

Una Experiencia Personal: Pico y Pala en la Explotación y Visualización de Datos Electorales

Roberto Gabino Camana Fiallos

Facultad de Ingeniería en Sistemas
Universidad Tecnológica Indomérica

robertocamana@yahoo.es, robertocamana@uti.edu.ec

Resumen. En este trabajo se presentó una experiencia sobre la aplicación de la minería y la visualización de un conjunto de datos pertenecientes a electorales presidenciales de Ecuador, en el año 2013. Este fue el resultado de 4 años de experiencias, inició con la formación del investigador dentro de un posgrado en minería de datos, esto dio como resultado la adquisición de conocimientos para emprender investigaciones en este campo de las ciencias de la computación hasta los actuales momentos. Se detallaron dos experimentos, la primera, el estudio sobre resultados electorales, utilizando una técnica de minería de datos, con el fin de optimizar las campañas políticas. La última fue la visualización del ausentismo en las elecciones presidenciales, por medio de una aplicación *on line*, con el propósito de implementar estrategias para disminuir la ausencia de la ciudadanía, como: campañas publicitarias y programas de educación. Estos experimentos dieron lugar a que más investigadores se hayan inmiscuido en el campo de la minería de datos y la visualización, en Ecuador en cuanto a estos campos, está en fase de evolución.

Palabras Clave: Ausentismo, árboles de decisión, experiencia, minería de datos, visualización.

Abstract In this work experience on the application of mining and visualization of a data set pertaining to presidential elections in Ecuador, in 2013 was presented. This was the result of 4 years of experience, started with the formation of the researcher within a Masters in data mining, this resulted in the acquisition of knowledge to undertake research in the field of computer science to the present time. Two experiments, the first study on the election results, using a data mining technique, in order to optimize political campaigns were detailed. Last was absenteeism display in the presidential election, through an online application in order to implement strategies to reduce the lack of citizenship, such as advertising campaigns and educational programs. These experiments led to more researchers have interfered in the field of data mining and visualization, in Ecuador regarding these fields, is under development.

Keywords: absenteeism, decision trees, experience, data mining, visualization.

1 Introducción

El aumento del volumen de la información que se encuentra informatizado en bases de datos digitales ha crecido de forma vertiginosa en los últimos años. Esta información en su mayoría histórica, es procedente de transacciones y procesos de producción [1]. Uno de esos procesos son los electorales, que introducen técnicas de procesamiento automático de datos.

Debido al manejo de grandes volúmenes de datos, la votación, los escrutinios y los resultados electorales juegan un papel decisivo en época de elecciones. Además, este

fue uno de los motivos para incursionar en la extracción y visualización del conocimiento de base de datos electorales, siendo esta la primera fuente de datos provenientes del portal web del Consejo Nacional Electoral [2].

Debido a los grandes volúmenes de datos que se generan las elecciones, es necesario buscar cierto tipo de información a partir de análisis y es ahí donde entra en acción la minería de datos para estudiar procesos que lleven a descubrir patrones de conducta, se utilizan métodos de aprendizaje supervisado y no supervisado, estadística entre otras técnicas [3]. Para luego sean utilizados esos resultados en la construcción de estrategias.

Terminado el proceso de la minería de datos, el análisis, verificación de resultados obtenidos coherentes se compararon con los obtenidos por los análisis estadísticos. Por lo tanto, la visualización de información es la primera opción para mostrar los datos de forma ordenada, creativa y visual [4]. Además, permite obtener visualizaciones rápidas, novedosas y, sobre todo, que faciliten la interpretación de las representaciones por parte de los usuarios [5].

En Ecuador, en los actuales momentos los campos de: minería y la visualización de la información, están en franco desarrollo. Muestra de ello es la aplicación de técnicas de minería de datos y la visualización investigadas e implementadas por el autor.

El objetivo de esta investigación fue dar a conocer las experiencias y resultados hallados en el desarrollo de proyectos de investigación, aplicando la minería de datos y la visualización de la información sobre los datos de las elecciones presidenciales en Ecuador en el periodo 2009 y 2013. Además estas investigaciones se constituyen como pioneros, puesto que no existen otros similares en el país, como los propuestos por el autor.

2 Motivación y Justificación

En el año 2007 surgió la idea de investigar temas concernientes a las elecciones en el Ecuador, por ello se empezó a coleccionar datos electorales presidenciales hasta la actualidad, como un insumo propio, afición por los datos e investigación. De este modo al autor le proporcionó un panorama mucho más amplio, para trabajar y analizar en cuanto a datos electorales se refiere.

Este hecho le permitió al autor complementar su profesión de Ingeniero en Sistemas y Docencia, con un postgrado en minería de datos, dotándole de conocimientos y de herramientas para el estudio y tratamiento de datos con la finalidad de extraer conclusiones e información relevante de ellos.

En este artículo se recogió los frutos de investigaciones realizadas por el autor, producto de dos proyectos de investigación: “Aplicación de la minería de datos para obtener patrones de comportamiento del electorado” y el segundo proyecto “La tecnología de Minería de Datos y su incidencia en el análisis avanzado de datos históricos electorales en Ecuador”, además se pretendió por un lado, mostrar la utilidad de las técnicas y herramientas para el análisis de datos electorales y por otro la visualización de datos electorales con una nueva gráfica interactiva, permitiendo visualizar en tiempo real y en la web.

Para el desarrollo de la investigación, se realizó en equipo. Por un lado, estudiantes se encargaron de realizar el trabajo de la extracción de datos históricos y actuales, depuración y limpieza, hasta construir un almacén de datos, que sirvió de base para realizar los diferentes modelos. Por el otro, el autor y docente de la cátedra de base de datos, trabajó con la creación de modelos con algoritmos de aprendizaje supervisado y aprendizaje no supervisado, y posteriormente desarrollado las visualizaciones para la toma de decisiones de candidatos, organizaciones políticas, consultores, encuestadores, en cara a las próximas elecciones. Todos los actores participaron activamente en cada una de las etapas del desarrollo de la investigación, sin embargo; consideramos necesario aclarar los motivos por los cuales trabajamos de esta manera.

Para que no quede en mera investigación académica, esta tuvo la vinculación: universidad y sociedad, con organizaciones políticas y consultoras de marketing político, entregándoles los instrumentos desarrollados para su uso y puesta en ejecución para sus intereses.

Además, con este artículo se busca generar motivación e interés de estudiantes y de profesionales inmersos en la minería de datos y descubriendo del conocimiento, campo de las ciencias de la computación. Para ayudar a mejorar y optimizar campañas electorales.

3. Metodología Aplicada

La metodología para la minería de datos que se siguió es la Extracción de Conocimiento, conocido como Knowledge Discovery in Databases (KDD). Para el desarrollo de visualizaciones se siguió la metodología planteada por Ben Fry [6], los pasos de estas dos metodologías se describen a continuación:

A. Para la Minería de Datos

Se planteó la metodología Knowledge Discovery in Databases (KDD) para el desarrollo de proyectos de minería de datos, porque este método se ajustó eficientemente a la resolución de problemas de la minería de datos. Además, esta metodología se refiere a un proceso no-trivial de descubrir conocimiento e información potencialmente útil dentro de los datos contenidos en algún repositorio de información [7]. La Fig. 1, ilustra las etapas del proceso KDD:

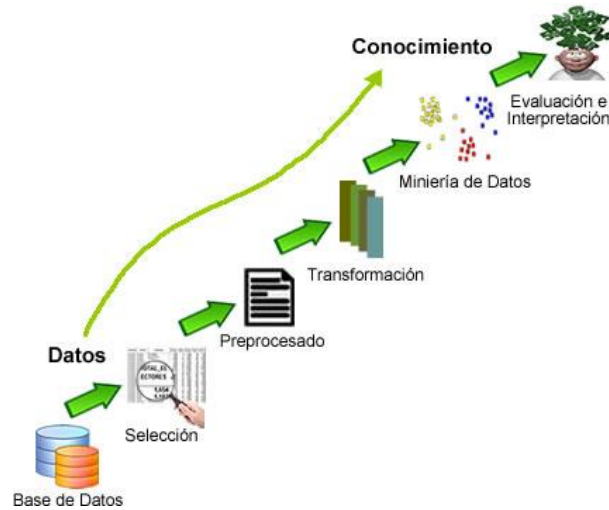


Figura 1. Proceso KDD

Para el desarrollo de las etapas KDD, se planteó una pregunta ¿Cuáles son los sitios más óptimos en el Ecuador para realizar campañas políticas?, las etapas del proceso KDD se dividió en 5 y que a continuación se resumen:

1ra. Fase: Selección

- Datos públicos provenientes de la única fuente primaria, siendo esta la del CNE [2].
- Elecciones presidenciales del Ecuador febrero 2013. El número de casos para este estudio fueron de 1.258 Juntas Receptoras del Voto (JRV).

2da. Fase: Reprocesamiento

- Se buscó la distribución de los datos, su simetría, la normalidad y las correlaciones existentes.
- Limpieza de los datos, es decir eliminar o se corrigen los datos incorrectos.

3ra. Fase: Transformación

- Generación de nuevas variables a partir de las ya existentes.
- Se utilizó Weka para definir las nuevas variables.

4ta. Fase: Minería de Datos

- Búsqueda de patrones de interés en una determinada forma de representación,
- Determinar el objetivo de minería de datos (para este ejemplo se utilizó aprendizaje supervisado).

5ta. Fase: Evaluación e interpretación

- Se evaluó patrones y fue analizado por los expertos.
- Fue necesario volver a las fases anteriores para una nueva iteración, con el objetivo de utilizar nuevos datos con otros algoritmos.

B. Para la Visualización de la información

Para el desarrollo de los proyectos de visualización se siguió la metodología de Ben Fry [6], porque permitió construir visualizaciones interactivas, la Fig. 2, muestra los pasos que se siguieron:

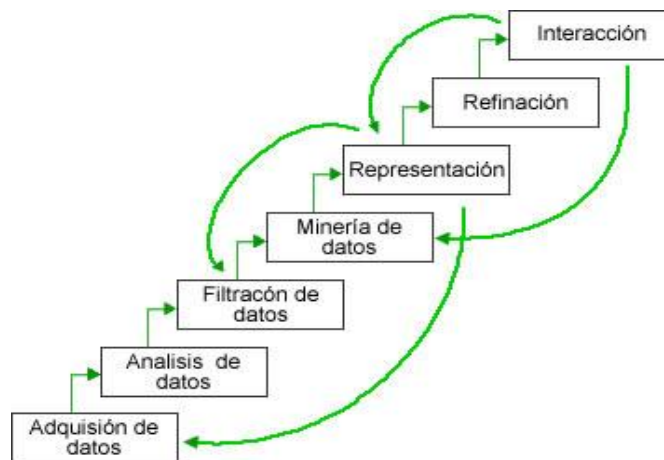


Figura 2. Metodología Ben Fry.

En la Fig. 2, se muestra los pasos de la metodología que se siguió. En primer lugar partiremos de una pregunta puntual, de este modo se simplificó el problema: ¿Qué provincias del Ecuador y emigrantes en el exterior tuvieron mayor ausentismo por hombre y mujer en las elecciones presidenciales 2013?. A continuación se resumieron los siguientes pasos:

Paso 1: Adquisición de datos

- La extracción de datos de las elecciones del año 2013, fue a través del portal web del Consejo Nacional Electoral [2].
- Archivos con todos los datos extraídos, con 40.451 casos de estudios.

Paso 2: Análisis de datos

- Conversión de los datos a un formato de datos: CSV.
- Se incluyó a las 24 provincias del Ecuador y 20 países donde residen emigrantes ecuatorianos.
- La herramienta informática utilizada fue Excel.

Paso 3: Filtración de datos

- Limitar el ausentismo por género (hombre y mujer) y por provincias.

- Limitar a 24 provincias del Ecuador y 3 zonas de emigrantes, con 55 casos para el estudio.

Paso 4: Minería de datos

- Extracción de datos significantes, como provincia, ausentismo y género (hombre y mujer).
- La herramienta de minería de datos utilizada fue Weka.

Paso 5: Representación

- Elegimos un modelo básico para la visualización de la información, por ejemplo el gráfico de burbuja.
- Utilizamos la aplicación gratuita *on line* Many Eyes de IBM [8].

Paso 5: Refinación

- Se modificó la variable ausentismo en expresión de porcentaje, porque es más manejable que número de miles.
- Creación de mapas interactivos *on line*.

Paso 6: Interacción

- Mostrar la información textual al seleccionar un gráfico por provincia.
- Cambio dinámico tras seleccionar cualquier provincia.
- Visualización *on line* y dinámica: [Ausentismo por hombre y mujer](#) [9].

3 Resultados y Discusión

C. La minería de datos

Esta fase es de modelamiento, donde métodos o técnicas fueron aplicados con el objetivo de extraer patrones previamente desconocidos, válidos, nuevos, potencialmente útiles y comprensibles que están contenidos u “ocultos” en los datos [10].

Se realizaron varios experimentos, el que se destacó fue el uso algoritmos de aprendizaje supervisado como el árbol de decisión: J48 y RandomTree, que fueron comparados para determinar cuál de ellos fue el mejor clasificador de casos estudiados. A continuación se detallan los arboles de decisión a utilizar:

Árbol de decisión: J48

Es un clasificador basado en árboles para instancias (datos) representados como vectores de características (*features*), además los nodos prueban características que hay en una rama, para cada valor de la característica, las hojas especifican la categoría [11], como se observa en la Fig. 3.



Figura 3. Ejemplo del árbol de decisión.

Descritas como regla, se obtuvo: $Tungurahua \wedge Pishilata \rightarrow La\ Joya$

Árbol de decisión: RandomTree

Siguiendo los conceptos de [12], un RandomTree es un árbol dibujado al azar de un juego de árboles posibles. En este contexto "al azar" significa que cada árbol en el juego de árboles tiene una posibilidad igual de ser probado. Otro modo de decir esto consiste en que la distribución de árboles es 'uniforme'.

Comparación entre árboles de decisión

Conocidos como trabajan los árboles de decisión: j48 y RandomTree, se procedió a la generación de los árboles en Weka con los siguientes resultados, como se observan a continuación (Tabla 2):

TABLA II
Resultados del árbol de decisión J48

| Criterio | a | b | c | d |
|----------------------|-------|---|---|---|
| MIP AIS | 932 | 1 | 0 | 0 |
| CREO | 319 | 2 | 0 | 0 |
| PSP | 2 | 0 | 0 | 0 |
| OTROS | 2 | 0 | 0 | 0 |
| Error medio absoluto | 0.185 | | | |

| | |
|------------------------------|-------|
| Precisión de la clase del c0 | 0.738 |
| Precisión de la clase del c1 | 0 |
| Precisión de la clase del c2 | 0 |
| Precisión de la clase del c3 | 0 |

El algoritmo decisión j48, fue sometido a prueba, utilizó valores para el árbol creado de 3 nodos y 8 hojas, con un tamaño de 15. Como se observa en una sola clase (a = MIPAIS) clasificó todas las juntas receptoras del voto o casos de estudio.

En el segundo árbol de decisión RandomTree, se obtuvo los siguientes resultados (Tabla 3):

TABLA III
Resultados del árbol de RandomTree

| Criterio | a | b | C | d |
|------------------------------|----------|----------|----------|----------|
| MIPAIS | 930 | 3 | 0 | 0 |
| CREO | 56 | 265 | 0 | 0 |
| PSP | 1 | 0 | 1 | 0 |
| OTROS | 1 | 0 | 0 | 1 |
| Error medio absoluto | 0.0268 | | | |
| Precisión de la clase del c0 | 0.941 | | | |
| Precisión de la clase del c1 | 0.989 | | | |
| Precisión de la clase del c2 | 1 | | | |
| Precisión de la clase del c3 | 1 | | | |

Tras ejecutar el algoritmo RandomTree con los parámetros nodos 3, hojas 8 y el tamaño 15, se observa en la Tabla 3 que es un árbol que clasifica correctamente cada clase (a = MIPAIS, b = CREO, c = PSP y d = OTROS). Además se obtuvo un árbol en forma gráfica como en la Fig. 4.

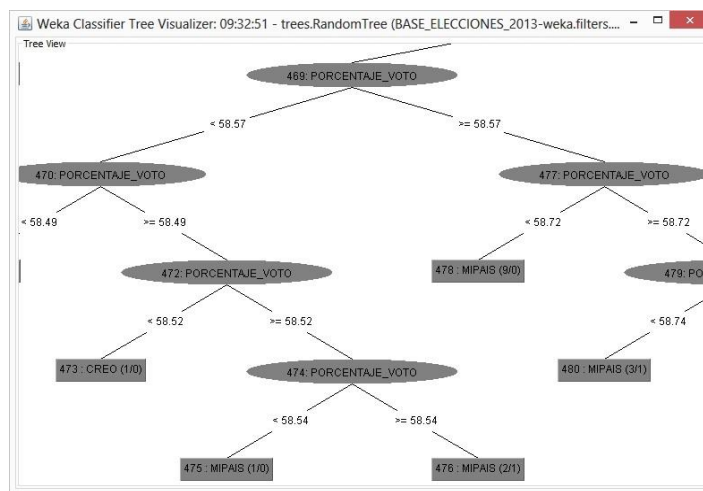


Figura 4. Árbol del algoritmo Random Tree.

Este algoritmo al generar varias hojas (cuadrados) da la posibilidad para clasificar más casos correctamente y con un error mínimo de clasificación de casos, como se muestra en la comparación de árboles de decisión.

Así mismo se obtuvo una muestra de los sitios donde se deben realizar las campañas políticas para las dos organizaciones políticas más representativas MPAIS, lista 35 y CREO, lista 21 (Tabla 4):

TABLA IV
Lista de sitios para campañas políticas

| PROVINCIA | CANTON | PARROQUIA | 35 | 21 |
|------------|------------|-----------------------|----|----|
| Azuay | Gualaceo | Gualaceo | ✓ | ✓ |
| Bolívar | Echeandia | Echeandia | ✓ | |
| Bolívar | Caluma | Caluma/San Antonio | | ✓ |
| Cañar | La Troncal | La Troncal | ✓ | ✓ |
| Carchi | Montufar | San José | ✓ | |
| Carchi | Montufar | Gonzalez Suarez | | ✓ |
| Chimborazo | Guamote | Guamote | ✓ | |
| Chimborazo | Colta | Columbe | | ✓ |
| Cotopaxi | Pujili | Pulili | ✓ | |
| Cotopaxi | Salcedo | San Miguel de Salcedo | | ✓ |
| El Oro | Pasaje | Ochoa León/Matríz | ✓ | |

| | | | | |
|--------------------|--------------------|-----------------------|---|---|
| El Oro | Santa Rosa | Santa Rosa | | ✓ |
| Esmeraldas | Quininde | Rosa Zarate | ✓ | ✓ |
| Galápagos | San Cristobal | Pto Baquerizo Moreno | ✓ | ✓ |
| Guayas | Duran | Eloy Alfaro/ Duran | ✓ | |
| Guayas | Milagro | Milagro | | ✓ |
| Imbabura | Ibarra | Sagrario | ✓ | |
| Imbabura | Otavalo | Jordan | | ✓ |
| Loja | Catamayo | Catamayo | ✓ | ✓ |
| Los Ríos | Vinces | Vinces | ✓ | |
| Los Ríos | Babahoyo | Clemente Baquerizo | | ✓ |
| Manabí | El Carmen | El Carmen | ✓ | |
| Manabí | Chone | Chone | | ✓ |
| Morona Santiago | Gualaquiza | Gualaquiza | ✓ | ✓ |
| Napo | Archidona | Archidona | ✓ | ✓ |
| Orellana | Joya de los Sachas | La Joya de los Sachas | ✓ | ✓ |
| Pastaza | Pastaza | Puyo | ✓ | ✓ |
| Pichincha | Cayambe | Cayambe | ✓ | ✓ |
| Santa Elena | La Libertad | La Libertad | ✓ | ✓ |
| Sto Dgo Tsachillas | Santo Domingo | Santo Domingo | ✓ | ✓ |
| Sucumbíos | Shushufindi | Shushufindi Central | ✓ | ✓ |
| Tungurahua | Pelileo | Pelileo | ✓ | ✓ |
| Zamora Chinchipe | Zamora | Zamora | ✓ | ✓ |

Se observa en la Tabla 4, que son 24 parroquias del Ecuador las más idóneas para hacer campañas políticas por las organizaciones políticas: MPAIS y CREO.

La distribución del voto, es la más importante en toda elección popular, ya que estos resultados servirán de base para las próximas elecciones, es decir conociendo los puntos de concentración de los electores por cualquiera de las dos organizaciones políticas, se determinará efectuar campañas con mayor éxito, sobre este electorado que será decisivo en las elecciones.

D. Las Visualizaciones

La aplicación de herramientas para la visualización de conjuntos de datos electorales es un reto de la computación gráfica, sobre todo si se trata de mostrar miles de datos en una forma atractiva y dinámica [4]. A continuación se muestran en la Fig. 5 un tipo de visualización seguida por la metodología Ben Fry:

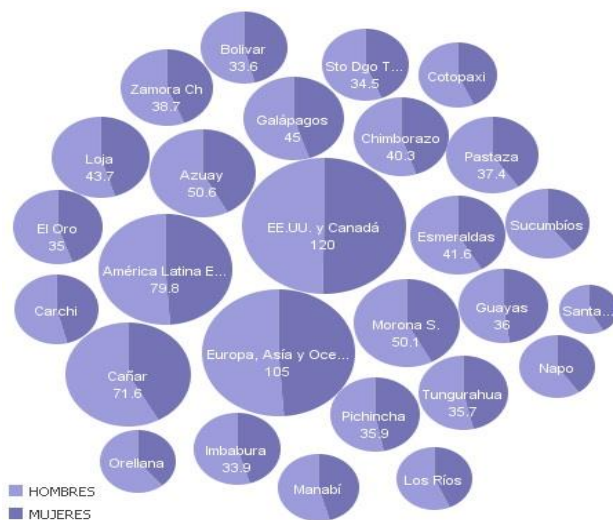


Figura 5. Ausentismo en las elecciones presidenciales 2013.

El objetivo de esta visualización fue mostrar en un solo vistazo la información por provincias y zonas que agrupan a países de emigrantes ecuatorianos, donde hubo ausentismo en las elecciones presidenciales 2013. Como se observa en la Fig. 2, la zona EEUU y Canadá, registró el mayor ausentismo en hombres con el 59.93%, en mujeres el 60.35% y en territorio nacional la provincia de Cañar registró mayor ausencia de hombre con el 41.5% y la mayor ausencia de mujeres se obtuvo en la misma provincia antes mencionado, alcanzando el 30.08%.

También se tomó una muestra de las provincias y zonas de emigrantes ecuatorianos, con la finalidad de que se conozca de donde se obtuvieron los datos (Tabla 1).

TABLA I
DATOS DE AUSENTISMO POR PROVINCIAS

| N° | PROVINCIAS Y ZONAS | HOMBRES | MUJERES |
|----|--------------------|---------|---------|
| 1 | Azuay | 29.27% | 21.38% |
| 2 | Bolivar | 18.12% | 15.50% |
| 3 | Cañar | 41.50% | 30.08% |
| 4 | Carchi | 17.31% | 14.61% |
| 5 | Chimborazo | 22.10% | 18.16% |
| 6 | Cotopaxi | 16.15% | 12.31% |
| 7 | El Oro | 19.39% | 15.62% |
| 8 | Esmeraldas | 24.31% | 17.28% |
| 9 | Galápagos | 24.79% | 20.24% |
| 10 | Guayas | 18.86% | 17.19% |
| 11 | Imbabura | 18.58% | 15.32% |

| | | | |
|-----|-----------------------------------|--------|--------|
| 12 | Loja | 23.88% | 19.83% |
| 13 | Los Ríos | 14.98% | 11.44% |
| 14 | Manabí | 16.31% | 13.63% |
| 15 | Morona S. | 29.00% | 21.09% |
| 16 | Napo | 15.63% | 10.58% |
| 17 | Orellana | 15.61% | 9.93% |
| 18 | Pastaza | 22.25% | 15.20% |
| 19 | Pichincha | 19.05% | 16.83% |
| 20 | Santa Elena | 9.10% | 6.54% |
| 21 | Sto Dgo Tsachillas | 19.36% | 15.10% |
| 22 | Sucumbíos | 18.70% | 12.21% |
| 23 | Tungurahua | 19.11% | 16.61% |
| 24 | Zamora Ch | 21.78% | 16.95% |
| 25E | América Latina El Caribe y Africa | 40.91% | 38.94% |
| 26E | EE.UU. y Canadá | 59.93% | 60.35% |
| 27E | Europa, Asia y Oceanía | 53.79% | 51.45% |

Lo importante de esta visualización no es solamente sus gráficos, sino la interesante y novedosa información que se pudo obtener [4]. Además, este instrumento servirá como diagnóstico para implementar estrategias (campañas publicitarias y programas de educación) para reducir el ausentismo electoral que proviene de la falta de interés de los ciudadanos, emigración y analfabetismo hacia las elecciones presidenciales.

4 Conclusiones

El análisis de datos electoral puede ser útil para optimizar campañas políticas que servirán en cara a las próximas elecciones. Además, la aplicación de árboles de decisión, que de una forma matemática por medio de algoritmos seleccionó sitios puntuales en el territorio ecuatoriano, donde se podrían realizar proselitismos políticos (campañas), de esta forma ahorrar tiempo y recursos. Estas investigaciones se complementaron con nuevas formas de representar la información.

La representación, por medio de la visualización, es la opción para representar la información más versátil, que de una forma tradicional (barras, círculos, etc.). La nueva forma de visualizar la información es; dinámica, por ello se presenta una visualización del ausentismo de las elecciones presidenciales de 2013, con la finalidad de conocer que provincias son las más sensibles al ausentismo, para implementar estrategias a corto plazo como campaña publicitaria y a largo plazo programas de educación.

Reconocimiento

El presente artículo no hubiera sido posible sin el apoyo de la Facultad de Ingeniería en Sistema de la Universidad Tecnológica Indoamérica, así como el aporte valioso de investigadores (analistas), auxiliares de investigación (estudiantes) y

evaluadores externos, como también el Consejo Nacional Electoral (CNE).

Referencia

- [1] R. Camana, "Digitalizando nuestra vida cotidiana," in *Generar Conocimiento el Pensar y Hacer*, Ambato, 2012, pp. 73-74
- [2] Consejo Nacional Electoral [online], 2013 Disponible en: <http://resultados.cne.gob.ec/Results.html?RaceID=1&UnitID=1&IsPS=0&LangID=0>
- [3] R. Camana, "Aplicación de técnicas de minería de datos para la indagación y estudio de resultados electorales," *CienciAmerica*, Año 1, N° 1, pp. 85-94, Enero, 2012.
- [4] R. Camana, "Visualización de datos electorales presidenciales: herramientas para la visualización de grandes conjuntos de datos," *CienciAmerica*, Año 2, N° 2, pp. 37-45, Diciembre, 2013
- [5] L. Freeman, *Visualizing social networks* [online]. EEUU: Universidad de California, 2013 Disponible en: <http://www.cmu.edu/joss/content/articles/volume1/Freeman.html>
- [6] B. Fry, "Visualizing Data: Exploring and Explaining Data with the Processing Environment," O'Reilly Media, 1 edition, Febrero, 2013.
- [7] WebMining [online], 2013 Disponible en: <http://www.webmining.cl/2011/01/proceso-de-extraccion-de-conocimiento/>
- [8] Many Eyes – IBM [online], 2013. Disponible en: <http://www-958.ibm.com/software/analytics/manyeyes/>
- [9] IBM [online], 2014. Disponible en: <http://www-958.ibm.com/software/analytics/manyeyes/visualizations/ausentismo-x-hombre-y-mujer>
- [10] Web Mining. [online], 2014. Disponible en <http://www.webmining.cl/2011/01/proceso-de-extraccion-de-conocimiento/>
- [11] T. Arredondo, *Árboles de decisión* [online]. Chile: Universidad Técnica Federico Santa María, 2014 Disponible en: <http://profesores.elo.utfsm.cl/~tarredondo/info/soft-comp/Arboles%20de%20Decision.pdf>
- [12] Gla.ac. "Page Lab" [online], 2013. Disponible: <http://taxonomy.zoology.gla.ac.uk/rod/cplite/ch6.pdf>