Análisis del diagnóstico de enfermedades mentales en la ciudad de Bogotá, mediante técnicas de minería de datos

Analysis of the diagnosis of mental illness in the city of Bogotá, using data mining techniques

Cindy Nayid Vega Santamaría
Ingeniera Industrial
Corporación Tecnológica Industrial Colombiana TEINCO
cindy.vega@teinco.edu.co
Bogotá-Colombia

Aneider Sahedy Angulo
Ingeniero en Control
Universidad Distrital Francisco José de Caldas
aneider.anguloa@gmail.com
Bogotá-Colombia

Pedro Alfonso Mariño
Estadístico
Universidad Manuela Beltrán
pedro.marino@docentes.umb.edu.co
Bogotá-Colombia

Resumen

El presente artículo propone un análisis de la aplicación de una técnica de minería de datos para clasificar las localidades de la ciudad de Bogotá, donde se presenta con mayor incidencia los trastornos mentales en sus habitantes. Se utilizó la información de la encuesta multipropósito del DANE, y se empleó métodos de clasificación. La información correspondiente de la base de datos extraída, permitió hacer un primer filtro a la clasificación, en esta se tomaron solamente las variables significativas para el objeto de estudio. El modelo presenta la localidad donde la incidencia en el diagnóstico de trastornos mentales es más significativa en sus habitantes. Consecuentemente se evidenciara el proceso analítico y se expondrán los niveles de incidencia de trastornos mentales de acuerdo al resultado obtenido en la simulación.

Palabras claves: Enfermedades mentales; minería de datos; población y trastornos mentales.

Abstract

The present article proposes an analysis of the application of a data mining technique to classify the localities of the city of Bogotá, where the mental disorders in their inhabitants present themselves with greater incidence. Data from the DANE multipurpose survey was used, and classification methods were used. The corresponding information of the extracted database allowed to make a first filter to the classification, in this only the variables were considered significant for the object of study. The model presents the locality where the incidence in the diagnosis of mental disorders is more significant in its inhabitants. Consequently the analytical process will be evidenced and the levels of incidence of mental disorders will be exposed according to the result obtained in the simulation.

Keywords: Mental Illnesses; data Mining; population and mental disorders.

Revista Mundo Fesc, 13, Enero- Junio 2017.
ISSN (Printed) 2216-0353, ISSN (Online) 2216-0388

Forma de citar: Vega, C.N., Angulo, A.S., Mariño, P.A. (2017). Análisis del diagnóstico de enfermedades mentales en la ciudad de Bogotá, mediante técnicas de minería de datos. Mundo Fesc, 13, 35-47
Recibido: *2 Octubre de 2016.*
Aceptado: *12 Diciembre de 2016.*

1. Introducción

En Bogotá, surgen necesidades de diferente índole, para sus habitantes y una de las principales labores del distrito es identificar estas problemáticas para cada localidad, para ello se cuenta con diferentes entidades, que permitan identificar aspectos de relevancia en la cotidianidad de los habitantes de la ciudad de Bogotá. Para ello se pretende visualizar estos aspectos mediante revisiones de encuestas realizadas a las personas por el Departamento Administrativo Nacional de Estadística o DANE para el año 2015. El problema es encontrar información útil, orientada a este análisis en estos grandes volúmenes de información. La solución al alcance de hacer fácil esta tarea de extraer el conocimiento es con la minería de datos. La técnica de clasificación o segmentación de clientes permite hacer una aplicación práctica a la mira del estudio. En este trabajo se propone la aplicación de técnicas de minería de datos que "surgen como las mejores herramientas para realizar exploraciones más profundas y extraer información nueva, útil y no trivial que se encuentra oculta en grandes volúmenes de información" (Moreno, Miguel, Garcia , & Polo, 2001, pág. 2); de igual manera (Marulanda, Lopez , & Mejia , 2017), define la Minería de Datos (MD) como el proceso de descubrir conocimiento útil y entendible, desde grandes bases de datos almacenados en distintos formatos, por medio de modelos inteligibles a partir de los datos. La detención de variables que pueden estar involucradas en el diagnostico de enfermedades o trastornos mentales de los habitantes de una localidad en la ciudad de Bogotá, los datos pueden mostrar una exagerada dimensionadad y dentro del objeto de estudio no pueden estar registradas otras características que puedan influir en la clasificación. El objetivo final es realizar una clasificación de las localidades donde se presenta con mayor incidencia los trastornos de este tipo.

A partir de la revisión de condiciones de los habitantes de Bogotá, el cual es una realidad latente en el día a día de la ciudad, que se olvidó del tiempo libre, del ocio sano y la familia, que sucumbe a la violencia o a la tristeza absoluta de una sociedad que tacha de desadaptados a personas normales que un día se enfermaron y se estrellaron con las barreras del acceso para recibir un tratamiento adecuado (Malaver, 2014).

Se propone un estudio en el cual se identifiquen las variables que afecta al grupo de habitantes de las diferentes localidades, cuyos aspectos económicos, de salud, de vivienda o temáticas sociales, pueden influir en el diagnóstico de enfermedades o trastornos mentales, la información que se toma de la encuesta Multipropósito para Bogotá y sus municipios aledaños, constituye la materia prima del análisis.

Este artículo está organizado de la siguiente forma: En la siguiente sección se hace una descripción general del método que se tendría en cuenta para la clasificación. En la sección 3, se describe como se realizó la recopilación de datos y las fuentes de información que se tuvieron en cuenta para el desarrollo. La sección 4 explica las tareas de pre-procesado de los datos. La sección 5, se realiza un análisis preliminar de la variable objeto de estudio respecto a otras variables particularmente asociadas al diagnóstico del trastorno mental, la sección 6 describe los diferentes ejercicios de minería de datos que se han trabajado y los resultados obtenidos de las mismas. En la sección 7, se hace una interpretación de los resultados y finalmente se describen las conclusiones del trabajo realizado.

2. Método

El método usado en el presente estudio, está relacionado con las técnicas de minería de datos, que se muestran en la figura 1.

2.1 Recopilación de datos: En esta fase se debe tener en cuenta la descripción del objeto de estudio a solucionar por la técnica de minería de da-

tos, seguido de esto se debe realizar una revisión de fuentes primarias en la cual se logre conseguir la base de datos necesaria para la evaluación de registros de datos, entre ellos datos anómalos, atípicos y en ultimas la ausencia de registros.

Figura 1. Proceso efectuado en minería de datos. Fuente: (Marquez, Romero , & Ventura , 2012). Adaptado por autores.

2.2 Pre-procesado de datos: En esta fase se debe evaluar el algoritmo que permita identificar datos faltantes o anómalos para que en el siguiente paso, se realice la minería sin contratiempos, con el fin de preservar la información relevante para una clasificación eficiente, para lo cual se verificaron los campos para eliminar los que no contenían información que permitieran realizar el proceso de predicción (Mosquera, Parra, & Cartrillon, 2016).

2.3 Minería de datos: En esta fase se debe identificar el algoritmo y la herramienta a utilizar para evaluar la clasificación de los datos con respecto al objeto de estudio planteado.

2.4 Interpretación de datos: En esta fase, se debe evaluar todo el proceso realizado con el fin de identificar el comportamiento de las variables planteadas y la solución de la clasificación por medio de técnicas de minería de datos.

3. Recopilación de Datos

En esta etapa se compila la información disponible de la encuesta Multipropósito del Departamento Administrativo Nacional de Estadística (Departamento Administrativo Nacional de Estadística, 2015), después de un minucioso análisis del objeto de estudio se selecciona las bases de datos correspondientes a la encuesta que contienen las variables relevantes al estudio, este conjunto de datos están categorizados en capítulos, donde se seleccionaron las bases de datos incluidas en cada capítulo de la encuesta como se muestra en la tabla 1.

Tabla 1

Capítulos seleccionados de DANE, multipropósito 2014

Capítulos y variables seleccionadas		
Capítulo	Descripción	Variables
A	Contiene datos de identificación	6
C	Contiene datos de condiciones habitacionales	10
E	Contiene datos de composición del hogar	8
L	Contiene datos de condiciones de vida	6
F	Contiene datos de condiciones de salud	21

Fuente: Autores.

Posterior a esta selección se extrajeron específicamente las variables que corresponden al objeto de estudio, información que finalmente permite conformar el conjunto de datos de análisis, algunas de las variables definidas que se presentan en la tabla 2.

4. Pre-procesado de datos

En esta fase se prepara el conjunto de datos desarrollando algunas tareas típicas que permitan transformar los datos originales a una forma adecuada para ser usados posteriormente en las técnicas definidas en minería de datos.

Tabla 2

Variables seleccionadas en el estudio

Variable	Categorías	Medición
Estrato	Socioeconómico imputado	Ordinal
Sexo	1. Hombre 2. Mujer 3. Inter	Categórica
Estado civil	Casado(a), unión libre	Categórica
Afiliación a SS	1. Si 2. NO	Categórica
Estado actual General	1. Mb, 2B, 3R 4M 5Mm	Categórica
Ha diagnosticado problemas de salud	1. Si 2. NO	Categórica
Ha diagnosticado depresión, ansiedad	1. Si 2. NO	Categórica
Ha diagnosticado hiperactividad	1. Si 2. NO	Categórica
Recibe atención medica periódica	1. Si 2. NO	Categórica
Tiene alguna limitación permanente	1. Si 2. NO	Categórica
Fue hospitalizado en los últimos meses	1. Si 2. NO	Categórica
Cuanto gasto en hogar en medicamentos	Rangos	Numérica
Acudió a un servicio de urgencias	1. Si 2. NO	Categórica

Fuente: Autores.

El estudio de estas actividades se ha definido en integración de variables, que consiste en la compilación de los datos de la encuesta en un solo conjunto de datos.

Limpieza de datos: que consiste en extraer únicamente la información de las localidades de Bogotá, excluyendo las áreas y zonas fuera del perímetro.

5. Análisis preliminar de la variable diagnóstico de trastorno mental

En este apartado se realiza un análisis del comportamiento de la variable respuesta presencia del trastorno mental de los pacientes, comparando su comportamiento con otras variables del estudio, acciones que permite ver comportamientos de la variable que serán definitivos a la hora de seleccionar las variables que deben incluirse en el proceso de minería.

En la figura 2, se muestra cómo se comporta la variable objeto de estudio respecto al género del paciente y su estado de salud, se puede observar que a pesar de que la persona presentan estado de salud bueno o estable muestran un trastorno mental latente en ambos géneros.

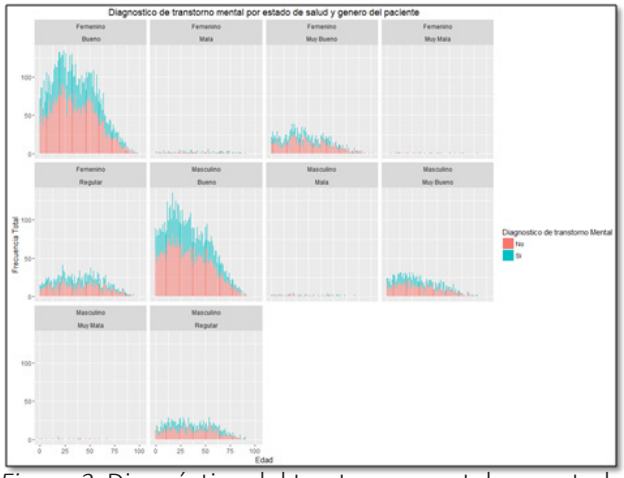

Figura 2. Diagnóstico del trastorno mental por estado de salud y género del paciente
Fuente: Autores (realizado en WEKA).

En la figura 3, se muestra la variable objeto de estudio respecto a la presencia o ausencia de tratamiento en el diagnóstico del trastorno mental por género del paciente, las Gráficas permiten concluir que en ambos géneros se presentan pacientes que aunque están diagnosticados con algún tipo de enfermedad o trastorno mental, no se encuentran en tratamiento, razones que pueden llegar a incrementar su enfermedad.

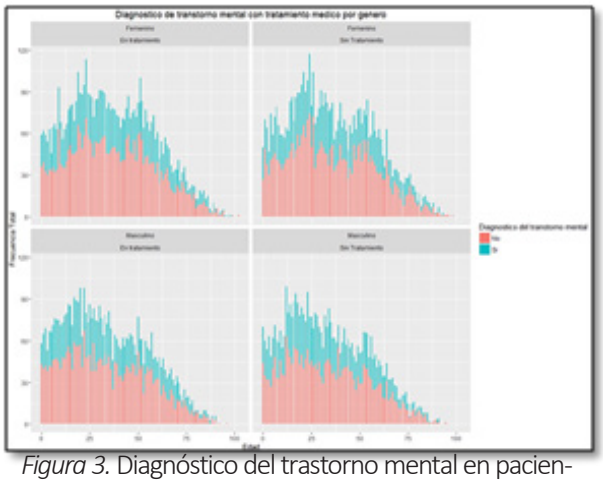

Figura 3. Diagnóstico del trastorno mental en pacientes con tratamiento por género del paciente
Fuente: Autores (realizado en WEKA).

En la figura 4, se visualiza como está distribuida la variable diagnóstico del trastorno mental en la ciudad de Bogotá, por estrato socioeconómico y género, las Gráficas permiten concluir que en

los estratos socioeconómicos medios y bajos, hay una presencia marcada de algún tipo de enfermedad o trastorno mental, ambos géneros presentan esta tendencia, razones que pueden ser objeto de estudios posteriores en cada una de las localidades de Bogotá.

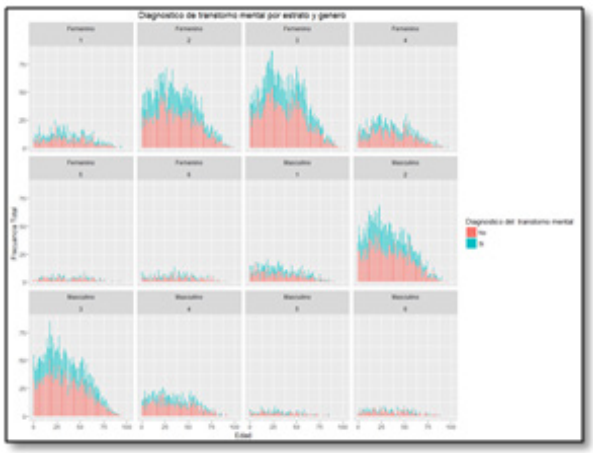

Figura 4. Diagnóstico del trastorno mental en pacientes por género y estrato

Fuente: Autores (realizado en WEKA).

A. Algoritmo KNN para imputación de valores faltantes

El método KNN pertenece al grupo de métodos para tareas de clasificación de datos que se pueden encontrar dentro de minería de datos, estos son fundamentalmente dependientes de la distancia y en consecuencia poseen características propias; como la cercanía, la lejanía y la magnitud de longitud, entre otras (Rodriguez, Rojas, & Franco , 2012).

El objetivo de la clasificación es encontrar un modelo, para predecir la clase a la que pertenecería cada registro, esta asignación es una clase que se debe hacer con la mayor precisión posible.
Por lo general, el conjunto de datos, se divide en dos conjuntos al azar, uno para entrenamiento y el otro de prueba, Un conjunto de prueba (tabla de testing) se utiliza para determinar la precisión del modelo.

B. Enlace simple o del vecino más cercano

Una característica de estos métodos es que buscan una matriz de similaridades de tamaño (nxn), que secuencialmente, se mezclan los casos más cercanos; aunque cada uno tiene su propia forma de medir las distancias entre grupos o clases. Un segundo aspecto es que cada paso o etapa en la conformación de grupos puede presentarse visualmente por un gráfico denominado dendograma.

Se inicia con tantos grupos como instancias se tengan, se juntan los dos casos que estén a la menor distancia entre este y otro grupos, la distancia que se utiliza es la distancia Euclidea, según la siguiente ecuación (1):

$$\delta(x,y) = \sqrt{\sum_{i=1}^{p}(xi - yi)^2} \qquad (1)$$

El método utilizado en este estudio, usa la medida de los 10 casos más similares dentro de cada variable, esta rellena con los valores calculados con la medida. El algoritmo implementado está disponible en (Zhang, 2012), este algoritmo usa una variación a la distancia euclidea y presenta la siguiente ecuación (2):

$$\delta(x,y) = \sqrt{\sum_{i=1}^{p}\delta i(xi - yi)^2} \qquad (2)$$

Donde Si determina la distancia entre los valores de las instancias y está dada por la ecuación 3:

$$\delta i = \begin{cases} 1 & Si\ i\ es\ nominal\ y\ v1 \neq v2 \\ 0 & Si\ i\ es\ nominal\ y\ v1 = v2 \\ (vi - v2)^2 & si\ i\ es\ numerica \end{cases} \qquad (3)$$

Una vez se aplica el algoritmo de imputación y limpieza de datos se obtiene una data disponible para realizar el análisis.

Minería y experimentación

En esta sección se describen los ejercicios realizados y las técnicas de minería de datos utilizados para la obtención de la clasificación de enfermedades de trastornos mentales en las diferentes localidades de la ciudad de Bogotá. Para este ejercicio se tuvo en cuenta la herramienta WEKA (Waikato Enviroment for Knowledge Analysis), que cuenta con las herramientas necesarias para realizar transformaciones sobre los datos, tareas de clasificación, regresión, clustering, asociación y visualización (Hall, Frank, Holmes, Pfahringer, Reutemann, & Witten, 2009). Para el ejercicio se tuvo en cuenta tres algoritmos, IBK, BAGGIN y REP Tree.

A. Algoritmo IBK

Este es un esquema de aprendizaje basado en la instancia más cercana. Los métodos de vecinos más cercanos tienen una historia (originada en el último decenio de 1950). Trabajan de tal manera que se almacena todo el conjunto de datos y se asignan nuevos elementos de datos a la misma clase que su" vecino más cercano". Solo se considera el mejor vecino o un número de vecinos más cercanos votan por la clasificación del nuevo elemento de datos.

El paquete Weka ofrece la posibilidad de detectar cuantos vecinos dan los mejores resultados mediante la evaluación cruzada. El algoritmo de vecino más cercano es tardío en el proceso, especialmente cuando se considera más de un vecino más cercano. También requiere memoria para almacenar las instancias. Existen varios métodos de optimización que apuntan a la cantidad de datos almacenados.

Se basa en la suposición de que los ejemplos cercanos pertenecen a la misma clase. Su fase de aprendizaje es muy simple, pues se limita a almacenar los ejemplos del conjunto de entrenamiento. La de clasificación también es sencilla, aunque dispendiosa en eficiencia: busca los k ejemplos más cercanos a la instancia que se quiere clasificar y le asigna la clase más frecuente entre ellos (Caises & Navarro, 2010).

El algoritmo IBK está basado en instancias, consiste en almacenar los datos presentados. Cuando una nueva instancia es encontrada, un conjunto de instancias similares relacionadas es devuelto desde la memoria y usado para clasificar la instancia consultada (Villena, Crespo, & Garcia, 2006).
De acuerdo a la clasificación el modelo IBK Clasificador basado en instancia utilizando un vecino más cercano, presenta los resultados de las tablas 3 y 4.

Tabla 3
Algoritmo IBK

Estimaciones para la Clasificación con IBK	
Coeficiente de correlación	1
Error absoluto medio	0
Error cuadrático medio raz	0
Error relativo absoluto	0
Error de cuadrado relativo de la raz	0
Número total de instancias	1666

Fuente: Autores.

Tabla 4
Estimación instancias agrupadas

Instancias Agrupadas		
0	712	43%
1	117	7%
2	42	3%
3	54	3%
4	24	1%
5	80	5%
6	236	14%
7	382	23%
8	19	1%

Fuente: Autores.

B. Algoritmo BAGGING

En la misma relación se encuentra el algoritmo BAGGING, por Booststrap Aggregating, fue intro-

ducido por Breiman en 1996, es un método de agregación de modelos homogéneos que se basa en el valor mayoritario o el promedio según el caso. El método consiste en hacer varias muestras del conjunto de datos iniciales y promediar las predicciones hechas por los distintos clasificadores. Los resultados que arroja se muestran en las tablas 5 y 6.

Tabla 5
Algoritmo Bagging

Estimaciones para la Clasificación con BAGGING	
Coeficiente de correlación	0.2422
Error absoluto medio	1.0342
Error cuadrático medio raz	1.1998
Error relativo absoluto	92.6117%
Error de cuadrado relativo de la raz	97.4832%
Número total de instancias	1666

Fuente: Autores.

Tabla 6
Estimación instancias agrupadas

Instancias Agrupadas		
0	671	45%
1	105	5%
2	46	6%
3	24	1%
4	68	2%
5	80	4%
6	266	8%
7	562	17%
8	19	3%

Fuente: Autores.

C. Algoritmo REPTree

Es un algoritmo de aprendizaje rápido mediante arboles de decisión (Nasa, 2012). Construye un árbol de decisión usando la información de varianza y lo poda usando como criterio la reducción del error. Solamente clasifica valores para atributos numéricos una vez. Los valores que faltan se obtienen partiendo las correspondientes instancias. Los resultados se muestran en las tablas 7 y 8 respectivamente.

Tabla 7
Algoritmo reptree

Estimaciones para la Clasificación con REPTree	
Coeficiente de correlación	0.4514
Error absoluto medio	0.9054
Error cuadrático medio raz	1.0977
Error relativo absoluto	81.1201%
Error de cuadrado relativo de la raz	89.233%
Número total de instancias	1666

Fuente: Autores.

Tabla 8
Estación instancias agrupadas

Instancias Agrupadas		
0	712	43%
1	117	7%
2	42	3%
3	54	3%
4	24	1%
5	80	5%
6	236	14%
7	382	23%
8	19	1%

Fuente: Autores.

D. Algoritmo ZeroR

ZeroR es el método de clasificación más simple, este se basa en el objetivo e ignora todos los predictores. El clasificador ZeroR predice simplemente la categoría mayoritaria (clase) (Vijayarani & Muthulakshmi, 2013). Aunque no hay poder de predictibilidad en ZeroR, es útil para determinar un desempeño como un punto de referencia para otros métodos de clasificación. Este Algoritmo es el más primitivo en Weka. Modela el conjunto de datos con una sola regla. Dado un nuevo elemento de datos para la clasificación, ZeroR siempre predice el valor de categoría más frecuente en los datos de entrenamiento para problemas con un valor de clase nominal o el valor de clase promedio para problemas de predicción numérica. En algunos conjuntos de datos es posible que otros esquemas de aprendizaje induzcan a modelos que presentan peores resultados en los

nuevos datos que ZeroR, que es un claro indicador de sobre ejecución grave.

E. Matriz de confusión para los algoritmos evaluados por combinación de variables.

La Matriz de Confusión contiene información acerca de las predicciones realizadas por la Clasificación, en la figura 5, se comparan el conjunto de individuos en la tabla de aprendizaje o de testing, la predicción dada versus la clase a la que estos realmente pertenecen. Esta matriz de confusión está relacionada directamente con la variable objeto de estudio enfermedad de trastorno mental, evaluado bajo cuatro algoritmos Zero R, REPTree, IBK y Bagging. De esta manera se permite visualizar mediante una tabla de contingencia la distribución de errores cometidos por un clasificador.

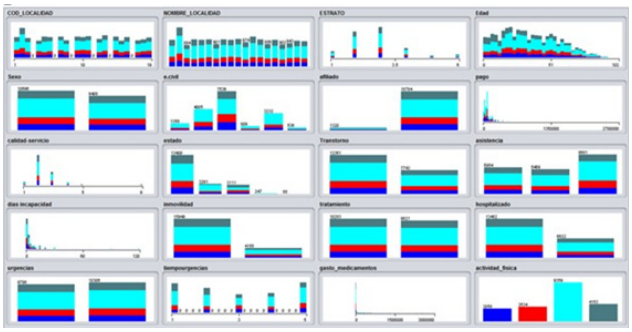

Figura 5. Gráfico de representación de variables por WEKA.
Fuente: Autores (realizado en WEKA).

En las siguientes tablas 9 y 10 se indicaran los factores de determinación de la clasificación de la matriz de confusión por algoritmo analizado en este trabajo, esto medido bajo todas las variables previstas en el inicio del estudio.

Tabla 9
Variable trastorno en pacientes con diagnostico por trastornos mentales

ZERO R			REPTree		IBK		Bagging	
FACTOR	**INSTANCIA**	**%**	**INSTANCIA**	**%**	**INSTANCIA**	**%**	**INSTANCIA**	**%**
Instancias correctamente clasificadas	12361	61%	13642	67,85%	19591	97,44%	15885	79,01%
Instancias incorrectamente clasificadas	7743	39%	6462	32,14%	513	2,55%	4219	20,98%
Estadísticas Kappa	0		0,2421		0,9461		0,5232	
Error absoluto medio	0,4736		0,4183		0,0256		0,3718	
Error medio cuadrático	0,4866		0,456		0,1593		0,4005	
Error relativo absoluto	100%		88,31%		5,39%		78,49%	
Error relativo cuadrado	100%		93,69%		32,74%		82,30%	
Número total de instancias	20104		20104		20104		20104	

Fuente: Autores

Tabla 10
Variable afiliación a sistema de seguridad y salud

ZERO R			REPTree		IBK		Bagging	
FACTOR	**INSTANCIA**	**%**	**INSTANCIA**	**%**	**INSTANCIA**	**%**	**INSTANCIA**	**%**
Instancias correctamente clasificadas	18784	93%	18807	93,54%	18807	93,54%	18794	93,48%
Instancias incorrectamente clasificadas	1230	65%	1237	6,45%	1297	6,45%	1310	651,00%
Estadísticas Kappa	0		0,0371		0,0371		0,0141	
Error absoluto medio	0,1227		0,1182		0,1182		0,1086	
Error medio cuadrático	0,2477		0,2428		0,2428		0,2265	
Error relativo absoluto	100%		96,30%		96,29%		88,46%	
Error relativo cuadrado	100%		98,05%		98,46%		91,45%	
Número total de instancias	20104		20104		20104		20104	

Fuente: Autores

Como se indica en la tabla 9, la variable trastorno (objeto de estudio), desarrollada bajo los cuatro algoritmos, evidencia que además se puede deducir que si el número de instancias correctas sube, entonces el error absoluto disminuye. Por ende en esta primera variable se puede deducir que el algoritmo que determina la mejor clasificación es el IBK, ya que es el que cuenta con más instancias clasificadas y el error absoluto es menor a la de los demás algoritmos.

Para el caso de la variable de afiliación representando sus resultados en la Tabla 10, se puede determinar que todos los algoritmos procesados cuentan con valores muy similares, es decir que el factor de elección del mejor algoritmo no es tan relevante como en otras variables, en este caso dos algoritmos muestran instancias iguales que permiten seleccionar entre estos dos la clasificación de la variable.

La variable estado de salud, representada en la tabla 11, indica que la de más instancias correctamente clasificadas está en el algoritmo IBK, en esta se encuentra el error absoluto mínimo que permite determinar el algoritmo de mejor clasificación para la variable estado de salud del paciente diagnosticado con trastorno mental.

La variable asistencia médica por trastornos mentales, el cálculo de los algoritmos para esta variable se muestran en la tabla 12, se puede observar que se encuentra mejor clasificada en el algoritmo IBK, ya que en este contiene las mayores instancias correctamente clasificadas y el mínimo error absoluto.

En la tabla 13 (ver pág. 44), se muestran los resultados para la variable movilidad se identifica como mejor algoritmo clasificador al IBK, ya que en este se concentra las mayores instancias correctamente clasificadas y el menor error absoluto.

Tabla 11
Variable estado de salud en pacientes con diagnostico por trastornos mentales

ZERO R			REPTree		IBK		Bagging	
FACTOR	**INSTANCIA**	**%**	**INSTANCIA**	**%**	**INSTANCIA**	**%**	**INSTANCIA**	**%**
Instancias correctamente clasificadas	13400	666534%	14264	70,95%	19420	96,59%	14939	74,30%
Instancias incorrectamente clasificadas	6764	33466%	5840	29,04%	684	3,40%	5165	25,69%
Estadísticas Kappa	0		0,2371		0,933		0,3396	
Error absoluto medio	0,202		0,1727		0,0137		0,1559	
Error medio cuadrático	0,3178		0,2931		0,1166		0,265	
Error relativo absoluto	100%		85,38%		6,77%		77,21%	
Error relativo cuadrado	100%		92,25%		36,69%		83,40%	
Número total de instancias	20104		20104		20104		20104	

Fuente: Autores

Tabla 12
Variable asistencia médica por trastornos mentales

ZERO R			REPTree		IBK		Bagging	
FACTOR	**INSTANCIA**	**%**	**INSTANCIA**	**%**	**INSTANCIA**	**%**	**INSTANCIA**	**%**
Instancias correctamente clasificadas	8671	43,18%	11125	55,33%	19077	94,89%	14102	70,14%
Instancias incorrectamente clasificadas	11423	56,81%	8979	44,66%	1027	5,10%	6002	29,85%
Estadísticas Kappa	0		0,2873		0,9216		0,5292	
Error absoluto medio	0,4346		0,3718		0,0341		0,3446	
Error medio cuadrático	0,4661		0,4294		0,1842		0,3914	
Error relativo absoluto	100%		85,56%		7,85%		79,31%	
Error relativo cuadrado	100%		92,11%		39,51%		83,96%	
Número total de instancias	20104		20104		20104		20104	

Fuente: Autores

En la tabla 14, se indica la variable tratamiento en pacientes con diagnostico por trastorno mental, se identifica como algún proceso que se ejerce para una enfermedad de algún tipo mental, esta variable medida bajo el algoritmo de clasificación permite identificar que el IBK, es el algoritmo que cuenta con el mejor clasificador de instancias seleccionadas bajo el mínimo error absoluto.

Los resultados mostrados para la variable hospitalización por trastorno mental diagnosticado se muestran en la tabla 15, las instancias mejor clasificadas están bajo el algoritmo IBK, ya que existe una precisión muy alta, para este algoritmo es de 97.36 % con un error absoluto mínimo. En el estudio también se tuvo en cuenta la variable atención.

Los resultados mostrados en la tabla 16, fue tomada bajo los cuatro algoritmos, dando como resultado que el algoritmo IBK, ya que obtuvo las mejores instancias clasificadas y el error absoluto mínimo.

La tabla 17 muestra los resultados de la variable actividad física, la cual evalúa, aquellos usuarios que indican su participación en algún tipo de actividad física, se estudió, bajo las cuatro variables, donde el algoritmo IBK, clasificó mejor las instancias y por ende obtuvo el menor error absoluto.

Como se evidencia en las tablas 15, 16 y 17 (ver pág. 45) el algoritmo que logra clasificar mejor las instancias en la evaluación de todas las variables, es el algoritmo IBK, esto también se ve reflejado en las matrices de confusión expresadas por todos los algoritmos tenidos en cuenta en el presente estudio, en la matriz de confusión permite identificar las proporciones de distribución dentro de la matriz.

Tabla 13
Variable inmovilidad o lesiones que afectan la movilidad en pacientes con diagnóstico por trastornos mentales

ZERO R			REPTree		IBK		Bagging	
FACTOR	**INSTANCIA**	**%**	**INSTANCIA**	**%**	**INSTANCIA**	**%**	**INSTANCIA**	**%**
Instancias correctamente clasificadas	15949	79,33%	16212	10,94%	19766	98,31%	16519	82,16%
Instancias incorrectamente clasificadas	4155	20,66%	1892	19,35%	338	1,68%	3585	17,83%
Estadísticas Kappa	0		0,1245		0,9486		0,2097	
Error absoluto medio	0,3279		0,2976		0,0169		0,266	
Error medio cuadrático	0,4049		0,385		0,1296		0,3451	
Error relativo absoluto	100%		90,76%		51,50%		81,11%	
Error relativo cuadrado	100%		95,00%		31,99%		85,22%	
Número total de instancias	20104		20104		20104		20104	

Fuente: Autores

Tabla 14
Variable seguimiento en el tratamiento en pacientes con diagnóstico por trastornos mentales

ZERO R			REPTree		IBK		Bagging	
FACTOR	**INSTANCIA**	**%**	**INSTANCIA**	**%**	**INSTANCIA**	**%**	**INSTANCIA**	**%**
Instancias correctamente clasificadas	10283	51,14%	13092	65,12%	19458	96,78%	16344	81,29%
Instancias incorrectamente clasificadas	9821	48,85%	7012	34,87%	686	3,21%	3760	18,70%
Estadísticas Kappa	0		0,2993		0,9357		0,6254	
Error absoluto medio	0,4997		0,4278		0,0322		0,3921	
Error medio cuadrático	0,499		0,4312		0,1791		0,41	
Error relativo absoluto	100%		85,60%		6,44%		78,45%	
Error relativo cuadrado	100%		92,27%		35,83%		82,01%	
Número total de instancias	20104		20104		20104		20104	

Fuente: Autores

Tabla 15

Variable hospitalizaciones en pacientes con diagnóstico por trastornos mentales

ZERO R			REPTree		IBK		Bagging	
FACTOR	INSTANCIA	%	INSTANCIA	%	INSTANCIA	%	INSTANCIA	%
Instancias correctamente clasificadas	13482	67,06%	14074	70,01%	19575	97,36%	15507	77,13%
Instancias incorrectamente clasificadas	6622	32,93%	5030	29,99%	529	2,63%	4597	22,86%
Estadísticas Kappa	0		0,1471		0,9404		0,3843	
Error absoluto medio	0,4418		0,405		0,0264		0,3584	
Error medio cuadrático	0,47		0,4489		0,1618		0,396	
Error relativo absoluto	100%		91,07%		5,90%		81,13%	
Error relativo cuadrado	100%		95,51%		34,43%		84,25%	
Número total de instancias	20104		20104		20104		20104	

Fuente: Autores

Tabla 16

Variable atención por urgencias en pacientes con diagnóstico por trastornos mentales

ZERO R			REPTree		IBK		Bagging	
FACTOR	INSTANCIA	%	INSTANCIA	%	INSTANCIA	%	INSTANCIA	%
Instancias correctamente clasificadas	10306	51,26%	13502	67,16%	19495	96,97%	16135	80,25%
Instancias incorrectamente clasificadas	9798	48,73%	6602	32,83%	609	3,02%	3969	19,74%
Estadísticas Kappa	0		0,342		0,9394		0,6046	
Error absoluto medio	0,4997		0,418		0,0303		0,3829	
Error medio cuadrático	0,4998		0,4551		0,7138		0,4053	
Error relativo absoluto	100%		83,64%		6,07%		76,62%	
Error relativo cuadrado	100%		91,05%		34,77%		81,08%	
Número total de instancias	20104		20104		20104		20104	

Fuente: Autores

Tabla 17

Variable actividad física

ZERO R			REPTree		IBK		Bagging	
FACTOR	INSTANCIA	%	INSTANCIA	%	INSTANCIA	%	INSTANCIA	%
Instancias correctamente clasificadas	9359	46,55%	10702	53,23%	18305	91,05%	12985	64,58%
Instancias incorrectamente clasificadas	10745	53,44%	9402	43,76%	1799	8,94%	7119	35,41%
Estadísticas Kappa	0		0,2264		0,9685		0,4271	
Error absoluto medio	0,3433		0,2998		0,0448		0,2734	
Error medio cuadrático	0,4143		0,3846		0,2114		0,3493	
Error relativo absoluto	100%		87,34%		13,04%		79,63%	
Error relativo cuadrado	100%		92,83%		51,03%		84,32%	
Número total de instancias	20104		20104		20104		20104	

Fuente: Autores

7. Interpretación de resultados

El ejercicio analítico de los cuatro algoritmos en minería de datos, los valores de coeficiente de relación varían siendo el algoritmo ZeroR el de menor valor y el algoritmo IBK el que arroja mayor valor, se observaron en los cuatro que las instancias agrupadas tuvieron el mismo valor porcentual y los mismos conjuntos de agrupación o clúster. Se logró observar que los principales atributos tienen en cuenta personas con problemas mentales, ya sea por localidad, estado, movilidad

y hostilización, en caso real el sistema de salud actual, tienen muchas relevancias en estos atributos, ya que los usuarios de acuerdo al pago y servicio hospitalario, se sienten afectados o beneficiarios en la calidad de la atención de los centros médicos.

Adicionalmente se generaron unas conclusiones con respecto al proceso logrado, estas son:

1. Se ha demostrado que los algoritmos de clasificación pueden utilizarse con éxito para predecir los diferentes rendimientos de variables identificadas en el objeto de estudio planteado.

2. Se ha mostrado la utilidad de las técnicas de selección de características cuando se dispone de muchos atributos, como se evidencio en este caso.
3. El uso de herramientas permite identificar de forma más eficiente datos anómalos, lo cual indica que en la minería de datos se realizar más rápido, la identificación de información relevante para el objeto de estudio.

8. Conclusiones

En este trabajo, se presentaron un conjunto de ejercicios con el objetivo de realizar una clasificación de enfermedades mentales por localidad en la ciudad de Bogotá, mediante técnicas de minería de datos. Con respecto a la recolección de datos no fue un proceso de fácil, debido a que se encontraron valores anómalos, faltantes y que por ende la calidad y fiabilidad de la información afecta de manera directa en los resultados obtenidos. El presente estudio es un inicio de otros estudios ya que permite realizar una identificación de los algoritmos de clasificación dependiendo del objeto de estudio, pero para lograr esto se requiere recursos; tales como tiempo, recurso económico, tecnológico y de compromiso con el proceso.

La técnica de minería de datos aporta de manera eficiente al trabajo con volúmenes altos de datos y permite la presentación, tabulación y análisis de resultados; la herramienta usada para el proceso fue efectiva y de fácil acceso ya que es una plataforma Open Source y permite acceso a niveles ascendentes todo dependiendo del enfoque del estudio y la necesidad que requiera la utilización de la técnica de minería de datos.

9. Referencias

Caises, Y., & Navarro, R. (2010). Transformación de Características para la Minería de Datos. *Trimestral, XVI(2), 1-13.*

Departamento Administrativo Nacional de Estadística. (27 de Julio de 2015). *Encuesta multiproposito.* Recuperado de http://www.dane.gov.co/index.php/estadisticas-portema/salud/calidad-de-vida-ecv/encuesta-multiproposito

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. (2009). The weka data mining software: an update. *Explorations newsletter,* XI(1), 10-18.

Malaver, C. (20 de Octubre de 2014). Así es el drama de la enfermedad mental en Bogotá. *El tiempo*. Recuperado de http://www.eltiempo.com/archivo/documento/CMS-14715516

Marquez, C., Romero , C., & Ventura , S. (2012). Predicción del Fracaso Escolar mediante Técnicas de Minería de Datos. *IEEE-RITA, VII*(3), 109-117.

Marulanda, C., Lopez , M., & Mejia , M. (2017). Minería de datos en gestión del conocimiento de pymes de Colombia. *Virtual*(50), 224-237.

Moreno, M., Miguel, L., Garcia , F., & Polo, J. (2001). Aplicación de técnicas de minería de datos en la construcción y validación de modelos predictivos y asociativos a partir de especificaciones de requisitos de software. *Proceedings of the II ADIS 2001 Wor-*

kshop on Decision Support in Software Engineering, (pp.1-14). Almagro. Universidad de Salamanca. Recuperado de http://www.sc.ehu.es/jiwdocoj/remis/docs/minerw.pdf

Mosquera, R., Parra, L., & Cartrillon, O. (2016). Metodología para la Predicción del Grado de Riesgo Psicosocial en Docentes de Colegios Colombianos utilizando Técnicas de Minería de Datos. *Información tecnológica, XXVII*(6), 259-272.

Nasa, C. (2012). Evaluation of different classification techniques for web. *International Journal of Computer Applications,* 52.

Rodriguez, J., Rojas, E., & Franco, R. (2012). Clasificación de datos usando el método k-nn. *Vinculos,* 4-18.

Vijayarani , S., & Muthulakshmi, M. (2013). Comparative analysis of bayes and lazy classification algorithms. *International Journal of Advanced Research in Computer and Communication Engineering,* 3118-3124.

Villena, J., Crespo, R., & Garcia, J. (2006). Inteligencia en redes de comunicaciones. Universidad Carlos III de Madrid. Recuperado de http://ocw.uc3m.es/ingenieria-telematica/inteligencia-en-redes-de-comunicaciones/material-de-clase-1/00-presentacion

Zhang, S. (2012). Nearest neighbor selection for iteratively knn imputation. J*ournal of Systems and Software,* LXXXV(11), 2541-2552.