinado programa y universidad tenga en la sociedad, al ser los egresados representantes directos del prestigio de calidad de una institución. Adicionalmente, puede contribuir a determinar si un egresado puede desempeñarse satisfactoriamente en un cargo que se relacione con el plan de estudios cursado en su carrera profesional.

Se presentan, además, algunos resultados de investigación relacionados con el área pertinente, la descripción de los algoritmos utilizados, la descripción de la base de datos y demás elementos metodológicos.

**Trabajos relacionados**

Rastrear el desenvolvimiento laboral y profesional de los egresados de la educación superior es una tendencia creciente en países que buscan mejorar la calidad y la pertinencia de los programas.



Existen una gran variedad de trabajos realizados en esta área, entre los cuales cabe destacar (MEN, 2007) los que se presentaron en el Seminario Internacional Pertinencia de la Educación: La Educación para la Competitividad. En ese espacio se socializaron las experiencias de sistemas de información de egresados en Italia, Alemania, Australia y Francia, casos que presentaron metodologías similares con puntos de encuentro: una consulta cuando los estudiantes se gradúan, otra consulta tres años después, y la última transcurridos cinco años. Esto permite comparar la información en temas como capacidades laborales, ingresos económicos de los egresados y estudios de posgrado.

En este mismo seminario se destaca el proyecto CEREQ, el cual creó 18 centros regionales universitarios que representan el 10% en instituciones de formación para el empleo en Francia. Uno de sus objetivos es construir y distribuir guías metodológicas a las instituciones de educación superior para crear encuestas y observatorios de egresados, en aras de contribuir al proceso de introducción de los egresados al entorno laboral y ser centro de consulta de jóvenes que van a ingresar a la educación superior.

En el marco del proyecto de cooperación universitaria de Alemania (Schomburg, 2011), por su parte, se realizó una encuesta unificada a egresados de 40 instituciones de ese país. La información que arrojó le sirve al sector académico para estudiar el diseño de nuevos programas, de acuerdo con la “empleabilidad” de los egresados, pero a su vez les sirve a otros grupos de interés como los estudiantes, los padres de familia y los docentes para tomar decisiones responsables.

En Suramérica, Porcel, Dapozo y López (2008) construyeron modelos predictivos del rendimiento académico de los alumnos de las carreras de la Facultad de Ciencias Exactas y Naturales y Agrimensura. Igualmente, Porcel, López, Dapozo y Caputo (2009) desarrollaron una relación entre el número de exámenes rendidos y el número de asignaturas aprobadas como indicador del rendimiento académico de alumnos universitarios.

Alcover, Benlloch, Blesa, Calduch, Celma, Ferri, y Hernández (2009) hicieron un análisis del rendimiento académico en los estudios de Informática de la Universidad Politécnica de Valencia aplicando técnicas de minería de datos. A su vez, Gallestey, De La Ristra, Gil y Guerra (2013), plantearon árboles de regresión y otras opciones metodológicas aplicadas a la predicción del rendimiento académico.

Dapozo, Porcel, López y Bogado (2007) utilizaron técnicas de preprocesamiento para mejorar la calidad de los datos en un estudio de caracterización de ingresantes universitarios; García, Alvarado y Jiménez (2007) aplicaron la predicción del rendimiento académico desde una comparación entre la regresión lineal y la regresión logística; y Thuraisingham (2000), Thomas (2011) y Tsai, (2013) realizaron aplicación de la técnica de minería de datos desde enfoques diversos, que van desde la gestión del conocimiento hasta el tema del bajo rendimiento estudiantil asociado.



## Descripción de la base de datos

La base de datos consta de 18 campos y 15.494 registros. En la tabla 1 se muestra la descripción de los atributos ver tabla 1.

Tabla 1. Estructura de la base de datos (elaboración propia)

| Atributo | Descripción |
|---|---|
| Facultad | Facultad a la que pertenece un egresado. |
| Carrera | Nombre de la carrera del egresado. |
| Estadocivi | Estado civil del egresado —soltero, casado o en unión libre—. |
| Genero | Especifica si el egresado es hombre o mujer. |
| Nhijos | Número de hijos del egresado. |
| Nivelpadre | Especifica el nivel educativo del padre del egresado, el cual puede ser: nunca estudió, primaria incompleta, primaria completa, secundaria incompleta, secundaria completa, educación técnica, educación universitaria incompleta, educación universitaria completa o educación de posgrado. |
| Nivelmadre | Igual a “Nivelpadre”. |
| Edadtbachi | Determina la edad del egresado cuando se graduó del colegio. |

| | |
|---|---|
| Edadpregra | Determina la edad con la que se graduó el egresado de su pregrado. |
| Tiempoemp | Especifica el tiempo que el egresado uso para para conseguir su primer empleo: menos de tres meses, entre tres y seis meses, más de seis, hasta 1 año, o más de 1 año. |
| Rasgos | Especifica la raza originaria del egresado: mestizo, indígena o afrocolombiano. |
| Deptoemp | Departamento donde trabaja el egresado. |
| Tipoempleo | Establece el tipo de empleo del egresado: empleado de empresa particular, empleado del gobierno, cuenta propia, patrón o empleador, u oficios del hogar. |
| Utileshabi | Determina si las habilidades adquiridas en el pregrado son valederas en su actividad profesional —muy útiles, útiles, poco útiles o nada útiles—. |
| Primeremp | Determina un sí o un no para el primer trabajo. |
| Tipocontra | Determina el contrato que actualmente tiene el egresado con su trabajo: contrato a término fijo, contrato a término indefinido, contrato de prestación de servicios u otro tipo de contrato. |
| Canalemple | Determina el canal que uso para la consecución del empleo: medios de comunicación, bolsa de empleo de la institución donde estudió, otras bolsas de empleo, redes sociales o servicio público de empleo. |
| Relacionad | Determina si el empleo actual del egresado está directamente relacionado con su profesión: directamente relacionado, indirectamente relacionado o nada relacionado. |



## Método

Con el propósito de extraer conocimiento a partir de los datos almacenados en el repositorio mencionado, y teniendo como factor fundamental la predicción de información frente a la utilidad de las habilidades del egresado de la Universidad de Caldas, el trabajo se dividió en las etapas más comunes de la minería de datos, apuntadas por Riquelme, Ruiz y Gilbert (2006) y Liao, Chu y Hsiao (2012).

### Selección

Se obtuvieron los repositorios de datos internos y externos que sirvieron de base para el proceso de minería de matos. Como fuente interna se seleccionó la base de datos histórica de los egresados de la Universidad de Caldas en los últimos diez años; y como fuente externa se seleccionó la encuesta de seguimiento a graduados, que pretende analizar el desarrollo profesional y personal de los graduados de educación superior. Con estos insumos se unificó la base de datos de los egresados con el modelo de encuestas, conformando una sola tabla para 15.494 estudiantes construida con un sistema gestor de base de datos.

**Procesamiento y limpieza**
El objetivo de esta etapa es obtener datos limpios. Así entonces, se eliminaron a través de programas *script* los valores indeterminados o nulos y se estandarizaron valores que eran incoherentes y afectaban el patrón de calidad de la base de datos. Por medio de consultas SQL se analizó y selecciono cuidadosamente la calidad de los datos contenidos en cada uno de los atributos de las tablas.

**Transformación de datos**
En esta etapa se discretizaron algunos de los atributos con valores continuos, como el tiempo de demora en conseguir empleo por parte del egresado y la ciudad o municipio donde labora, por el departamento asociado; es decir, los valores numéricos se transformaron en discretos o nominales con el fin de disminuir el número de valores distintos de estos atributos.

**Minería de datos**
Proceso de descubrimiento de patrones en la utilidad de las habilidades del egresado. Para tal efecto se utilizaron las siguientes tareas de clasificación, los cuales se discutirán seguidamente: Clasificador OneR, algoritmo J48, Bayes.NaiveBayes y Stacking.



## Resultados y discusión

En la minería de datos los procedimientos de clasificación desarrollan un modelo agregado por reglas (si-entonces) y se aplican cabalmente. En concordancia, el efecto de aplicar el algoritmo de clasificación se direcciona a comparar la clase predicha con la clase real de las instancias.

Este proceso de minería de datos busca reglas para definir si un ítem o un evento pertenecen a una clase de datos en particular. Para el conjunto de datos se cuenta con un conjunto apropiado de atributos predictivos, de tal manera que el modelo busca identificar los egresados con mayor propensión a justificar una percepción “muy útil”, “poco útil” o “nada útil” de las destrezas adquiridas en algún programa cursado.

Para el desarrollo de la investigación se utilizaron los algoritmos que se presentan a continuación.

**Clasificador OneR**
Este algoritmo tiene como particularidad la selección del atributo que mejor revela la clase de salida. Las características propias de este método de clasificación se

resumen en su rapidez y buenos resultados, en contraste con otros algoritmos más complejos.

Para este caso se aplicó a la predicción de graduandos con la variable “utilidad” (Utileshabi) en la base de datos, como se observa en la figura 1.

```
=== Classifier model (full training set) ===

nivelpadre:
        Educación de postgrado  -> Muy útiles
        Educación técnica       -> Muy útiles
        Educación universitaria completa        -> Muy útiles
        Educación universitaria incompleta      -> Muy útiles
        Nunca estudió   -> Útiles
        Primaria completa       -> Nada útiles
        Secundaria completa     -> Muy útiles
        Secundaria incompleta   -> Muy útiles
        Educación tecnológica   -> Muy útiles
(13144/15494 instances correct)


Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       13144               84.8328 %
Incorrectly Classified Instances      2350               15.1672 %
Kappa statistic                          0.6633
Mean absolute error                      0.0758
Root mean squared error                  0.2754
Relative absolute error                 30.0067 %
Root relative squared error             77.4737 %
Coverage of cases (0.95 level)          84.8328 %
Mean rel. region size (0.95 level)      25      %
Total Number of Instances            15494
```

Figura 1. Clasificador OneR aplicado (elaboración propia)

Los resultados del algoritmo OneR muestran que la mejor predicción posible con un solo atributo es el nivel de educación del padre (nivelpadre en la base de datos). Su umbral está fijado en la aparición del valor categórico “Muy útiles” cuando su valor de clasificación oscila, según la tabla 2.

Tabla 2. Clasificación nivelpadre-utilehabi (elaboración propia)

| Nivelpadre | Utileshabi |
|---|---|
| Educación de postgrado | Muy útiles |
| Educación Técnica | Muy útiles |
| Educación Universitaria completa | Muy útiles |
| Secundaria completa | Muy útiles |

| | |
|---|---|
| Secundaria incompleta | Muy útiles |
| Educación tecnológica | Muy útiles |

La tasa de aciertos de 84,83% es representativa sobre el propio conjunto de entrenamiento. Si se analiza la matriz de confusión (figura 2), se puede observar que los valores de la diagonal son los aciertos, y el resto los errores. De los 13.144 alumnos clasificados en la utilidad de sus habilidades, 12.091 son correctamente clasificados como 'Muy útiles'.

```
=== Confusion Matrix ===

     a     b     c     d   <-- classified as
 10037    30     0    46 |     a = Muy útiles
  1598  2054     0   211 |     b = Útiles
   308     0     0     9 |     c = Poco útiles
   148     0     0  1053 |     d = Nada útiles
```

Figura 2. Matriz de confusión (elaboración propia)



**Algoritmo J48**

Implementación libre en Java del algoritmo C4.5. Utiliza el concepto de entropía de la información para la escogencia de variables que mejor clasifiquen a la variable estudiada. Este algoritmo genera un árbol de decisión estadístico. El atributo con la mayor ganancia de información normalizada se elige como parámetro de decisión. Cuando todas las muestras en la lista pertenecen a la misma clase, se crea un nodo de hoja para el árbol de decisión.

El parámetro más importante que se debe tener en cuenta es el factor de confianza para la poda, que influye en el tamaño y capacidad de predicción del árbol construido. Cuando la probabilidad es menor, se exige que la diferencia en los errores de predicción antes y después de podar sea más significativa para no simplificar (figura 3).

```
=== Summary ===

Correctly Classified Instances       14262               92.0485 %
Incorrectly Classified Instances      1232                7.9515 %
Kappa statistic                          0.8342
Mean absolute error                      0.0704
Root mean squared error                  0.188
Relative absolute error                 27.8695 %
Root relative squared error             52.88   %
Coverage of cases (0.95 level)          98.7995 %
Mean rel. region size (0.95 level)      43.7056 %
Total Number of Instances            15494
```

Figura 3. Factor de confianza (elaboración propia)

Se muestra una clasificación representativa del 92,04%, contra un 7,95% de instancias no clasificadas. Así entonces se genera un árbol de decisión, como se observa en la figura 4.



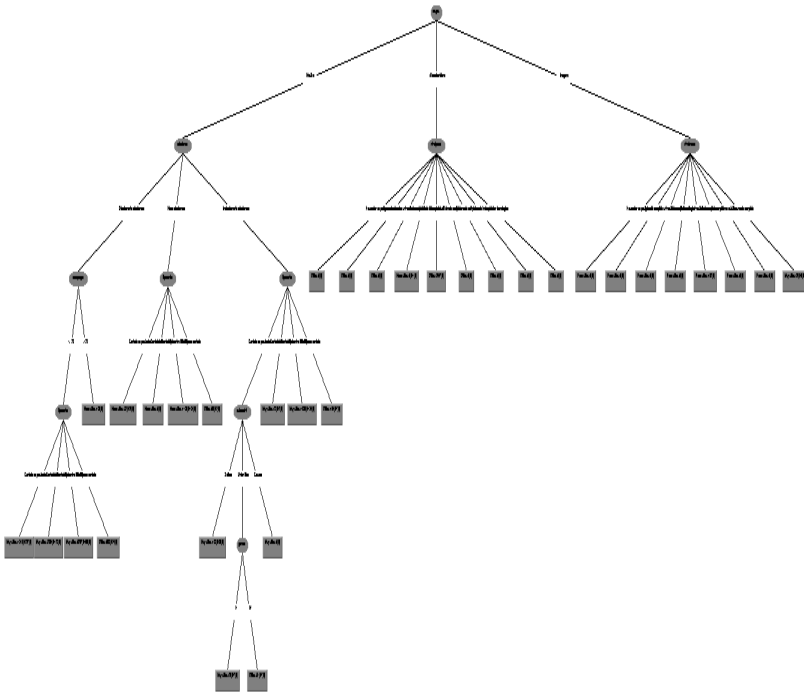

Figura 4. Árbol de decisión (elaboración propia)

**Bayes.NaiveBayes**

Este es un algoritmo muy usado en procesos de clasificación; por su efectividad en el aprendizaje inductivo se lo considera como uno de los más eficientes dentro de la minería de datos. Este algoritmo trabaja basado en la hipótesis de que todos los atributos son independientes entre sí, claro está, si el valor de la variable clase es conocido.

En el caso de la base de datos de egresados, se utilizó este algoritmo para representar la distribución de una mezcla de componentes cuyas variables se asumen independientes. Por tanto, se genera un único nodo raíz que apunta a la clase y en el que todos los atributos son nodos hoja. Su efectividad dependerá de la independencia de sus atributos (ver figura 5).

```
=== Summary ===

Correctly Classified Instances       12702               81.9801 %
Incorrectly Classified Instances      2792               18.0199 %
Kappa statistic                          0.6556
Mean absolute error                      0.0916
Root mean squared error                  0.2861
Relative absolute error                 36.2304 %
Root relative squared error             80.4946 %
Coverage of cases (0.95 level)          89.2281 %
Mean rel. region size (0.95 level)      29.0419 %
Total Number of Instances            15494
```



Figura 5. Mezcla de componentes (elaboración propia)

Se tiene una clasificación representativa del 81,98% contra un 18,01% de instancias no clasificadas, por lo que se genera la matriz de confusión de la figura 6.

```
=== Confusion Matrix ===

    a    b    c    d   <-- classified as
 8886   34 1147   46 |    a = Muy útiles
  993 2568   91  211 |    b = Útiles
   87   26  195    9 |    c = Poco útiles
  143    5    0 1053 |    d = Nada útiles
```

Figura 6. Matriz de confusión de mezcla de componentes (elaboración propia)

Los valores de la diagonal establecen los aciertos, siendo la clasificación “Muy útiles” la mejor instanciada.

### *Stacking*

Este algoritmo, considerado un metaclasificador, combina varios modelos y construye un nuevo conjunto con los generados. Trabaja utilizando el resultado mayoritario, siempre y cuando la precisión de los clasificadores sea representativa. Se parte de los datos y se crean varios clasificadores heterogéneos, generando salidas que se usan como atributos de un nuevo clasificador, como se observa en la figura 7.

```
=== Summary ===

Correctly Classified Instances       10113               65.2704 %
Incorrectly Classified Instances      5381               34.7296 %
Kappa statistic                          0
Mean absolute error                      0.2527
Root mean squared error                  0.3555
Relative absolute error                100      %
Root relative squared error            100      %
Coverage of cases (0.95 level)          97.954  %
Mean rel. region size (0.95 level)      75      %
Total Number of Instances           15494
```

Figura 7. Resultados *stacking* (elaboración propia)

Se tiene una clasificación aceptable del 65,27%, contra un 34,72% de instancias no clasificadas. Se aplica entonces la matriz de confusión, como se observa en la figura 8.



```
=== Confusion Matrix ===

     a     b     c     d   <-- classified as
 10113     0     0     0 |     a = Muy útiles
  3863     0     0     0 |     b = Útiles
   317     0     0     0 |     c = Poco útiles
  1201     0     0     0 |     d = Nada útiles
```

Figura 8. Matriz de confusión *stacking* (elaboración propia)

Se aprecia que no se logran clasificar ciertas variables, lo que lleva a escoger un algoritmo de mayor criterio en la clasificación. Con la obtención de resultados en los procesos de selección se consolida la información que se presenta en la tabla 3.

**Tabla 3.** Resultados clasificaciones (elaboración propia)

| Algoritmo | Fiabilidad | Estimación | Kappa |
|---|---|---|---|
| OneR | 84,83% | 131.144 | 0,66 |
| J48 | 92,04% | 14.262 | 0,83 |
| Bayes | 81,98% | 12.702 | 0,65 |
| Stacking | 65,27% | 10.113 | 0 |

Los resultados muestran que el mayor porcentaje de fiabilidad se logra con el algoritmo J48 con un valor de 92,04%, corroborado con un valor de Kappa de 0,83, y da una estimación de 14262 registros de egresados. En contraste, y a pesar de su condición de metaclasificador, *Stacking* muestra la menor fiabilidad: 62,27% con una estimación de 10.113 registros de egresados.

Entre las reglas de clasificación más representativas según la interpretación del mejor modelo están:

- Si el rasgo del egresado es “Mestizo”, y la relación de su empleo frente a su carrera de estudio está “Directamente relacionado” y la edad es menor de “28” años, con tipo de contrato igual “Contrato a término indefinido” o “Contrato a término fijo” o “Contrato de prestación de servicios”, su percepción de las habilidades y destrezas adquiridas durante sus estudios en la universidad es de tipo A, como se observa en la tabla 3.



**Tabla 3.** Regla de clasificación (elaboración propia)

| Atributo utileshabi | Tipo |
|---|---|
| Muy útiles | A |
| Útiles | B |
| Poco útiles | C |
| Nada útiles | D |

- La regla de clasificación encuentra 8967 egresados que cumplen la condición; por tanto, el 57,87% de egresados están en este rango.
- Si el rasgo es “Mestizo” mayor de “28” años, el 0,87% de los egresados cree que sus habilidades no son útiles.

- Si la relación de su empleo frente a su carrera de estudio está “Nada relacionado” y la edad supera los “28” años con rasgo de raza “Mestizo”, el 8,4% considera que las habilidades adquiridas son “Nada útiles”.

- Si el rasgo del egresado es “Afrocolombiano” y el nivel de educación de su padre es “Nunca estudió” el 13,11% de estos estudiantes cree que las habilidades adquiridas son “Útiles”.

## Conclusiones

Aplicar técnicas de minería de datos requiere de una precisa transformación de los mismos; en este caso, esa fue la etapa más costosa en tiempo, conforme a la irregularidad en la calidad de dichos datos, ya que los registros presentaban incoherencias y una gran cantidad de los datos de cierto atributo no eran persistentes. Igualmente, se encontraron muchos datos nulos o faltantes y otros redundantes.

Algunas variables categóricas, como el estrato del egresado, no representaron un papel importante en los algoritmos de clasificación. Adicionalmente, el tipo de contrato y la relación entre la carrera y el puesto de trabajo son influyentes en la percepción de la calificación de las habilidades y destrezas adquiridas en los estudios.



## Referencias

Alcover, R., Benlloch, J., Blesa, P., Calduch, M., Celma, M., Ferri, C. & Hernández Orallo, J. (2007). “Análisis del rendimiento académico en los estudios de Informática de la Universidad Politécnica de Valencia aplicando técnicas de Minería de Datos”. XII Jornadas de Enseñanza Universitaria de la Informática 2007. Recuperado de http://bioinfo.uib.es/~joemiro/aenui/procjenui/j en2007/alanal.pdf

Bacallao Gallestey C., Parapar De La Ristra, J., Roque Gil M., Bacalloa Guerra J. “Arboles de regresión y otras opciones metodológicas aplicadas a la predicción del rendimiento académico”. Revista Cubana de Educación Médica Superior, vol.18, N°3. 2004. Disponible en: http://bvs.sld.cu/revistas/ems/vol18_3_04/ems 02304.htm

Candás J. (2006). Minería de datos en bibliotecas: bibliominería. Universitat de Barcelona. Recuperado de http://www.ub.edu/bid/17canda2.htm.

Dapozo, G., Porcel, E., López, M. V. & Bogado, V. (2007). Técnicas de preprocesamiento para mejorar la calidad de los datos en un estudio de caracterización de ingresantes

universitarios. IX Workshop de Investigadores en Ciencias de la Computación (WICC 2007). Trelew. Chubut. Argentina.

García Jiménez, M. V., Alvarado Izquierdo, J. M. & Jiménez Blanco, A. (2000). La predicción del rendimiento académico: regresión lineal versus regresión logística. *Psicothema, 12*(2), 248-252. Recuperado de http://redalyc.uaemex.mx/redalyc/pdf/727/72797059.pdf

Liao, S.-H., Chu, P.-H., & Hsiao, P.-Y. (2012). Data mining techniques and applications – A decade review from 2000 to 2011. *Expert Systems with Applications, 39*, 11303–11311.

Ministerio de Educación Nacional (2007). *Seminario Internacional sobre Pertinencia de la Educación: La Educación para la Competitividad.* Recuperado el 1 de septiembre de 2015, de http://www.graduadoscolombia.edu.co/html/1732/article-170872.html

Natek, S., & Zwilling, M. (2014). Student data mining solution–knowledge management system related to higher education institutions. *Expert Systems with Applications, 41*, 6400–6407.

Pérez-Palacios, T., Caballero, D., Caro, A., Rodríguez, P., & Antequera, T. (2014). Applying data mining and Computer Vision Techniques to MRI to estimate quality traits in Iberian hams. *Journal of Food Engineering, 131*, 82–88.



Porcel, E. A., Dapozo, G. N. & López, M. (2008). *Técnicas clásicas de minería de datos aplicadas al estudio del rendimiento académico de alumnos de primer año de carreras de la FACENA.* Corrientes (Argentina): Universidad Nacional del Nordeste.

Porcel, E., López, M. V., Dapozo, G. & Caputo, L. (2009). Relación entre el número de exámenes rendidos y el número de asignaturas aprobadas como indicador del rendimiento académico de alumnos universitarios. XXII Encuentro Nacional de Docentes de Investigación Operativa (ENDIO). XX Escuela de Perfeccionamiento en Investigación Operativa (EPIO).

Riquelme, J., Ruiz, R., & Gilbert, K. (2006). Minería de Datos: Conceptos y Tendencias. *Revista Iberoamericana de Inteligencia Artificial*, 11-18.

Schomburg H. (2011). Contribution to the conference “Lavore dei laureati: I Numeri, Le competenze”, 9 de junio. Genoa, Italia.

Thomas, S. (2011). Association rule module for data mining. Patente de Estados Unidos US 7,962,483 B1.

Tsai, H. (2013). Knowledge management vs. data mining: Research trend, forecast and citation approach. *Expert Systems with Applications*, *40*, 3160-3173.

Thuraisingham, B. (2000). A primer for understanding and applying Datamining. *IT Professional, 2* (1), 28-31.