ISSN 0717-9103
ISSN Online 0718-8307

MINERÍA DE DATOS EN LA ENCUESTA PERMANENTE DE HOGARES 2009, UNIVERSIDAD NACIONAL DEL LITORAL, ARGENTINA.

APPLICATION OF DATA MINING TO PANEL SURVEYS 2009, NATIONAL UNIVERSITY OF THE LITORAL, ARGENTINA.

Denis Lizazo Torres[1], Roberto Delfor Meyer[1], Verena Torres Cárdenas[2]
[1]Sistema de Monitoreo Social del Litoral. Secretaría de Extensión. Universidad del Litoral Santa Fé, Argentina. rmeyer@unl.edu.ar.
[2]Departamento de Biomatemática. Instituto de Ciencia Animal. Ministerio de Educación Superior. vtorres@ica.co.cu. San José de las Lajas, La Habana, Cuba.

RESUMEN

En este trabajo se aplicó por primera vez la Minería de Datos a la información generada por la Encuesta Permanente de Hogares del año 2009, del Observatorio Social de la Universidad del Litoral en Santa Fe, Argentina. El objetivo fue utilizar algoritmos de clasificación que poseen enfoques descriptivos, y es una de las tareas más utilizada en Minería de Datos para obtener nuevos conocimientos sobre la familia santafecina. La base de datos fue confeccionada con la información en el relevamiento de 2009, denominado Onda 2009, y se procesó con el software de minería de datos Weka 3.6.2. Los algoritmos de clasificación que mejores resultados aportaron fueron: ZeroR, Ridor y J48, pertenecientes a los grupos de Reglas y Árboles de Decisión, siendo el algoritmo de clasificación J48 el que mejor clasifica a la Encuesta Permanente de Hogares 2009 del Observatorio Social de la Universidad Nacional del Litoral.

Palabras Claves: Minería de Datos, Weka, Algoritmos de Clasificación, Relevamiento.

ABSTRACT

In this work it was applied the Data Mining for the first time, to the information generated by the Permanent Survey of Homes of the year 2009, of the Social Observatory of the University of the Litoral in Santa Fe, Argentina. The objective was to use classification algorithms that possess descriptive focuses and it is one of the tasks more used in Data Mining, to obtain new knowledge on the family of Santa Fe. The database was made with the information of the Permanent Survey of Homes of the year 2009 and it was processed with the software of data mining Weka 3.6.2. The classification algorithms that better results contributed were: ZeroR, Ridor and J48, belonging to the groups of Rules and Trees of Decision, being the classification algorithm J48 the one that better it classifies to the Permanent Survey of Homes 2009, of the Social Observatory of the National University of the Litoral.

Keyword: Data Mining, Weka, Classify Algorithms, Survey.

Autor para correspondencia: de.lizazotorres@yahoo.com.ar
Recibido: 27.01.2011 Aceptado: 13.08.2011

INTRODUCCIÓN

Desde el siglo pasado, la información ha ido ocupando un sitio preponderante en el proceso de evolución de la sociedad humana, convirtiéndose en un importante elemento y base del conocimiento. En las últimas décadas del Siglo XX, con la expansión de las nuevas Tecnologías de la Información y la Comunicación, tuvo lugar lo que se ha dado en llamar Sociedad de la Información. Cisnero *et al.* (2002) plantean que en todos los ámbitos de la vida social y económica la cantidad de información que se genera actualmente puede convertirse en conocimiento, teniendo en cuenta los mecanismos de su producción, tratamiento y distribución.

Según Cabrera (2004), la información, como un recurso estratégico, deviene en la sociedad actual como factor primordial del éxito de cualquier organización. Cada año el volumen de información aumenta exponencialmente, por lo que en ocasiones resulta difícil encontrar o inferir el conocimiento contenido en dicha información. Actualmente, la comunidad científica está enfrascada en la búsqueda y desarrollo de herramientas y técnicas que permitan procesar la gran cantidad de información disponible y obtener conocimientos acorde a las exigencias actuales.

La Minería de Datos es una de las herramientas que ha emergido para el procesamiento de grandes cantidades de información. Sus algoritmos permiten obtener conocimiento relevante desde grandes cantidades de datos, y es el resultado de la combinación de los avances de las Tecnologías de la Información y la comunicación con la Estadística, las Bases de Datos, el Reconocimiento de Patrones, la Inteligencia Artificial, Aprendizaje Automático y los Sistemas para el Apoyo a la Toma de Decisiones. Witten & Frank (2000), conceptualizan a la Minería de Datos como el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos.

Con la expansión de la Sociedad de la Información, y formando parte de ésta, tiene lugar la aparición y desarrollo de los Observatorios Sociales, los cuales realizan estudios de múltiples fenómenos en el contexto de las ciencias sociales. Es en este marco que se creó, en el año 2000, un sistema de monitoreo de opiniones de demandas sociales en la Secretaría de Extensión de la Universidad Nacional del Litoral en Santa Fe, Argentina, para favorecer los procesos de toma de decisiones de las autoridades gubernamentales. Este Sistema de Monitoreo Social del Litoral es desarrollado a través de cuestionarios aplicados a los diferentes grupos sociales que aportan información sobre la región de Santa Fe-Santo Tomé, mediante un "Panel Detallista". Para Fuentes (2007), se trata de un instrumento de medición, investigación y análisis de la situación socio –económico- laboral, en un ámbito regional determinado, compuesto por un panel de actores sociales que brinda información oportuna e información especializada, oportuna y de calidad para una mejor toma de decisiones.

El panel genera en forma continua y dinámica información que, al ser analizada, permite captar los cambios en las opiniones que se producen en la sociedad y el análisis de la percepción social de temáticas de interés público, que afectan a la sociedad en su conjunto y que favorecerán los procesos de toma de decisión de las organizaciones involucradas. Los relevamientos desarrollados en los años 2005, 2006, 2007 y 2009 fueron analizados a través de análisis de clusters, algoritmo de k-medias, análisis de discriminantes y tablas de contingencia.

El objetivo de este trabajo es presentar los resultados obtenidos al procesar por vez primera mediante la Minería de Datos, a través de algoritmos de Clasificación, utilizando el software Weka, la información proveniente de la Encuesta Permanente de Hogares efectuada en el año 2009, del Observatorio Social "Sistema de Monitoreo Social del Litoral". Se define el mejor algoritmo de clasificación al procesar la Onda 2009, todo lo cual permitió obtener nuevos conocimientos, reglas y relaciones para establecer las características de la familia santafecina, con el propósito de utilizarlos en futuros relevamientos.

ISSN 0717-9103
ISSN Online 0718-8307

MATERIALES Y MÉTODOS

La aplicación de la Encuesta Permanente de Hogares constituye la base del Sistema de Monitoreo Social del Litoral. Esta encuesta se basa en un cuestionario de respuestas múltiples, conformado por secciones que recopilan información que incluye, entre otros aspectos, la siguiente información:

A. datos referidos al grupo conviviente en el hogar, (A1, A2, ..., A7),
B. actividades recreativas o culturales por parte de familia, uso de Internet, capacidad de ahorro, (B1, B2 ..., B4),
D. relación de la familia con el lugar donde vive, (D1, D2, D3),
E. tipo de servicio de salud que usa, cobertura por obra social, especialidades que consulta y guardia de urgencia que utiliza. (E1, E2, ..., E7).

Otros aspectos que se preguntan en el cuestionario son:

F. la percepción que tiene la familia sobre los tipos de delitos que han sufrido, niveles de delitos en el barrio, actuación del gobierno provincial respecto a este flagelo, causas que lo propician y las medidas a tomar para enfrentarlo, (F1, F2, ..., F9),
H. actividades de capacitación, cumplimiento de su objetivo, modalidad y cursos que desearía realizar, (H1 y H2).

En el caso de la onda 2009 de la Encuesta Permanente de Hogares del Sistema de Monitoreo Social del Litoral, por primera vez se realizó el análisis de la base de datos aplicando las novedosas técnicas presentes en la minería de datos, como una vía para obtener nuevo conocimiento, reglas y relaciones, a partir de clasificar a la familia santafecina en base a las variables presentes en el cuestionario.

De la población de 104931 hogares se seleccionó una muestra de 1091 de éstos para la Onda del año 2009, a partir de cuotas geográficas previamente establecidas en la estratificación de la ciudad, según Necesidades Básicas Insatisfechas en base a datos censales. La información revelada por los encuestadores contenía las respuestas dadas a las 240 variables de la encuesta por las 1091 familias, para un total de 261840 registros iniciales.

Se implementó un Data Warehouse para preparar los datos fuentes. Para, en una primera etapa de organización, realizar un primer acercamiento, comprenderlos, consolidarlos e integrarlos con el propósito de organizarlos. En el Data Warehouse se consideraron variables o atributos discretos, categóricos nominales y de tipo cadena.

La aplicación de la Minería de Datos siguió los criterios de Merlino *et al.*, (2005), quienes definen las etapas de este proceso como:

a) Obtención de Datos
b) Transformación de los Datos
c) Aplicación de la Técnica de Explotación de Datos
d) Evaluación de los Resultados Obtenidos

Para la implementación de estas etapas se utilizó el software de Minería de Datos Weka, Versión 3.6.2 (2010); herramienta programada en Java por investigadores de la Universidad de Waikato, Nueva Zelanda, bajo licencia GPL. Su condición de Software Libre ha favorecido que sea una de las suites más utilizadas en el área en los últimos años.

El primer procesamiento del Data Warehouse en Weka consistió en la selección, limpieza y transformación de los datos. Esto significó eliminar de la base de datos todas las variables o atributos que tuvieran menos del 50% de las mediciones por ser incompletas; se eliminaron datos por presentar inconsistencias o por ser irrelevantes para el proceso. Otras instancias

fueron corregidas, completadas y estandarizadas. Se realizó una selección de las variables que más información podían aportar al análisis, luego de lo cual la cantidad de atributos quedó reducida en 152; para las mismas 1091 familia, resultaron en total 165832 registros.

En Weka, los atributos de tipo numéricos del Data Warehouse fueron transformados en atributos nominales a partir de la discretización; los atributos de cadena conservaron su tipo. Se definió el atributo A3_TipoFlia (Tipo de Familia), como atributo de clase de instancia, convirtiéndose en la variable objetivo a predecir, ya que se trata de caracterizar a la familia santafesina en función del resto de los atributos presentes en el cuestionario de la Onda 2009. La elección de esta variable como atributo de clase se hizo considerando que es la familia la mejor representación de la realidad social, a partir de las opiniones que vierte sobre temáticas de interés público que afectan a toda la sociedad en su conjunto. Es la familia la que aporta información sobre cambios en hábitos y costumbres de actividades físicas, deportivas, recreativas y culturales; o las preferencias acerca de la vivienda, o cambios de hábitos y costumbres, así como cambios en los roles e imagen de las instituciones públicas y privadas, al mismo tiempo es receptora de las gestiones de instituciones, servicios y programas.

La clasificación es un proceso en el que un atributo individual es asignado a un grupo a partir de sus características y rasgos; proporciona modelos que clasifican las nuevas instancias con un valor concreto para su clase. Tiene un enfoque descriptivo cuando se trata de conocer las variables y valores más significativos de las instancias de cada tipo de clase. Es quizás la tarea más utilizada en Minería de Datos y, a partir de definir un atributo llamado clase de instancia, se utiliza al resto de los atributos para predecir la clase. El objetivo es maximizar la razón de precisión de la clasificación para nuevas instancias. Hernández *et al.* (2004)

De las opciones que brinda Weka para validar el proceso, se empleó la variante de Percentage Splits, opción que divide los datos en dos grupos, de acuerdo con el porcentaje indicado (50% en este estudio). El valor indicado es el porcentaje de instancias que se utilizan para construir el modelo, el que es evaluado luego sobre el restante porcentaje de los datos; comparando la clase predicha con la clase real de las instancias. Las instancias en cuestión utilizadas para la construcción del modelo y las que se utilizan para evaluar el resultado obtenido son seleccionadas aleatoriamente por la propia herramienta.

Para medir la calidad del algoritmo seleccionado se utilizaron los parámetros de Weka que más explicación aportan a la descripción de los resultados. Estos son:

- ✓ Distribución de errores cometidos por los algoritmos en la clasificación de las instancias.
- ✓ Proporción de Verdaderos Positivos (TP Rate): Proporción de ejemplos que fueron clasificados de una clase dada, de entre todos los casos que de verdad tienen esa clase.
- ✓ Proporción de Falsos Positivos (FP Rate): Casos clasificados de una clase determinada, pero que en realidad pertenecen a otra.
- ✓ Precisión: Mide el número de términos correctamente clasificados respecto al total de los predichos, sean éstos verdaderos o falsos.
- ✓ Cobertura (Recall): Mide la proporción de términos correctamente reconocidos respecto al total de términos reales; es inversamente proporcional a la precisión.
- ✓ Total de F-measure: Caracteriza con único valor la bondad de un clasificador o algoritmo.

ISSN 0717-9103
ISSN Online 0718-8307

RESULTADOS

Los algoritmos de clasificación en Weka, que mejores resultados y nuevo conocimiento aportaron en la clasificación de la Onda 2009 del Observatorio Social de la UNL, pertenecen a las familias de:

- Árboles de Decisión: Tienen la función de clasificar los datos y predecir el comportamiento de manera estadística, a partir de construir diagramas lógicos que categorizan y representan una serie de condiciones de forma sucesiva.
- Reglas: Técnicas de clasificación de clases sobre la base de un conjunto de reglas que siguen el principio "Si ... Entonces ...", en orden jerárquico.

Un Árbol de Decisión es un conjunto de condiciones organizadas en una estructura jerárquica, de tal manera que la decisión final a tomar se puede determinar siguiendo las condiciones que se cumplen desde la raíz del árbol hasta algunas de sus hojas. Señalan Hernández, Ramírez, & Ferri (2004), una de las grandes ventajas de los árboles de decisión es que, en su forma más general, las operaciones posibles a partir de una determinada condición son excluyentes. Estos mismos autores plantean que las Reglas pueden ser consideradas una generalización de los árboles de decisión, los cuales pueden expresarse como un conjunto de reglas.

El primer clasificador utilizado fue el ZeroR, perteneciente al grupo de clasificadores de Reglas y en la tabla Nº 1 se muestran los principales resultados de aplicar este algoritmo.

Tabla Nº 1. Distribución de la Clasificación de las Instancias al Aplicar el Algoritmo ZeroR.

Instancias Correctamente Clasificadas			386		70.83%	
Instancias Incorrectamente Clasificadas			159		29.17%	
Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
1	1.00	1	0.70	1.00	0.82	0.50
2	0.00	0.00	0.00	0.00	0.00	0.50
3	0.00	0.00	0.00	0.00	0.00	0.50
4	0.00	0.00	0.00	0.00	0.00	0.50
5	0.00	0.00	0.00	0.00	0.00	0.50
Total	0.70	0.70	0.50	0.70	0.58	0.50

Los algoritmos de clasificación de Weka, cuyos resultados superaron el 70.83% de aciertos en la clasificación correcta de las instancias para la Onda 2009, criterio aportado por el Algoritmo ZeroR, son considerados como algoritmos adecuados para clasificar la Onda 2009. Cumplieron esta condición los algoritmos J48 del grupo de los Árboles de Decisión y el Ridor de la familia de las Reglas.

El algoritmo J48 se basa en el C4.5, uno de los algoritmos de aprendizaje de árboles de decisión más populares y efectivos, ya que genera un árbol de decisión a partir de los datos mediante particiones realizadas recursivamente. Considera todas las pruebas posibles que pueden dividir el conjunto de datos y selecciona la prueba que genera mayor obtención de información.

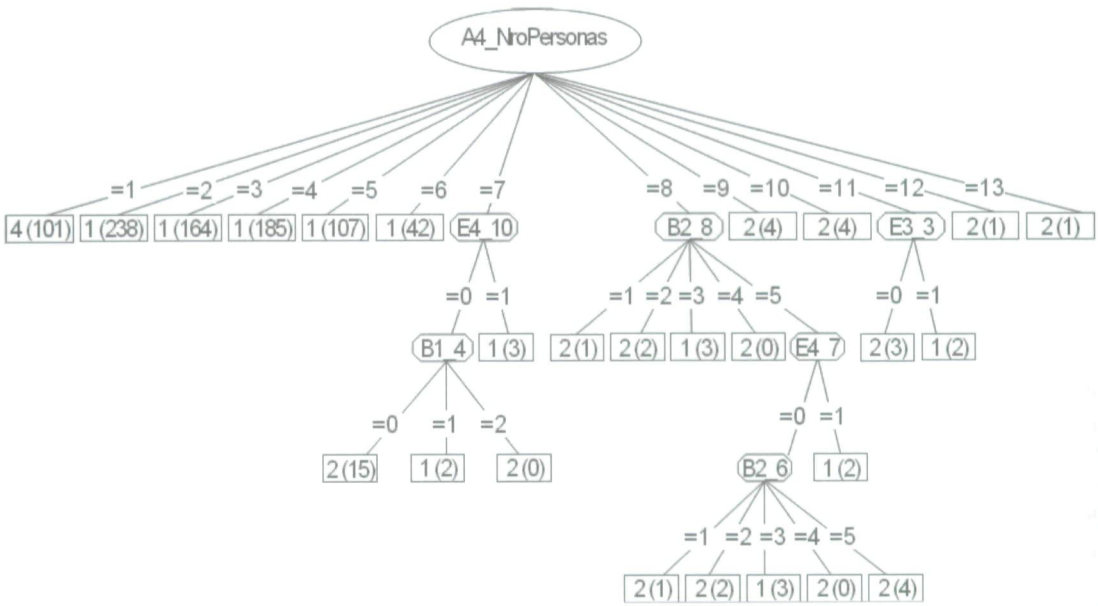

Figura N° 1. Árbol de Decisión Generado por el Algoritmo J48.

En el árbol de decisión de la Onda 2009 (Figura 1), con el algoritmo J48 está la distribución por número de personas (Atributo A4) de los diferentes tipos de familia para el atributo de clase Tipo de Familia. Se observa como la Familia Tipo (padre y/o madre e hijos), la Familia Parental (más de una familia conviviendo, parientes entre sí) y la Familia Unipersonal, son las más representadas en la base de datos y que la tipología del grupo habitante la constituyen desde 2 hasta 13 miembros para las familias Tipo y Parental. Las relaciones que aportan nuevo conocimiento reflejan que las familias compuestas por 7, 8 u 11 integrantes se vinculan a los atributos de la sección B correspondiente a actividades Culturales y Recreativas, así como de la sección E de Salud.

La clasificación de la Onda 2009, utilizando el algoritmo J48, incluido dentro de la familia de algoritmos de Árboles de Decisión, aparecen en la tabla N° 2.

Tabla N° 2. Distribución de la Clasificación de las Instancias al Aplicar el Algoritmo J48.

Instancias Correctamente Clasificadas		435		79.81%		
Instancias Incorrectamente Clasificadas		110		20.18%		
Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
1	0.95	0.58	0.79	1.00	0.86	0.74
2	0.16	0.04	0.47	0.16	0.24	0.68
3	0.00	0.00	0.00	0.00	0.00	0.42
4	1.00	0.00	1.00	1.00	1.00	1.00
5	0.00	0.00	0.00	0.00	0.00	0.59
Total	0.79	0.42	0.73	0.79	0.75	0.75

En el caso de estudio de la Onda 2009, el algoritmo de Reglas Ridor clasifica correctamente al 80% de las instancias. Este algoritmo Ridor de clasificación de Weka es una implementación de la regla de aprendizaje Riple-Down y genera las mejores reglas para el modelo, conjuntamente con sus mejores excepciones. El comportamiento de los parámetros con la aplicación de este algoritmo que crea un total de 63 Reglas, aparecen en la tabla N° 3.

ISSN 0717-9103
ISSN Online 0718-8307

Tabla Nº 3. Distribución de la Clasificación de las Instancias al Aplicar el Algoritmo Ridor.

Instancias Correctamente Clasificadas			436		80.00%	
Instancias Incorrectamente Clasificadas			109		20.00%	
Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
1	1.00	0.68	0.78	1.00	0.87	0.71
2	0.00	0.00	0.00	0.00	0.00	0.63
3	0.00	0.00	0.00	0.00	0.00	0.63
4	1.00	0.00	1.00	1.00	1.00	1.00
5	0.00	0.00	0.00	0.00	0.00	0.58
Total	0.80	0.48	0.64	0.80	0.71	0.72

Las reglas generadas fueron:

A3_TipoFlia=5 (19)
Except (D2_C=0) => A3_TipoFlia=4 (566) [289]
Except (B3_2=0) Y (E5_A=1) => A3_TipoFlia=1 (277) [145]
Except (E5_GuardNiños=3) Y (F9_7=0) => A3_TipoFlia=2 (15) [6]
Except (E4_1=1) => A3_TipoFlia=1 (71) [37]
Except (E4_2=1) Y (H1=2) Y (B1_6=0) Y (F4=3) Y (A3_B_TipoVivienda=1) => A3_TipoFlia=2 (13) [3]
Except (E6=1) Y (B2_11=5) Y (E4_8=1) => A3_TipoFlia=1 (31) [10]
Except (F3_1=3) Y (F5=2) => A3_TipoFlia=1 (19) [13]
Except (B2_1=2) Y (E4_4=0) => A3_TipoFlia=1 (26) [12]
Except (A3_B_TipoVivienda=1) Y (F9_8=0) => A3_TipoFlia=4 (76) [25]
Except (B3_2=0) Y (D1_A=4) => A3_TipoFlia=1 (162) [90]
Except (F3_1_12_AccPolicBarr=0) Y (E1_TipoServicPublic=1) => A3_TipoFlia=2 (10) [1]
Except (E4_1=1) Y (H1=2) Y (B4=2) => A3_TipoFlia=2 (6) [2]
Except (E4_1=1) => A3_TipoFlia=1 (102) [48]
Except (D2_C=0) Y (B1_12=0) Y (E4_8=1) Y (E3_2=0) Y (B2_8=5) => A3_TipoFlia=2 (9) [1]
Except (D1_A=4) Y (B1_13=0) Y (F9_10=1) Y (H1=2) Y (B2_11=5) => A3_TipoFlia=2 (14) [4]
Except (B3_2=0) => A3_TipoFlia=1 (59) [23]
Except (E7=2) => A3_TipoFlia=3 (37) [16]
Except (E4_3=0) => A3_TipoFlia=1 (109) [51]
Except (B3=1) => A3_TipoFlia=1 (20) [14]
Except (E1ServicSalud=1) => A3_TipoFlia=1 (5) [3]
Except (F9_1=1) Y (E3_1=0) => A3_TipoFlia=2 (6) [1]
Except (B3_2=1) => A3_TipoFlia=2 (53) [17]
Except (E5_GuardNiños=0) Y (F3_1_12_AccPolicBarr=3) Y (B1_12=0) => A3_TipoFlia=1 (20) [7]
Except (A4_NroPersonas=2) Y (B1_13=0) => A3_TipoFlia=1 (12) [4]
Except (B2_17=3) Y (E4_2=0) => A3_TipoFlia=1 (13) [5]
Except (B2_1=5) Y (F8_2=0) Y (B1_14=7) => A3_TipoFlia=1(10) [3]
Except (B1_12=1) => A3_TipoFlia=1 (7) [1]
Except (B2_17=1) Y (E5_GuardiAdult=1) => A3_TipoFlia=1 (5) [1]
Except (F5=3) Y (B2_9=5) Y (D1_A=4) => A3_TipoFlia=1 (12) [3]
Except (E5_GuardNiños=1) => A3_TipoFlia=2 (27) [5]
Except (E6=1) Y (B1_13=0) => A3_TipoFlia=4 (39) [14]
Except (B3=1) Y (E4_6=1) => A3_TipoFlia=1 (87) [27]
Except (A4_NroPersonas=6) Y (F9_4=0) => A3_TipoFlia=2 (5) [1]
Except (A4_NroPersonas=7) => A3_TipoFlia=2 (3) [1]
Except (E4_1=1) => A3_TipoFlia=1 (89) [36]
Except (D1_A=4) Y (A4_NroPersonas=7) Y (A2_TipoEncuest=3) => A3_TipoFlia=2 (6) [1]
Except (D1_A=4) Y (F9_10=1) Y (H1=2) Y (B2_11=5) => A3_TipoFlia=2 (8) [4]
Except (B3=1) => A3_TipoFlia=1 (58) [31]
Except (E3_3=1) => A3_TipoFlia=1 (33) [16]

Except (B4=2) Y (B1_10=0) Y (E4_4=0) => A3_TipoFlia=1 (17) [5]
Except (B2_4=3) Y (D3=1) Y (F8_8=1) => A3_TipoFlia=1 (14) [4]
Except (B1_11=1) Y (E4_2=1) => A3_TipoFlia=3 (8) [2]
Except (E4_3=0) Y (F8_5=0) => A3_TipoFlia=1 (58) [28]
Except (E4_1=1) => A3_TipoFlia=3 (17) [9]
Except (B2_6=5) => A3_TipoFlia=1 (134) [55]
Except (D1_A=4) Y (E5_GuardNiños=3) Y (F8_1=0) => A3_TipoFlia=2 (9) [1]
Except (E4_3=0) => A3_TipoFlia=1 (61) [24]
Except (D1_A=4) Y (B1_12=0) Y (F6=4) Y (E3_2=0) => A3_TipoFlia=2 (9) [2]
Except (B1_11=1) => A3_TipoFlia=4 (9) [4]
Except (B3=1) => A3_TipoFlia=1 (127) [69]
Except (F5=3) => A3_TipoFlia=2 (34) [15]
Except (E7=4) Y (E5_GuardNiños=0) => A3_TipoFlia=1 (15) [9]
Except (D1_A=3) => A3_TipoFlia=1 (12) [4]
Except (F9_7=0) Y (E1ServicSalud=1) Y (E4_10=0) => A3_TipoFlia=1 (14) [2]
Except (E5_GuardNiños=0) => A3_TipoFlia=1 (14) [5]
Except (B2_3=1) => A3_TipoFlia=1 (6) [2]
Except (E1ServicSalud=1) => A3_TipoFlia=1 (27) [14]
Except (B4=2) => A3_TipoFlia=1 (13) [3]
Except (E5_GuardNiños=3) Y (B1_10=0) => A3_TipoFlia=2 (4) [1]
Except (Cluster=5) Y (A3_B_TipoVivienda=1) Y (E3_3=0) => A3_TipoFlia=2 (6) [2]
Except (F8_2=1) Y (E3_1=1) => A3_TipoFlia=1 (6) [1]
Except (E4_4=1) Y (A2_TipoEncuest=3) => A3_TipoFlia=2 (2) [3]

Estas reglas expresan la cantidad de familias que pertenecen a una tipología determinada, a partir del cumplimiento de determinadas condiciones. Estas condiciones son dadas en forma de excepciones y se corresponden con las respuestas dadas a las preguntas presentes en el cuestionario.

DISCUSIÓN

La tabla Nº 1 muestra que fueron clasificadas correctamente el 70.83% de las instancias. En relación a los restantes parámetros de exactitud del modelo, se tiene que TP Rate=1, significa que acierta para el 100% de los atributos de la Clase 1 que corresponde con la Familia Tipo, es la clase mayoritaria en la base de datos. Para las clases 2 Familia Parental, 3 Familia Ampliada, 4 Familia Unipersonal y 5 Familia Plurinuclear, la proporción de positivos verdaderos es 0%. Este resultado es lógico, por cuanto el algoritmo ZeroR sólo acierta para la clase mayoritaria que en la Onda 2009 es la Familia Tipo.

FP Rate=1. El 100% de los casos falsos, clasificados como Clase 1, Familia Tipo, en realidad pertenecen a las otras clases. La Precision=0.708 de pureza de los valores significa que el 70.8% del total de mediciones predichas, sean verdaderas o falsas, están correctamente clasificadas. En relación con la Cobertura, el 100% de las instancias de la Clase 1 están correctamente clasificadas como pertenecientes a esta clase, respecto al total de términos reales. El parámetro F-Measure evalúa al algoritmo de clasificación ZeroR con una bondad del 58.70%.

Tal y como muestra la tabla Nº 2, al aplicar el algoritmo J48 fueron clasificadas correctamente el 79.81 de las instancias. El TP Rate obtiene un 95.30% de aciertos positivos para la Clase 1; para la Clase 2 es el 16.80% y en la Clase 4, el 100%; mientras que las Clases 3 y 5 tienen un 0% de aciertos. El FP Rate, para la Clase 1 es de 58.50% de falsos positivos, un 47.10% para la Clase 2 y nuevamente el 100% de casos falsos para la clase 4. También los falsos positivos

son de 0% para las clases 3 y 5. La precisión clasifica correctamente el 79.80% de verdaderos o falsos en la clase 1. Este parámetro se comporta con un 47.10% en la Clase 2, un 100% en la clase 4 y para las Clases 3 y 5 es el 0%. En cuanto a la Cobertura, el 100% de las instancias clasificadas que forman parte de la Clase 1, efectivamente corresponden a esta clase, y para la Clase 2 la cobertura es de 16.80%, en tanto que para la Clase 4 es del 100%; sigue siendo del 0% para las Clases 3 y 5. El valor de la bondad del algoritmo J48, de acuerdo con el parámetro F-Measure, es del 75.00%.

El comportamiento de los parámetros al aplicar el algoritmo Ridor, tabla Nº 3, indican que el parámetro TP Rate, las Clases 1 y 4 fueron evaluadas correctamente para el 100% de los casos. En las Clases 2, 3 y 5 los verdaderos positivos fueron de un 0%. En el caso de FP Rate, la Clase 1 tiene un 58.50% de falsos positivos, un 47.10% para la Clase 2 y nuevamente el 100% de casos falsos para la clase 4. También los falsos positivos son de un 0% para las clases 3 y 5. En relación a la Precisión, en la Clase 1 están correctamente clasificadas para verdaderas o falsas, el 79.80%. Este parámetro se comporta con un 47.10% en la Clase 2, un 100% en la clase 4 y sigue siendo 0% para las Clases 3 y 5. La Cobertura es del 100% de las instancias clasificadas para la Clase 1, por lo que todas las instancias clasificadas para esta clase efectivamente corresponden a ella. Para la Clase 2 es el 16.80%, y el 100% para la Clase 4 y sigue siendo del 0% para las Clases 3 y 5. La bondad total del algoritmo Ridor es de un 71.20% según el parámetro F-Measure.

Tabla Nº4. Principales Resultados de la Aplicación de Algoritmos en la Onda 2009.

Algoritmos	Instancias Bien Clasificadas %	Instancias Mal Clasificadas %	Precision %	F-Measure %
ZeroR	70.83	29.17	50.20	58.70
J48	79.81	20.18	73.90	75.00
Ridor	80.00	20.00	64.40	71.20

La comparación de los tres mejores algoritmos aplicados a la Onda 2009 del Observatorio Social de la Universidad Nacional del Litoral, tabla Nº 4, muestra que en cuanto al porcentaje de instancias correctamente clasificadas es el algoritmo de Reglas Ridor el que mejores resultados presenta; la diferencia entre éste y el algoritmo J48 de los Árboles de Decisión es mínima. Sin embargo, el algoritmo J48 supera ampliamente al ZeroR y al Ridor en cuanto a la precisión con un 73.90%. La media armónica ponderada de precisión y exhaustividad F-Measure que mejor comportamiento presenta corresponde también al algoritmo J48.

CONCLUSIONES

La Minería de Datos es una tecnología innovadora, que resulta en un buen punto de encuentro entre los investigadores. En este trabajo se han aplicado las técnicas y algoritmos de clasificación contenidos en Weka, software de minería de datos, aplicados a los datos generados por la Encuesta Permanente de Hogares 2009, del Observatorio Social de la Universidad Nacional del Litoral, y se determinó que el algoritmo de clasificación J48 de la familia de Árboles de Decisión es el que mejor clasifica la Onda 2009.

Este resultado es útil porque da información sobre la distribución del número de personas y los tipos de familia en función del atributo de clase. También resulta un nuevo conocimiento, derivado de aplicar dicho algoritmo, la relación existente entre la composición por integrantes de las familias y su vinculación con secciones específicas en el cuestionario. Tanto la aplicación del algoritmo J48, como las reglas obtenidas con el algoritmo Ridor, permiten profundizar en las características de la familia en la región de Santa Fe-Santo Tomé.

A pesar de que el proceso de Minería de Datos a la Onda 2009, presente en este trabajo, solo involucró el proceso de clasificación, es un punto de partida en la aplicación de las técnicas de Minería de Datos a la Encuesta Permanente de Hogares, y es indispensable e interesante realizar nuevos procesamientos de esta información empleando para ello las técnicas y algoritmos de Agrupamiento, Asociación y Selección de Atributos; procesamientos en los que se está trabajando actualmente, los cuales están vinculados además con la Encuesta Permanente de Hogares realizada por el Observatorio Social de la Universidad Nacional del Litoral, en el año 2010.

La aplicación de estas técnicas y los resultados obtenidos permiten tener nuevos elementos sobre las características de los santafesinos y, al mismo tiempo, evaluarlos y tener criterios sobre el análisis que esta sociedad haga de los temas que en ella incidan y son de su interés. Con ello, la Minería de Datos puede favorecer y contribuir con los procesos de toma de decisiones, lo que es muy importante en función de predecir futuros comportamientos en los nuevos relevamientos que se realicen.

BIBLIOGRAFÍA

Cabrera, I.M. (2004). Comportamiento de las nuevas tecnologías de la información y su impacto en el trabajo bibliotecario. En: 1er. Foro social de información, documentación y bibliotecas (pp. 1). Buenos Aires: Argentina.

Cisnero, I., García, C. & Lozano, I.M. (2002). ¿Sociedad de la Información ⇔ Sociedad del Conocimiento? La educación como mediadora. Disponible en: http://tecnologiaedu.us.es/edutec/paginas/43.html (Consultado 14 de Noviembre de 2010).

Fuertes, P. (2007). Observatorios Socio Económicos Laborales - Estudio de Sistematización. Ministerio de Trabajo y Promoción del Empleo - Programa de Estadísticas y Estudios Laborales, Agencia Suiza para el Desarrollo y la Cooperación – COSUDE, Centro de Servicios para la Capacitación Laboral y el Desarrollo – CAPLAB. (pp. 5). Lima: Perú.

Hernández, J., Ramírez, Mª.J. & Ferri, C. (2004). ¿Qué es la minería de datos? In: Introducción a la Minería de Datos. (1ra. Ed. pp. 25-283). Madrid, España.

Merlino, H., Britos, P., Ierache, J., Diez, E. & García-Martínez, R. (2005). Un Método de Transformación de Datos Orientado al Uso de Explotación de Información. En: II Workshop de Ingeniería del Software y Bases de Datos XI Congreso Argentino de Ciencias de la Computación. (pp. 22-32). Entre Ríos: Argentina.

Weka Machine Learning Software in Java. 2010. Projects. Disponible en: http://sourceforge.net/projects/weka/ (Consultado 14 de Junio de 2010).

Witten, I.H. & Frank, E. 2000. Data Mining. Practical Machine Learning Tools and Techniques with Java Implementation. Ed. Morgan Kaufmann Publisher. San Francisco CA, USA p. 7.