Artículo de investigación

Aplicación de una metodología adaptada de minería de datos en información del sector público*

Application adapted methodology of data mining in public sector information

Johnny Alexander Salazar Cardona, Marcelo López Trujillo*****

**Master en Ingeniería Computacional. Docente de carrera, Escuela de Administración y Mercadotecnia del Quindío. Director grupo de investigación Inge-Soft. Facultad de Ingenierías, Avenida Bolívar N 3-11, Quindío – Armenia, Colombia.

***PhD Ingeniería Informática. Profesor Asociado, Universidad de Caldas, Manizales, Colombia. Grupo de investigación GITIR. Facultad de Ingenierías, Caldas – Manizales, Colombia.

Resumen

En este trabajo se muestran los resultados obtenidos en la aplicación de procesos de descubrimiento de conocimiento y minería de datos en información del sector público, enfocado a la descripción del estado monetario y laboral de los ciudadanos en Colombia, aplicando la metodología establecida obtenida en resultados previos del proyecto de investigación magistral de los autores de este artículo. Dicha metodología está establecida en el campo de los datos abiertos que son liberados en un estado gubernamental de gobierno abierto, con el fin de descubrir el conocimiento oculto en un conjunto de datos que serán publicados y compartidos a la ciudadanía en general, aplicando algoritmos predictivos y descriptivos de minería de datos. Con la aplicación de dicha metodología se obtuvieron tendencias en los datos a partir del departamento, municipio, nivel académico de la persona y su género, definiendo reglas puntuales para cada uno de los atributos relacionado con atributos de filtro, además se constató que la metodología establecida es una guía clave para el éxito del proyecto, ya que enmarca todo el entorno necesario para el tratamiento de los datos y el descubrimiento de conocimiento inmerso en estos, pero según sus características puede generar cuellos de botella cuando el volumen de datasets a analizar es muy alto.

Palabras clave: Algoritmos descriptivos, algoritmos predictivos, datos abiertos, gobierno abierto, minería de datos.

Marco de referencia para la implementación del mapa de ruta establecido en los lineamientos nacionales de apertura de datos del sector público y su integración con procesos de descubrimiento de conocimiento e inteligencia de negocios. Grupo de investigación GITIR, Universidad de Caldas

Recibido: 10/02/2016
Revisado: 23/03/2016
Aceptado: 01/12/2016

Correspondencia de autor:

alexander9052@gmail.com
mlopez@ucaldas.edu.co

Cómo citar:

Salazar, J.A., López, M. Aplicación de una metodologia adaptada de minería de datos en información del sector público. *UGCiencia* 22, 199-212.

Abstract

This paper describes the results obtained in the application of process knowledge discovery and data mining in public sector information, focused on the description of the monetary and employment status of citizens in Colombia, using the methodology established in a previous research project masterly thesis of the authors of this article. This methodology is set to the topic of open data that are released in a state of open government, with the aim of discover the hidden knowledge in a set of data that will be published and shared with citizens in general, applying predictive and descriptive algorithms of data mining. With the application of this methodology we obtained trends in the data from the state, city, academic level of the person and its gender, defining specific rules for each attributes related to filter attributes, in addition it was found that the established methodology is a key guide to the success of the project, it is covers all the necessary environment for the treatment of the data and the discovery of knowledge immersed in them, but according to its characteristics can generate bottlenecks when the volume of datasets to analyze is very high.

Keywords: Data mining, descriptive algorithms, open government, open data, predictive algorithms.

Introducción

Cuando un gobierno tiene sistematizados todos los elementos transaccionales que involucran a la ciudadanía en general (e-gov o gobierno electrónico), debe enfocarse en alcanzar un nivel de madurez de gobierno abierto u o-gov , el cual se centra en la publicación o liberación de los datos públicos para que la ciudadanía tenga acceso a estos, los utilice y les realicen auditoria según la naturaleza de la información (Gómez & Gascó, 2012; Sourouni, Kourlimpinis, Mouzakitis, & Askounis, 2010). Cuando una entidad del sector gubernamental desea liberar un conjunto de datos a la ciudadanía para que esta pueda ser descargada para su uso, debe pasar por un proceso de transformación para poder aplicarles procesos de descubrimiento de conocimiento y que se adapten a los estándares del nivel de gobierno abierto en el cual se encuentre un determinado dataset. Para esto, se definió una metodología adaptada de minería de datos (Salazar, 2015) basada en KDD (Knowledge Discovery in Data Base) (Usama Fayyad, Gregory Piatesky-Shapiro, & Padharaic Smyth, 1996a, 1996b) y fue validada sobre un conjunto de datos que contienen la información sobre la medición del estado monetario con base en su situación laboral en Colombia.

En la validación y aplicación de la metodología establecida, se esperaba encontrar relaciones entre el ingreso total, la edad, el nivel académico, el estrato, factor de expansión anual, entre otros, según el departamento, municipio, posición laboral, y género. Para lograr este objetivo se definieron una serie de atributos sobre el dataset de trabajo, clasificados en 3 categorías. La primera de ellas, son los atributos de segmento, que permitió subdividir el gran conjunto de datos que se tenía en unos más pequeños facilitando el trabajo. Con esto se tendrían resultados independientes y así se evitaría el ruido en los resultados. Luego, se definieron los indicadores claves de rendimiento (Rice, Abshire, Christakis, & Sherman, 2010) que permitirían visualizar el estado de estos indicadores comparados entre sí, especificando las dependencias y relaciones que se quieren encontrar. Además estos indicadores permitirían facilitar el proceso de entendimiento por parte de un tercero, si se desea mostrar los resultados en un Dashboard (Ogan & Oana, 2012), indicando claramente si algún aspecto de interés se encuentra en buen o mal estado. Finalmente se especificaron los indicadores de filtro que permitirían filtrar los resultados según la sección de datos seleccionados, adicionalmente si se desean mostrar los datos en un Dashboard permitirá filtrar los resultados en tiempo real, facilitando aún más la comprensión de estos (Salazar, 2015).

Materiales y métodos

Este artículo tiene un enfoque de investigación descriptiva, en el cual se detalla la aplicación de una metodología de minería de datos para el sector gubernamental previamente establecida a través de cada una de sus etapas, conllevando al análisis predictivo y descriptivo para el descubrimiento de conocimiento o validación de este según corresponda. El conjunto de datos utilizados para el marco de esta investigación, describen el estado monetario y laboral de los ciudadanos en Colombia, pero como los datos fueron otorgados por una entidad pública de la ciudad de Manizales - Colombia, se priorizarán ciertos resultados en este sector.

Preparación inicial de los datos

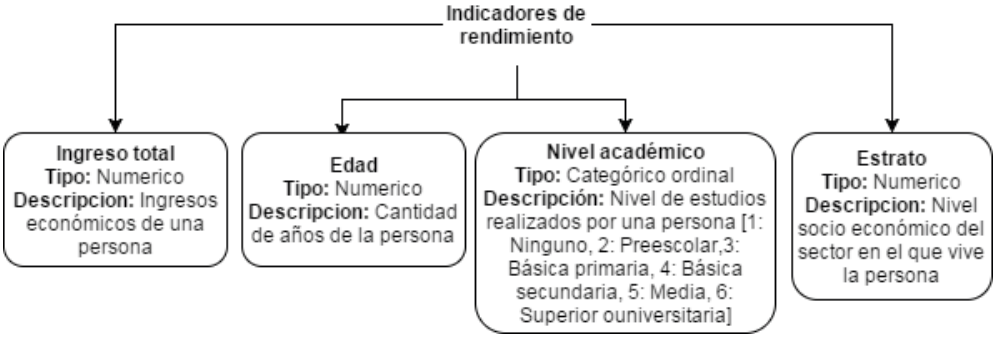
El conjunto de datos de trabajo era de alta dimensionalidad con un total de 135 atributos y 822.087 registros,

encontrándose en formato SAV (Saved File), este era un archivo propio del software estadístico SPSS (*Statistical Package for the Social Sciences*) para el análisis de datos, pero este tipo de extensión no se encontraba entre los diferentes formatos establecidos en la metodología guía, por lo que fue transformado a formato CSV (*comma-separated values*). Se decidió tener el archivo en este formato, debido a que es ligero y puede ser visualizado fácilmente en un software ofimático sin importar la gran dimensionalidad que tenga el dataset, siendo fácilmente manipulable por parte de un tercero que esté interesado en ver los datos en bruto, y adicionalmente, es un formato soportado por la herramienta elegida de minería de datos para el proceso de descubrimiento de conocimiento. El formato XLS (eXceL Spreadsheet) aunque también es visible fácilmente por una herramienta ofimática, no es un formato ligero que soporte ágilmente grandes volúmenes de información por lo que fue descartado.

Comprensión del dominio del problema

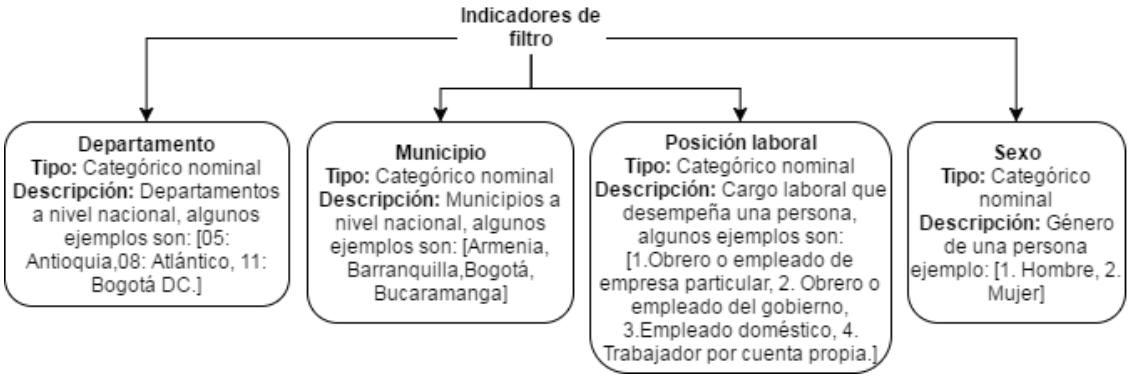
De los 135 atributos posibles se determinó que 27 podrían ser relevantes en el proceso. Se debe entender que el conjunto de datos de trabajo contiene información sobre la medición de pobreza monetaria con base a su situación laboral en el año 2010, de personas que se encuentran trabajando, personas desempleadas, estudiantes, pensionados, etc. En los diferentes campos del dataset se encuentra un atributo que indica el estado laboral actual de la persona (especificando lo descrito), por lo que fue seleccionado como atributo de segmento, ya que permitió dividir la información en personas que se encuentran trabajando, desempleados, estudiantes, entre otros, permitiendo tener información relevante por cada una de las categorías existentes. Los atributos que fueron seleccionados como indicadores de filtro y de rendimiento buscaban permitir centrar esfuerzos para el cumplimiento del objetivo establecido que se desea alcanzar con el proceso de descubrimiento de conocimiento (fig. 1 y 2)

Figura. 1 Indicadores de rendimiento



Fuente: elaboración propia

Figura. 2 Indicadores de filtro



Fuente: elaboración propia

Creación de la base de datos de trabajo o selección de datos

El formato al cual se transformó el conjunto de datos que se aplicaría minería de datos fue ARFF (Attribute-Relation File Format), que permite ahorrar tiempo a la herramienta de minería de datos que sería utilizada debido a que no tendría la necesidad de determinar el tipo de datos y sus posibles valores ahorrando gran recurso computacional (fig. 3) (Carreño, 2008).

Figura. 3 Atributos seleccionados para el proceso

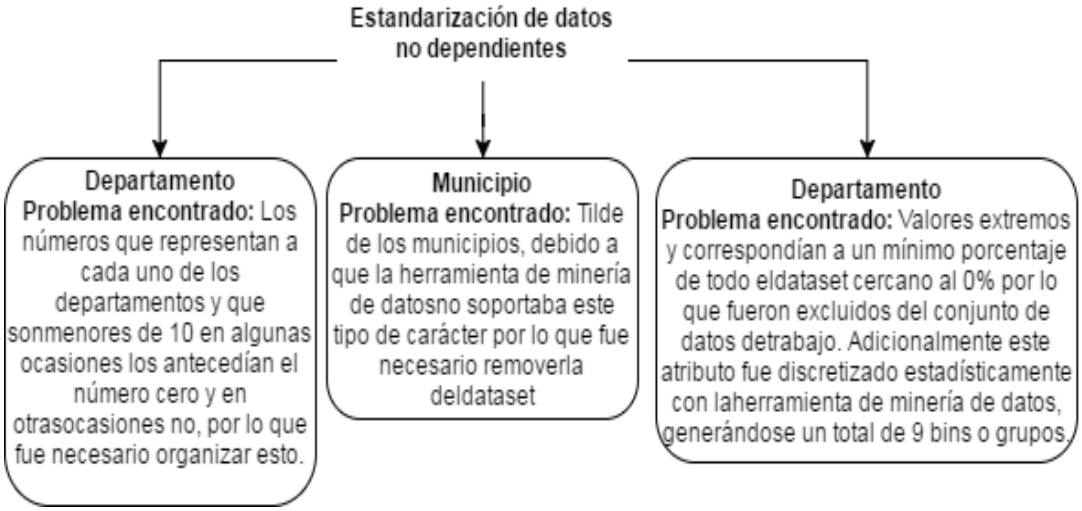
```
@attribute DPTO {5,8,11,13,15,17,18,19,20,23,25,27,41,44,47,50,52,54,63,66,68,70,73,76}
@attribute Municipio(DOMINIO) {ARMENIA,BARRANQUILLA,BOGOTA,BUCARAMANGA,CALI,CARTAGENA,CUCUTA,FLORENCIA,IBAGUE,MANIZALES,
@attribute CAPITAL {0,1}
@attribute ESTRATO {1,2,3,4,5,6}
@attribute Sexo(P6020) {1,2}
@attribute Anos(P6040) numeric
@attribute Entidad_salud(P6090) {1,2,9}
@attribute Nivel_educativo(P6210) {1,2,3,4,5,6,9}
@attribute Posicion_laboral(P6430) {1,2,3,4,5,6,7,8,9}
@attribute horas_x_semana(P6800) numeric
@attribute No_personas_empresa(P6870) {1,2,3,4,5,6,7,8,9}
@attribute Pension(P6920) {1,2,3}
@attribute otro_trabajo(P7040) {1,2}
@attribute horas_otro_trabajo(P7045) numeric
@attribute posicion_otro_trabajo(P7050) {1,2,3,4,5,6,7,8,9}
@attribute desea_mas_horas(P7090) {1,2}
@attribute tiene_disponible_mas_horas(P7120) {1,2}
@attribute otros_ingresos_por_propiedades(P7500S1) {1,2,9}
@attribute otros_ingresos_pension_subsidio(P7500S2) {1,2,9}
@attribute otros_ingresos_instituciones(P7500S3) {1,2,9}
@attribute ingresos_cesantias(P7510S6) {1,2,9}
@attribute Ingresos_trabajo(P6500) numeric
@attribute Ingreso_otro_trabajo(P7070) numeric
@attribute ingreso_segunda_actividad(ISAES) numeric
@attribute ingreso_total(INGTOT) numeric
@attribute ingreso_total(INGTOT_NOMINAL)(P7500S3) {Menor_de_300000,Entre_300_y_620,Entre_620_y_900,Entre_900_y1500,Entre
@attribute Factor_expansión _anual(FEX_C) numeric
@attribute FEX_C_NOMINAL2 {Entre_0_y_50,Entre_50_y_100,Entre_100_y_150,Entre_150_y_200,Mayor_de_200}

@data
15,TUNJA,1,4,1,70,1,3,?,?,?,?,?,?,?,?,?,?,1,2,2,?,?,?,?,150000,Menor_de_300000,6.975405,Entre_0_y_50
18,FLORENCIA,1,5,1,48,2,4,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,290000,Menor_de_300000,4.884113,Entre_0_y_50
```

Fuente: elaboración propia

Limpieza y pre-procesamiento de los datos

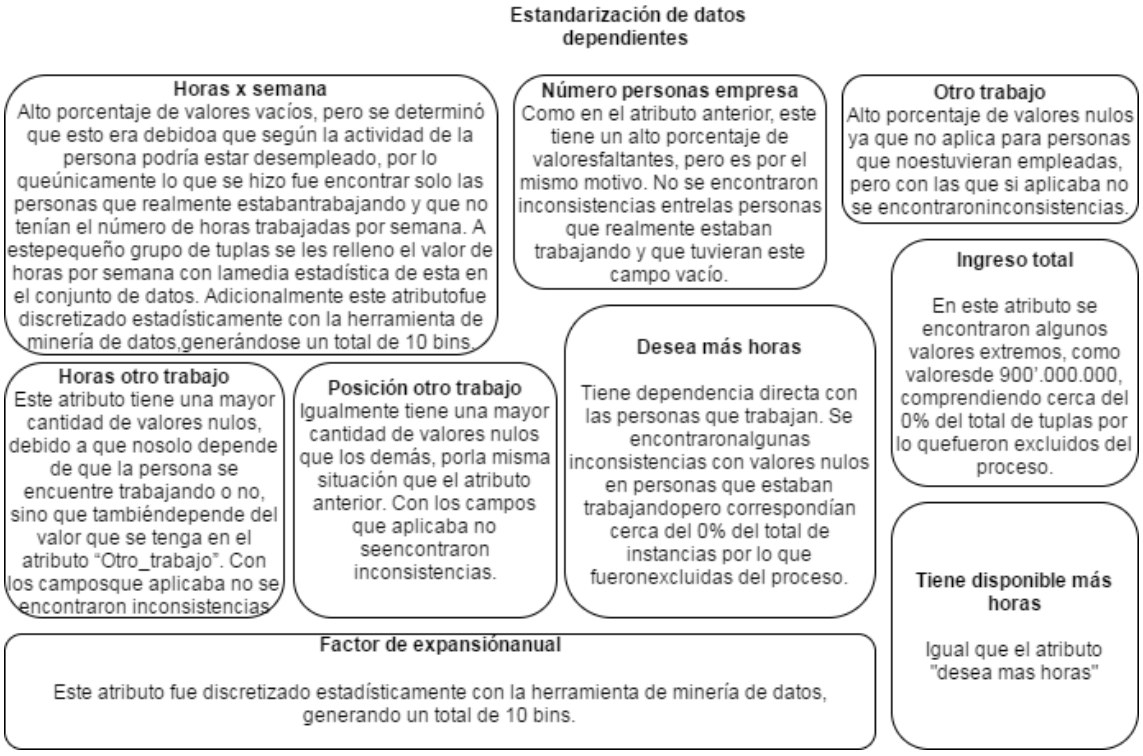
Se organizaron algunos elementos del conjunto de datos que generarían inconsistencia en los resultados de descubrimiento de conocimiento en cada uno de los atributos. Se describen a continuación los atributos que fueron procesados, entre los que se encontraban atributos no dependientes de otros (fig. 4) y atributos que era dependientes según su actividad, lo que conllevó a tener muchos valores vacíos para su tratamiento (fig. 5) (Hernandez & Rodrigez, 2008).

Figura. 4 Estandarización de datos no dependientes



Fuente: elaboración propia

Figura. 5 Estandarización de datos dependientes



Fuente: elaboración propia

Transformación de los datos

Hasta este punto se tenía un conjunto de datos limpio para trabajar, con los diferentes atributos que se creen relevantes, sin datos extremos, nulos o no estandarizados que puedan generar ruido en los resultados de minería de datos. El dataset contaba con un total de 822.087 registros y 28 atributos, lo cual indicaba que seguía teniendo una alta dimensionalidad para un proceso ágil de minería, con datos que realmente muestren información segmentada a los usuarios según sus necesidades.

Reducción de instancias

Con los posibles valores del atributo de segmento se subdividió el dataset principal en 6 subdatasets, cada uno con los registros que tengan el valor especificado para el atributo de segmento. Con esto se generaron los siguientes conjuntos de datos con sus respectivo número de instancias: Personas que se encuentran trabajando con un total de 446.437, Personas que están buscando trabajo 25.267, estudiantes con un total de 120657, oficios del hogar 152.945, incapacitados permanentes con 8.929 y otro para ninguno de las categorías anteriormente descritas con 47.852 registros. Para establecer de estos subdatasets cuál sería tomado para realizar el proceso de descubrimiento de conocimiento, se determinó que lo más acertado sería encontrar conocimiento sobre las personas que actualmente buscan trabajo, con el fin de describir este sector crítico de la población.

Hasta este momento se logró realizar una reducción a nivel de instancias, disminuyendo en un 97% el total de estas al segmentar la información, adicionalmente con esto se podría obtener una información más centrada en un determinado sector de todo el conjunto de datos de trabajo. Aunque se redujo significativamente el total de todas las instancias aún no se había realizado la reducción de dimensionalidad a nivel de atributos, los cuales eran 28.

Reducción de atributos

Antes de realizar esta reducción se realizó un proceso de nivelación del atributo clase, es decir, para cada posible valor de los indicadores de filtro se buscó que la cantidad de estos en el dataset fuera proporcional, con el fin de que no se tuviera tendencias muy marcadas sobre un determinado valor del atributo. Esta nivelación fue realizada aplicando un proceso de remuestreo balanceado, eligiendo instancias aleatorias para cada posible valor del atributo clase hasta tener el total de instancias iniciales (Gutiérrez, 2012; Orallo, Quintana, & Ramirez, 2005).

Para la reducción a nivel de atributos se trabajó sobre los indicadores de filtro, tomando como base el dataset segmentado solo con los datos de las personas que actualmente están buscando trabajo y se generó un nuevo conjunto de datos de trabajo por cada uno de los diferentes indicadores de filtro.

Sobre este indicador de trabajo se realizó una reducción de dimensionalidad a nivel de atributos, teniendo en cuenta solo los atributos que son relevantes estadísticamente por cada uno de los indicadores clave, mejorando en etapas posteriores la precisión de los resultados y disminuyendo el costo computacional requerido para la obtención de resultados. Esta relación estadística se realizó de manera supervisada por cada uno de los indicadores clave, tomando el indicador analizado como atributo clase, obteniendo así los atributos que son relevantes para predecir el valor de dicho indicador. Para cada uno de los conjuntos de datos que se crearon se realizó la reducción de dimensionalidad a nivel de atributos respecto a cada uno de los indicadores clave, y fueron aplicados diversos algoritmos, detectando así cuáles métodos de búsqueda son los que mejores resultados obtuvieron: Algoritmo evaluador *cfs_sub_set_eval* con el método de búsqueda *exhaustive_search*, evaluador *chi_cuadrado* con el método de búsqueda *ranker* y el evaluador *info_gain* con el método de búsqueda *ranker*. Con los algoritmos descritos se obtuvieron los siguientes resultados en la reducción de dimensionalidad a nivel de atributos:

Dataset con indicador clave departamento: De los 28 atributos disponibles solo 8 tuvieron una relación directa con departamento los cuales son: factor de expansión anual, estrato, capital, nivel educativo, ingreso total, entidad de salud, sexo, años y el deseo de una persona por trabajar más horas a la semana.

Dataset con indicador clave municipio: Solo 7 tuvieron una relación directa con estrato los cuales son: factor de expansión anual, estrato, nivel educativo, ingreso total, entidad de salud, años y sexo

Dataset con indicador clave posición laboral: Solo 12 tuvieron una relación directa con estrato los cuales son: número de personas en la empresa, ingresos trabajo, ingreso total, municipio, sexo, otros ingresos, pensión, subsidio, horas por semana, ingresos otro trabajo, desea más horas, factor expansión anual, nivel educativo y si se encuentra viviendo en una ciudad capital.

Dataset con indicador clave sexo: Solo 10 tuvieron una relación directa con estrato los cuales son: nivel educativo, años, municipio, ingreso total, entidad de salud, estrato, otro trabajo, pensión, posición laboral y expansión anual nominal.

Algoritmos aplicados

Debido a que se definieron los diferentes indicadores clave de rendimiento y los atributos relevantes en función a cada uno de ellos, se decidió hacer énfasis en los modelos predictivos, como los arboles de decisión y generación de reglas supervisadas, para así encontrar relaciones entre los valores de los atributos relevantes sobre los valores del indicador de filtro. Adicionalmente se aplicaron procesos descriptivos como reglas de asociación (Reyes & García, 2005) y *clustering* para constatar los resultados obtenidos en los modelos predictivos (Usama Fayyad, Gregory Piatesky-Shapiro, & Padhraic Smyth, 1996; Maimon & Rokach, 2010).

Se aplicaron diferentes algoritmos sobre los múltiples conjuntos de datos tratados. Finalmente, se dejaron unos pocos que arrojaron los mejores resultados sobre los datasets tratados. Estos algoritmos fueron: el clasificador *Part* y el clúster Dbscan. La aplicación de estos sobre los diferentes conjuntos de datos se realizó aplicando

validación cruzada con un valor de partición de 10 (Hernandez Orallo, 2005). Buscando un proceso de optimización y de obtención de mejores resultados con estos algoritmos se parametrizaron de la siguiente manera:

Clasificador DTNB (Decision table/naive bayes): Para este algoritmo se determinó que la medida de evaluación sería la exactitud para los valores nominales o discretizados y la medida para los valores numéricos sería RMSE (root-mean-square error).

Clúster DBScan (Density-Based Scan): Para este algoritmo se determinó que el valor de la épsilon fuera 0.9

Resultados

Resultados con algoritmos predictivos

De los algoritmos predictivos aplicados a los diferentes dataset divididos a partir del atributo de segmento, se obtuvieron una gran cantidad de resultados de cada uno de los indicadores de rendimiento, de estos se mostraran solo algunos ejemplos de la gran lista de relaciones y tendencias entre datos para cada uno de los atributos de filtro como lo es departamento (tabla 1), municipio (tabla 2), posición laboral (tabla 3) y género (tabla 4), a partir de cada uno de los indicadores de rendimiento.

Tabla 1. Resultados por departamento

Valor de filtro	Indicador de rendimiento	Uno de los múltiples resultados obtenidos
Antioquia	Según ingresos	Las personas con un estrato económico de 6, con factor de expansión anual mayor a 53 tienden a tener salarios no mayores a $833.333. (100% de clasificaciones correctas)
	Según estrato	Personas que cotizan en alguna entidad de salud y que tienen un factor de expansión anual entre 50.02 y 99.41 tienden a ser estrato 3. (92% de clasificaciones correctas)
	Según edad	Personas con un factor de expansión anual mayor a 79.93 y con un nivel educativo de 9 grado tienden a ser personas mayores a 39 años (85% de clasificaciones correctas)
Cundinamarca	Según ingresos	Hombres que son de estrato 5, con factor de expansión anual entre 50 y 100 tienden a no ganar más de 100.000 pesos. (75% de clasificaciones correctas)
	Según estrato	Las personas que tiene un nivel educativo de básica primaria, con un factor de expansión anual de entre 150 y 200 y que pagan una entidad de salud tienden a ser estrato 2. (75% de clasificaciones correctas)
Norte de Santander	Según ingresos	Los hombres de estrato 2 con factor de expansión anual entre 16.50 y 35.84 no ganan más de 441.666 pesos. (83% de clasificaciones correctas)
	Según estrato	Personas con ingresos inferiores a 425.000 pesos, con un nivel educativo de hasta grado 9, y factor de expansión anual entre 35.84 y 58.69 tienden a ser de estrato 2. (87% de clasificaciones correctas)

	Según edad	Personas de estrato 2, con nivel académico de grado 11, y con factor de expansión anual de máximo 50 tienden a ser menores de 23 años. (75% de clasificaciones correctas)
Valle del Cauca	Según ingresos	Las mujeres de estrato 2, con un nivel académico de grado 11, con un factor de expansión anual entre 50.02 y 79.93 tienden a no ganar más de 166.000 pesos. (80% de clasificaciones correctas)
	Según estrato	Personas con un nivel académico de 11, que son mayores de 20 años, que no pagan una entidad de salud y que tienen un factor de expansión anual mayor a 50.02 tienden a ser de estrato 3. (100% de clasificaciones correctas)
Caldas	Según ingresos	Las mujeres de estrato 3, con un nivel académico de pregrado o superior, con factor de expansión anual a 16.50 tienden a ganar $480.000. (100% de clasificaciones correctas)
	Según estrato	Hombres con ingresos de 233.333 pesos, que tienen un factor de expansión anual inferior a 16.50, con un nivel académico de básica primaria y que pagan alguna entidad de salud tienden a ser estrato 2. (71% de clasificaciones correctas)
	Según edad	Las personas con factor de expansión anual no mayor a 16.50, que son de estrato 3, que tiene un nivel académico de grado 9, que pagan una entidad de salud no superan los 46 años. (83% de clasificaciones correctas)

Fuente: elaboración propia

Tabla 2. Resultados por municipio

Valor de filtro	Indicador de rendimiento	Uno de los múltiples resultados obtenidos
Bucaramanga	*Según ingreso*	Personas de estrato 1, con factor de expansión anual entre 39.67 y 52.67 tienden a ganar entre 400.000 y 500.000 pesos. (90% de clasificaciones correctas)
	Según estrato	Personas con un nivel académico de pregrado o superior, con ingresos entre 300.000 y 400.000 pesos, entre 20 y 27 años y con factor de expansión anual entre 26.68 y 39.67 tienden a ser de estrato 3. (100% de clasificaciones correctas)
	Según edad	Personas con un nivel académico de pregrado o superior, con ingresos inferiores a 100.000 pesos, y factor de expansión anual entre 26.68 y 39.67 tienden a tener entre 27 y 34 años. (87% de clasificaciones correctas)
Cali	*Según ingreso*	Personas entre 41-48 años, con nivel educativo de grado 9, con factor de expansión anual entre 78.67 y 91.66 tienden a no ganar más de 100.000 pesos. (80% de clasificaciones correctas)

	Según estrato	Personas con ingresos entre 300.000 y 400.000 pesos, con edad entre 41-48 años y con factor de expansión anual entre 65.67 y 78.67 tienden a ser estrato 3. (100% de clasificaciones correctas)
	Según edad	Personas de estrato 2, con ingresos totales entre 400.000 y 500.000 pesos, y con factor de expansión anual entre 65.67 y 78.67 tienden a tener entre 27 y 34 años. (100% de clasificaciones correctas)
Medellín	*Según estrato*	Personas con nivel académico de pregrado o superior, con ingresos entre 500.000 y 600.000 pesos y factor de expansión anual entre 26.68 y 39.67 tienden a ser de estrato 2. (72% de clasificaciones correctas)
Manizales	Según ingresos	Las mujeres de estrato 3, con un nivel académico de pregrado o superior, con factor de expansión anual a 16.50 tienden a ganar $480.000. (100% de clasificaciones correctas)
	Según estrato	Hombres con ingresos de 233.333 pesos, que tienen un factor de expansión anual inferior a 16.50, con un nivel académico de básica primaria y que pagan alguna entidad de salud tienden a ser estrato 2. (71% de clasificaciones correctas)
	Según edad	Las personas con factor de expansión anual no mayor a 16.50, que son de estrato 3, que tiene un nivel académico de grado 9, que pagan una entidad de salud no superan los 46 años. (83% de clasificaciones correctas)

Fuente: elaboración propia

Tabla 3. resultados por posición laboral

Valor de filtro	**Indicador de rendimiento**	**Uno de los múltiples resultados obtenidos**
Obrero o empleado	*Según ingreso*	Personas que trabajan de 42 a 53 horas por semana, y que tienen un factor de expansión anual inferior a 91.26 tienden a ganar entre 250.000 y 500.000 pesos. (100% de clasificaciones correctas)
Empleado del gobierno	*Según ingreso*	Personas que trabajan entre 32 y 42 horas por semana, que en donde trabajan son aproximadamente 9 empleados y que tienen un factor de expansión anual inferior a 91.26 tienden a ganar entre 500.000 y 750.000 pesos. (100% de clasificaciones correctas)
Peón	*Según ingreso*	Personas que trabajan entre 42 y 53 horas por semana, que en la empresa donde están tienen alrededor de 4 empleados, tienen ingresos entre 500.000 y 700.000 pesos. (93% de clasificaciones correctas)

Fuente: elaboración propia

Tabla 4. Resultados por sexo

Valor de filtro	Indicador de rendimiento	Uno de los múltiples resultados obtenidos
Hombres	*Según ingreso*	Hombres con un nivel académico de pregrado o superior, que pagan alguna entidad de salud, que tienen entre 63 y 70 años, y que son de estrato 6 tienen ingresos menores a 300.000 pesos. (80% de clasificaciones correctas)
	Según estrato	Hombres con un nivel académico de pregrado o superior, que tienen entre 49 y 56 años, que tienen ingresos entre 300.000 y 620.000 pesos y que no pagan una entidad de salud tienden a ser estrato 3. (92% de clasificaciones correctas)
	Según edad	Personas de estrato 4, que tienen un pregrado o superior, que tienen ingresos entre 620.000 y 900.000 pesos, tienen entre 34 y 49 años. (82% de clasificaciones correctas)
Mujeres	*Según ingreso*	Mujeres de estrato 6, que tienen entre 49 y 56 años, que cotizan en alguna entidad de salud, y que tienen un nivel académico de pregrado o superior tienden a ganar entre 600.000 y 900.000 pesos. (92% de clasificaciones correctas)
	Según estrato	Las mujeres con nivel académico de pregrado o superior, que tienen entre 34 y 41 años, y que tienen ingresos entre 900.000 y 1'500.000 pesos tienden a ser estrato 4. (83% de clasificaciones correctas)
	Según edad	Mujeres de estrato 3, con un nivel académico de grado 9, que tienen ingresos entre 300.000 y 620.000 pesos y que cotizan en alguna entidad de salud tienen entre 34 y 41 años. (82% de clasificaciones correctas)

Fuente: elaboración propia

Resultados con algoritmos descriptivos

Con el algoritmo descriptivo que se obtuvo los mejores resultados fue DBScan, y se utilizó para validar el conocimiento encontrado previamente con los algoritmos predictivos y adicionalmente, se utilizó para encontrar nuevo conocimiento que fuera sido omitido por el enfoque predictivo (tabla 5) (tabla 6).

Tabla 5. Distribución de datos aplicando clustering

# De clúster	% de datos del clúster	# de clúster	% de datos del clúster	# de clúster	% de datos del clúster	# de clúster	% de datos del clúster
0	14%	10	1%	20	1%	30	1%
1	1%	11	1%	21	1%	31	1%
2	8%	12	8%	22	1%	32	1%

3	1%	13	1%	23	1%	33	1%
4	8%	14	1%	24	1%	34	1%
5	8%	15	1%	25	1%	35	1%
6	1%	16	1%	26	1%	36	1%
7	8%	17	1%	27	1%		
8	15%	18	1%	28	1%		
9	1%	19	1%	29	1%		

Fuente: elaboración propia

Tabla 6 . Grupos encontrados aplicación de clustering

# De Clúster	Nivel Edu.	Ent. Salud	Ing. Total	Exp. Anual	Sexo	Pos. Laboral	Horas x Semana	Años	Estrato
0 (14%)	5	1	300>	(-inf-11]	1	4	(11.6-21.2]	(49-55.8]	2
2 (8%)	6	1	150-300	(37-43]	2	4	(11.6-21.2]	(21.8-28.6]	3
4 (8%)	6	2	150-300	(11-16]	1	4	(11.6-21.2]	(55.8-62.6]	6
7 (8%)	4	1	150-300	(11-16]	2	1	(11.6-21.2]	(21.8-28.6]	1
8 (15%)	6	1	300>	(11-16]	1	4	(11.6-21.2]	(28.6-35.4]	4
5 (8%)	5	1	150-300	(-inf-11]	1	4	(11.6-21.2]	(42.2-49]	3
12 (8%)	4	1	150-300	(11-16]	1	4	(11.6-21.2]	(35.4-42.2]	3

Fuente: elaboración propia

Discusión de resultados

Con la aplicación de esta investigación se esperaba encontrar relaciones entre el ingreso total, la edad, el nivel académico, el estrato, factor de expansión anual, entre otros, según el departamento, municipio, posición laboral, y género. Los resultados fueron los siguientes:

Algoritmos predictivos

Los anteriores resultados mostrados en la sección de resultados, indican claramente que hay patrones marcados para uno de los atributos de filtro según los diferentes indicadores claves de rendimiento. A partir de estos se pueden establecer futuros proyectos que analicen el trasfondo de dichos comportamientos con el fin de mejorar las condiciones de vida de los ciudadanos colombianos.

Algoritmos descriptivos

Los resultados obtenidos aplicando este enfoque no fueron los esperados. En total el algoritmo encontró un total de 945 grupos, el cual era un valor demasiado elevado y que no indica un comportamiento definido. Para solucionar este problema se decidió segmentar aún más la información, teniendo un dataset solo para la región de Manizales ya que sería la de mayor importancia según el sector de estudio, y con este nuevo conjunto de datos se obtuvo un total

de 1033 registros. Luego de aplicar el algoritmo Dbscan sobre este nuevo dataset se obtuvieron un conjunto de resultados (tabla 5).

Se obtuvieron un total de 37 grupos y aunque es menor que el resultado inicial siguen siendo una gran cantidad de grupos. Si se observa en detalle la mayor parte de los grupos encontrados solo contienen 1% del total de los datos analizados, por lo que estos fueron omitidos dejando únicamente los grupos más significativos. Con esta reducción se dejaron solo 7 grupos con unas características puntuales (tabla 6).

Con el grupo 0 y 2 se constató que el conocimiento descubierto para el sector de Manizales era realmente valido (Comparándolo con los resultados obtenidos en el proceso predictivo), y con los otros grupos se encontró nuevo conocimiento que con los algoritmos predictivos no se había identificado.

A partir de las aplicaciones de estos modelos se encontró que sí existe una relación entre el ingreso total, la edad, el nivel académico, el estrato, factor de expansión anual, entre otros, según el departamento, municipio, posición laboral, y género. Con estos resultados se puede dar una descripción clara y concisa a la ciudadanía en general sobre el comportamiento general del *dataset*, y a las entidades que quieran utilizar los datos en bruto para la toma de decisiones según su campo de acción, tendrán una idea previa sobre lo que encontrara, como también un enfoque sobre en qué dirección deben enfocar sus análisis para la obtención de resultados.

Conclusiones

El modelo adaptado de la metodología KDD aplicado en datos del sector público fue satisfactorio, teniendo como elemento central la definición del objetivo de la aplicación de descubrimiento de conocimiento para especificar los atributos de segmento, indicadores de filtro y de rendimiento. En la aplicación de este modelo, aunque los procesos descriptivos permiten tener una visión detallada de los datos, si se tienen claramente definidos los indicadores claves de rendimiento, los procesos predictivos ofrecen adicionalmente una descripción que pueden ser una solución muy efectiva, debido a que al ser un proceso supervisado teniendo como atributo clase cada uno de los indicadores claves de rendimiento, se obtiene resultados segmentados y centrados en los objetivos del proceso. Sin embargo, aunque el uso de procesos de descubrimiento de conocimiento en el entorno o-gov es efectivo para encontrar información adicional en los datos públicos, no es muy eficiente, dando paso a un proceso que puede generar un cuello de botella cuando se desee tratar grandes volúmenes de datasets, por lo que se recomienda o se deja abierta la posibilidad como trabajos futuros, implementar otros enfoques para descubrir conocimiento en dichos datos, como la aplicación de web semántica, con el fin de determinar cuál de los procesos evaluados es el más eficiente.

Referencias bibliográficas

Carreño, J. (2008). Descubrimiento de conocimiento en los negocios. *Panorama, 4*. 2(4).pp 59-76.

Fayyad, U. Piatesky-Shapiro, G. & Smyth, Padharaic. (1996a). *Advances in knowledge discovery and data mining*. Masachussetts: MIT Press.

Fayyad, U. Piatesky-Shapiro, G. & Smyth, Padharaic. (1996b). The KDD process for extracting useful knoledge from volumes of data. *Communications of the ACM, 39*(11), 27-34.

Fayyad, U. Piatesky-Shapiro, G. & Smyth, Padhraic. (1996). From data mining to knowledge discovery in databases. *IA Magazine, 17, No 3*, 37-54.

Gómez, C. & Gascó, M. (2012). *Y ahora… gobierno abierto: nuevos términos en la constante búsqueda por la transparencia y la rendición de cuentas.* Paper presented at the XVII Congreso Internacional del CLAD sobre la Reforma del Estado y de la Administración Pública, Cartagena, Colombia.30 oct. - 2 nov. 2012.

Gutiérrez, J. E. (2012). *Descubrimiento de conocimientos en la base de datos académica de la universidad autónoma de Manizales aplicando redes neuronales.* (Maestria en Gestion y Desarrollo de Proyectos de Software), Universidad Autónoma de Manizales, Manizales, Colombia.

Hernández, C. L. & Rodrigez, J. E. (2008). Preprocesamiento de datos estructurados. *Revista Vinculos, 8*(28), 27-48.

Hernandez, J. (2005). *Encycpledia of database technologies and applications* (Vol. 54). Valencia, España: Technical University of Valencia, Spain.

Maimon, O. & Rokach, L. (2010). *Data mining and knowledge discovery handbook* (2 ed. Vol. 1). New York: Springer.

Ogan, M. Y. & Oana, V. (2012). A review of dashboards in performance management: Implications for design and research. *International Journal of Accounting Information Systems, 13*, 41-59.

Orallo, J. H. Quintana, J. & Ramirez, C. (2005). *Introducción a la mineria de datos*. Madrid, España: Departamento de Sistemas Informáticos y Computación, Universidad Politécnica de Valencia.

Reyes, J. & García, R. (2005). El proceso de descubrimiento de conocimiento en bases de datos. *Ingenierias, 8*. pp. 37-47.

Rice, A, Abshire, D, Christakis, M, & Sherman, G. (2010). The Assessment Movement toward Key Performance Indicators. *Presented at the meeting of NASPA International Assessment and Retention Conference,*.

Salazar, J. (2015). *Marco de referencia para la implementación del mapa de ruta establecido en los lineamientos nacionales de apertura de datos del sector público y su integración con procesos de descubrimiento de conocimiento e inteligencia de negocios.* (Master en Ingenieria Computacional), Universidad de Caldas, Manizales, Colombia.

Sourouni, A. Kourlimpinis, G. Mouzakitis, S. & Askounis, D. (2010). Towards the government transformation: An ontology-based government knowledge repository. *Computer Standards & Interfaces, 32, No 2.*, 44.