Aplicación de Técnicas de Minería de Datos para la Indagación y Estudio de Resultados Electorales

Roberto CAMANA FIALLOS

Facultad de Ingeniería en Sistemas
Universidad Tecnológica Indoamérica
Bolívar 20-35 y Guayaquil, Ambato, Ecuador
robertocamana@yahoo.es,
robertocamana@uti.edu.ec

RESUMEN

En esta investigación se trata un problema de Minería de Datos, utilizando técnicas de aprendizaje supervisado y no supervisado. Se estudió un conjunto de datos reales de las elecciones presidenciales del 28 de abril de 2009, correspondientes a la zona No. 3, integrada por las provincias de Chimborazo, Cotopaxi, Pastaza y Tungurahua. El problema se dividió en dos partes. La primera, fue la selección de las Juntas Receptoras del Voto más representativas y la segunda, el análisis de distribución del voto. Los algoritmos de aprendizaje supervisado que se utilizaron fueron DesicionStrump, j48, LMT y perceptrón multicapa. Los algoritmos no supervisados que se utilizaron fueron EM (Expectation Maximization) y K-Medias. Las herramientas utilizadas fueron Crementine y Weka, un programa de código abierto que proporciona una gran variedad de algoritmos de aprendizaje muy útiles en la Minería de Datos. Programados en lenguaje Java, se crearon los programas: Preprocesamiento.java, Resultado.java, PorMesa.java, PorMesaGana.java y Distancias.java.

PALABRAS CLAVE

algoritmo de aprendizaje, datos electorales, exploración, minería de datos, verificación.

ABSTRACT

This research is focused on a data mining problem, using techniques of supervised and unsupervised learning. I studied a set of real data from presidential elections April 28, 2009, of area No. 3, which includes the provinces of Chimborazo, Cotopaxi, Pastaza and Tungurahua. The problem was divided into two parts. The first was the selection of the most representative polling stations and the second, the vote distribution analysis. The supervised learning algorithms used were: DesicionStrump, J48, LMT, and multilayer perceptron. The unsupervised algorithms used were: EM (Expectation Maximization) and K-Means. The tools used were Crementine and Weka, an open source program that provides a variety of useful learning algorithms in data mining. Programmed in Java language, a number of programs were created: Preprocesamiento.java programs, Resultado.java, PorMesa.java, PorMesaGana.java and Distancias.java.

KEYWORDS

learning algorithms, election data, exploration, mining, verification.

1. Introducción

Una definición de Minería de Datos es el descubrimiento eficiente de información valiosa, no-obvia de una gran colección de datos. Su objetivo es ayudar a buscar situaciones interesantes con los criterios correctos, complementando una labor que hasta ahora se ha considerado "intelectual" y de alto nivel, privativa de los gerentes, planificadores y administradores [1]. Además, pueden de realizar búsquedas fuera de horas pico, usando tiempos de máquina excedentes. En general, la metodología de la Minería de Datos se puede ver en la Figura 1.

La utilidad de la Minería de Datos ya no se pone en duda [1, 2], por lo cual esta tecnología está siendo aplicada por muchas herramientas de *software.* Las técnicas de aplicación varían de acuerdo a la herramienta.

El objetivo de esta investigación fue utilizar herramientas de minería de datos para la selección de las Juntas Receptoras del Voto más representativas y realizar un análisis de distribución del voto a nivel nacional durante las elecciones presidenciales de 2009.

Este análisis es muy importante, ya que brinda una clara señal para realizar sondeos a "pie de urna" o "boca de urna" de forma fiable, reduciendo el costo de los mismos. Así, bastaría con realizarlos en mesas o Juntas Receptoras del Voto más representativas, en futuras elecciones.

2. Metodología

Para iniciar esta investigación, se tomaron dos fuentes de datos: Datos históricos generales (únicamente por provincias) de las elecciones presidenciales (desde 1980 hasta 2004), proporcionados por el Consejo Nacional Electoral (CNE), y otros obtenidos desde su página web http://cne.gob.ec. Cabe indicar que estos datos tuvieron una disgregación por provincia, cantón y parroquia, de las elecciones de Presidente y Vicepresidente, en la primera vuelta de 2009.

Debido al gran volumen de datos que se generan las elecciones Presidenciales o de dignidad popular, en conjunto con el progreso tecnológico de la última década, se ha logrado que el análisis de los datos se demore menor tiempo [3]. La metodología a seguir consta de fases sucesivas que se cumplen en orden jerárquico. figura 1.

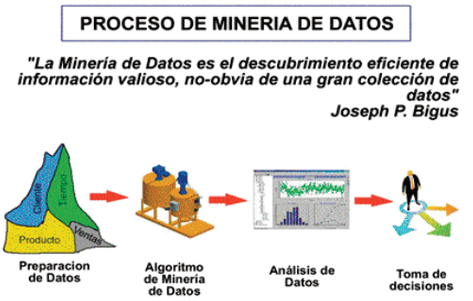

Figura 1. Metodología de Minería de datos

A continuación se describen cada una de las fases que se siguieron durante la realización de este trabajo:

2.1. Preparación de Datos

Esta fase consistió en la limpieza y selección de datos, iniciando con la eliminación del mayor número posible de datos erróneos, inconsistentes e irrelevantes.

2.1.1. Selección de los datos. Luego de haber realizado la limpieza, se seleccionó un conjunto inicial de datos, de cual se partió. Estos datos se almacenaron en un archivo de texto en el que cada patrón correspondía a una Junta Receptora del Voto (JRV). Existe un total de 4.379 JRV. Cada uno de estos patrones tiene 18 atributos, que se especifican en la Tabla 1:

Tabla 1. Atributos utilizados

Número	Nombre	Descripción
1	PRO	Provincia
2	CAN	Cantón
3	PAR	Parroquia
4	TOT	Total Electores Padrón
5	JUNH	Junta Hombres
6	JUNM	Junta Mujeres
7	MPAIS	Movimiento Patria Altiva I Soberana
8	PSP	Partido Sociedad Patriótica "21 de Enero"
9	PRIAN	Partido Renovador Institucional Acción Nacional

10	RED/MIPD	Alianza Izquierda Unida
11	MTM	Movimiento Triunfo Mil
12	MTF	Movimiento Tierra Fértil
13	MIJS	Movimiento Independiente Justo y Solidario
14	MITS	Movimiento de Integración y Transformación Social
15	VOT	Votación Total
16	TOTALB	Total Votos en Blanco
17	TOTALN	Total Votos Nulos
18	TOTALA	Total Ausentismo

Para llevar a cabo este proceso, en primer lugar, se transformó el archivo de texto a un archivo con extensión .arff, que es usado por el programa de código abierto Weka. En la primera línea aparece @relation datos_electorales_2009, que especifica el nombre de la relación. En las siguientes líneas están los diferentes atributos en la forma @attribute cert, si toma un valor numérico y en la forma: @attibute mesa {1, 2, 3, 4, …n}, si es una variable nominal y toma valores discretos. Cuando se han puesto todos los atributos se escribe @data y a continuación se pone un patrón por línea con los atributos separados por comas. Debido a la gran cantidad de partidos políticos que participan en las elecciones, se optó por limitar el análisis a MIPAIS, PSP y PRIAN, mientras que los demás partidos fueron agrupados con OTROS. Esta última agrupación obedece a que muchos partidos pequeños no obtuvieron un número significativo de votos.

2.1.2. Selección especifica de Juntas Receptoras del Voto (JRV) más representativas. Se obtuvieron datos reales sobre los resultados electorales correspondientes a las elecciones de Presidente y Vicepresidente, de abril de 2009, en la Zona N° 3 (Chimborazo, Cotopaxi, Pastaza y Tungurahua). El objetivo principal era seleccionar una serie de JRV, cuyos resultados fueran representativos.

Por otro lado, se extrajo otro tipo de conocimiento de los datos, tomando en cuenta la disgregación por provincia, cantón, parroquia y número de JRV (hombre o mujer) por votantes; esta información facilitó el análisis de selección de atributos específicos de juntas del voto más representativas. Este tipo de conocimiento se conoce como conocimiento oculto, o no es evidente, es desconocido a priori, pero puede ser muy útil.

2.2. Algoritmos de Minería de Datos

Un algoritmo de Minería de Datos es un conjunto de cálculos que permiten crear modelos de Minería de Datos. El algoritmo analiza primero los datos proporcionados, en búsqueda de patrones o tendencias [9]. Los algoritmos que se que se utilizaron fueron:

2.2.1. Algoritmos de aprendizaje supervisado (clasificación). La clasificación de datos desarrolla una descripción o modelo para cada una de las clases presentes en la base de datos. Existen muchos métodos de clasificación, entre los cuales se destacan los siguientes:

- Árboles de decisión. Los árboles de decisión son algoritmos de aprendizaje por inducción supervisada que pretenden modelar los datos de ejemplo mediante un árbol. Los nodos intermedios son los atributos de entrada de los ejemplos presentados, las ramas representan valores de dichos atributos y los nodos finales con los valores de la clase [5]. Para elegir qué atributos y en qué orden aparecen en el árbol, se utiliza una función de evaluación llamada ganancia de información (reducción de entropía del conjunto al clasificar usando un determinado atributo). Tienen como ventaja que son fáciles de programar pues se traducen en regla if-else. Se trabajan con atributos nominales únicamente, como el ID3, y que trabajan también con atributos numéricos, como el C4.5 (j48 en Weka) [6].

- Perceptrón multicapa. Es un tipo de red neuronal. Las redes neuronales son una simplificación matemática bastada en el proceso estimulo/respuesta de las neuronas. Son capaces de aprender o modelar la relación existente entre sus entradas y salidas mediante la modificación de los valores de los pesos de las conexiones que unen las entradas con las neuronas y las neuronas entre sí [4].

El modelo más simple de red neuronal es el perceptrón simple, que no tiene capas de neuronas ocultas. La relación entre entradas y salida viene dada por:

$$y = F\left(\sum_{i=1}^{n} w_i x_i + \theta\right)$$

Siendo *F* la función de activación.

El problema de este tipo de estructura es que solo sirve para resolver problemas de complejidad lineal (relación lineal entre entradas y salidas), de ahí la necesidad de introducir capas de neuronas ocultas y dar lugar al perceptrón multicapa, como el que se usa en este estudio. El problema de entrenar las neuronas de las capas ocultas para actualizar los pesos de sus conexiones se resuelve con el algoritmo de *BackPropagation* (retropropagación) [5].

- Entrenamiento y validación. Para que evaluar el aprendizaje, se divide el conjunto de datos en dos grupos: entrenamiento y test. El primero para entrenar el modelo y el segundo para validarlo, comprobando en este último caso que el error es lo suficientemente pequeño. Suele usarse un 80% de muestras para entrenamiento y un 20% para test, y se pretende que ambos conjuntos sean capaces de representar al conjunto total de datos, de no ser así aparece el llamado error de muestreo. Para evitarlo puede usarse validación cruzada (*crossvalidation*): Se divide el conjunto inicial de datos en N partes, se entrena/evalúa N veces, cada vez con un conjunto de test diferente y se da como error final la media de las *N* medidas de error [7].

2.2.2. Algoritmos de aprendizaje no supervisado (agrupamiento). Un algoritmo de agrupamiento (en inglés, *clustering*) es un procedimiento de agrupación de una serie de vectores de acuerdo con un criterio de cercanía. Esta cercanía se define en términos de una determinada función de distancia, como la Euclídea [8]. Entre las cuales tenemos:

- k-Medias. Algoritmo de clasificación no supervisado, inventado por J. MacQueen en 1967, mediante el cual el espacio de patrones de entrada se divide en K clases o regiones, cada una representada por un punto llamado centroide. Dichos centros se determinan con el objetivo de minimizar las distancias euclideas entre los patrones de entrada y el centro más cercano. Los pasos para la aplicación del algoritmo son:

1. Se inicializan aleatoriamente los centros de las K clases.

2. Se asignan N_i patrones de entrada a cada clase i del modo:

- El patrón X(n) pertenece a la clase i sí:

$$\|X(n) - C(i)\| < \|X(n) - C_S\| \quad \forall\, s \neq i \; con \; s = 1, 2, \dots, K.$$

- Por tanto, cada clase tendrá asociado un determinado número de patrones de entrada, aquellos más cercanos al centro de la clase.

3. Se calcula la nueva posición de los centros de las clases como la media de todos los patrones que pertenecen a su clase, es decir:

$$c_{ij} = \frac{1}{N_i}\sum_{n=1}^{N} M_{in} x_j(n) \; para \; j = 1, 2, \dots p \; , i = 1, 2, \dots K$$

4. Se repiten los pasos 2 y 3 hasta que las nuevas posiciones de los centros no se modifiquen respecto a su posición anterior, es decir hasta que:

$$\|C_i^{nuevo} - C_i^{anterior}\| < \in \forall\, i = 1, 2, \dots, K$$

El algoritmo de K-medias es un método fácil de implementar y usar. Suele ser un algoritmo bastante eficiente en problemas de clasificación, pues converge en pocas iteraciones hacia un mínimo de la función, aunque podría tratarse de un mínimo local. Su principal inconveniente es su dependencia de los valores iniciales asignados a cada centro (mínimos locales).

- EM *(Expectation Maximization)*. Es un método no supervisado de aprendizaje. Se trata de un

estimador ML que maximiza la log-verosimilitud de los datos incompletos iterativamente maximizando la esperanza de la log-verosimilitud de los datos completos, donde los datos completos están formados por los observables (incompletos) y los no observables [9]. El problema es que no se sabe de qué distribución viene cada dato y no se conocen los parámetros de las distribuciones.

El algoritmo EM empieza adivinando los parámetros de las distribuciones y los usa para calcular las probabilidades de que cada objeto pertenezca a un *cluster* y usa esas probabilidades para re-estimar los parámetros de las probabilidades, hasta converger (se puede empezar adivinando las probabilidades de que un objeto pertenezca a una clase) [9].

El cálculo de las probabilidades de las clases o los valores esperados de las clases es la parte de *expectation*. El paso de calcular los valores de los parámetros de las distribuciones, es *maximization*, maximizar la verosimilitud de las distribuciones dados los datos [9].

Para estimar los parámetros, se debe considerar que se conocen únicamente las probabilidades de pertenecer a cada *cluster* y no los *clusters* en sí mismos. Estas probabilidades actúan como pesos.

Como condición de parada del algoritmo suele tomarse el momento cuando la log-verosimilitud no varía de manera significativa. Aunque EM garantiza convergencia, esta puede ser a un máximo local, por lo que se recomienda repetir el proceso varias veces [9].

2.3. Análisis de Datos

Consiste en el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes fuentes de datos, siendo los mismos obtenidos de distintos formatos. La minería de datos encuentra modelos explicables a partir de los datos. Para que este proceso sea efectivo debe ser automático o semiautomático. En el caso de esta investigación, el uso de patrones de descubrimiento ayuda a la investigación del comportamiento de electores en proyectos del Consejo Nacional Electoral. Los modelos son de dos tipos, predictivos y descriptivos; este último será aplicable a esta investigación, por lo que este modelo identifica patrones que explican en resumen las propiedades examinadas, no para predecir nuevos datos [10, 11].

2.3.1. Herramientas. Para la realización de este trabajo se usó fundamentalmente el programa de código abierto *Weka* y se programó código Java para dar formato a los ficheros de entrada a Weka, así como para calcular el resultado de las elecciones y las distancias de este resultado a las diferentes mesas o JRV. Los programas creados fueron: Preprocesamiento.java, Resultado.java,PorMesa.java, PorMesaGana.java y Distancias.java.

2.3.2. Construcción del modelo y validación. La construcción del modelo requiere que los datos preparados puedan ser utilizados iterativamente, en otras palabras, poder aplicar algoritmos y técnicas sobre diferentes vistas "minables" y de esta manera descubrir patrones de comportamiento de electores [9]. Para la presente investigación, se utilizó la herramienta de minería de datos Clementina 11. 1 y Weka, este se caracteriza por:

- **Acceso a Datos:** Fuentes de datos ODBC's, tablas de Excel y archivos planos.

- **Pre procesamiento de Datos:** Muestreo, particiones y reordenación de campos.

- **Técnicas Aprendizaje:** Arboles de decisión, C5.0 y el C&RT, redes neuronales, agrupamientos, reglas, de asociación, regresión lineal y logística.

- **Visualización de Resultados:** Soportes gráficos que permiten al usuario tener una visión global de todo el proceso, que comprende desde el análisis del problema hasta la imagen final del modelo aprendido [11].

2.4. Toma de Decisiones (evaluación)

Este es el último paso, es donde se evalúa el modelo, teniendo en cuenta el cumplimiento los criterios de éxito del problema, como la selección de una serie de Juntas Receptoras del Voto cuyos resultados sean repetitivos. Esto determinará los puntos claves para realizar sondeos "a boca de urna" de forma fiable, ya que las mismas podrían servir para futuras elecciones a cualquier dignidad de elección popular (Presidente, asambleístas, prefectos, entre otros).

2.4.1. Análisis de las relaciones entre variables.

La opción utilizada del software Weka, es la opción *Visualize*, que permite observar gráficamente la relación entre los diversos atributos (dos a dos). Esta opción permite explorar los datos y encontrar patrones interesantes.

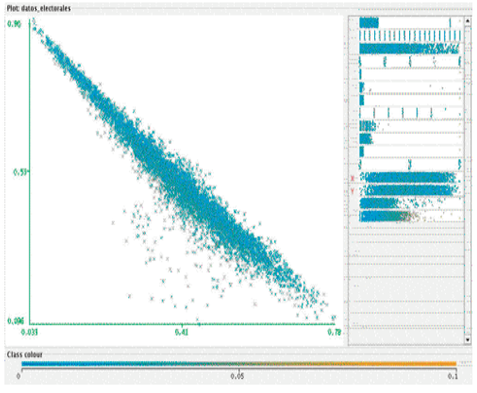

Figura 2. Votación Total (MPAIS de color verde y Votación Total (PSP) de color azul.

3. Resultados y Discusión

3.1. Obtención de las Juntas Receptoras del Voto más representativas usando *clustering*

Se escogieron las 25 Juntas Receptoras del Voto más representativas sin el uso de Weka. Para el efecto, se calculó el resultado global final de las elecciones a nivel nacional, además sumando los votos obtenidos por cada partido político: MIPAIS, PSP, PRIAN y OTROS, que fueron normalizados con respecto al total de votos emitidos para obtener los siguientes porcentajes:

MIPAIS: 0,52: ~ 52,4%
PSP: 0,29: ~ 29,9%
PRIAN: 0,95: ~ 9,5%
OTROS: 0,79: ~ 7.9%

Después, a través de la Distancia Euclidea se obtuvieron cada una de las Juntas Receptoras del Voto al resultado final de las elecciones, mediante la fórmula:

$$distancia = \sqrt{\begin{array}{c}((v\_mipais_i - v\_mipais_t)^2 + \\ (v\_psp_i - v\_psp_t)^2) + \\ (v\_prian_i - v\_prian_t)^2)\end{array}}$$

La distribución de las Juntas Receptoras del Voto en función de la Distancia Euclidea, el resultado final, puede observarse en la Figura 3.

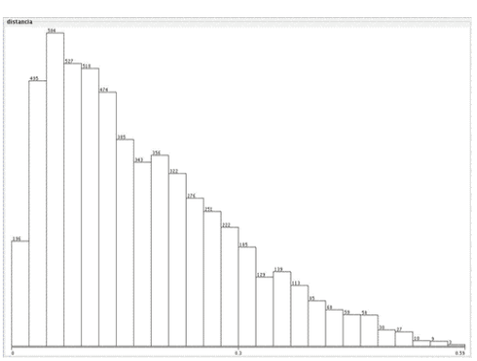

Figura 3. Distribución de las distancias del resultado de cada una de las JRV.

En la Figura 3, el algoritmo de *clustering* utilizado fue el K-MEDIAS, con el objetivo de seleccionar las Juntas Receptoras del Voto más parecidas al resultado final. Cuanto mayor sea el número de grupos elegidos, menor número de muestras aparecerán en cada uno de ellos. Se eligieron 20 *clusters*, mostrados en la Tabla 3:

Tabla 3. Clusters seleccionados.

Cluster	%	Cluster	%
0	63 (1%)	1	429 (7%)
2	362 (6%)	3	409 (7%)
4	159 (3%)	5	380 (6%)
6	311 (5%)	7	339 (6%)
8	236 (4%)	9	232 (4%)
10	382 (7%)	11	304 (5%)
12	352 (6%)	13	190 (3%)
14	262 (4%)	15	133 (2%)
16	318 (5%)	17	350 (6%)
18	366 (6%)	19	287 (5%)

En el *cluster* 13 se ubican aquellas Juntas Receptoras del Voto con resultados más cercanos en la Distancia Euclidea del resultado final. Este cluster agrupa un 3% del total, es decir 190 JRV, que podrían ser útiles para efectuar cualquier clase de sondeo. Si se selecciona un número inferior de JRV, como se observa en la Figura 5 y se muestra en la Tabla 3, se debe seleccionar un número mayor de *clusters*. Para el ejemplo, se utilizaron 100 *clusters*, se obtienen 25 JRV con una distancia de 0,992 en media y de 0,03 de la desviación estándar. Además Weka permitió obtener los centroides de cada *cluster* y ver las JRV.

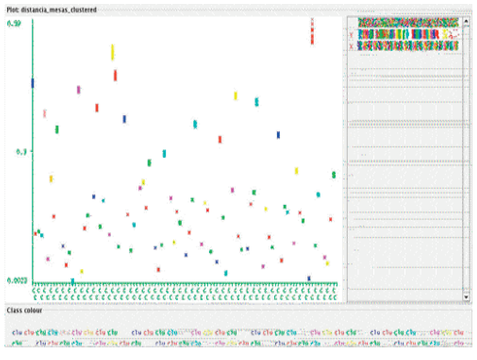

Figura 4. Centroides de los 100 *cluters* haciendo uso del algoritmo K-MEDIAS.

Al seleccionar el *cluster* de color azul en la Figura 4, que se encuentra en la parte de abajo del eje de las ordenadas, se observa que las JRV son las mismas que se buscan: 171, 121, 117, 107, 104, 102, 98, 87, 86, 80, 77, 69, 61, 53. Las cuales corresponden a las señaladas en la Tabla 4.

Tabla 4. 22 JRV mayormente representativas.

Nº mesa	Provincia	Cantón	Parroquia
171	Chimborazo	Riobamba	Lizarzaburo
121	Chimborazo	Riobamba	Velasco
117	Tungurahua	Ambato	La Matriz
107	Chimborazo	Riobamba	Maldonado
104	Cotopaxi	Pujili	Pujili
102	Cotopaxi	Saleedo	San Miguel de Salcedo
98	Cotopaxi	Latacunga	La Matriz
87	Pastaza	Pastaza	Puyo
86	Cotopaxi	La Mana	La Mana
86	Cotopaxi	Latacunga	Eloy Alfaro/San Felipe
86	Tungurahua	Ambato	La Merced
80	Tungurahua	Ambato	Huachi Chico
77	Tungurahua	Pelileo	Pelileo
69	Chimborazo	Riobamba	Veloz
61	Chimborazo	Guamote	Guamote
53	Chimborazo	Colta	Columba
53	Tungurahua	Ambato	Izamba

3.2. Obtención de las JRV más representativas realizando *clustering* directo sobre las muestras

A continuación se presenta otra forma de conseguir las 10 JRV representativas, mediante un agrupamiento directo, es decir sin calcular previamente las distancias. Para el mismo se utilizó el algoritmo K-Medias, que proporciona Weka y como entrada los porcentajes de voto por cada JRV para MIPAIS, PSP, PRIAN y OTROS. Al obtener un número de JRV representativas, se debe incrementar el número de clusters, para que existan menos JRV por *cluster*. Para evitar demoras en el tiempo de procesamiento, únicamente se utilizó MIPAIS y PSP como entradas, pues PRIAN y OTROS son menos significativos. Para concluir, una vez realizado el agrupamiento, se seleccionó el *cluster* con el centroide mas parecido al resultado final de las elecciones presidenciales y se tomaron como JRV con mesas representativas las asignadas a dicho *cluster.*

En primer lugar, a modo ilustrativo (salen demasiadas JRV por *cluster*) se muestra el resultado del algoritmo EM con 6 *clusters* para los atributos MIPAIS y PSP:

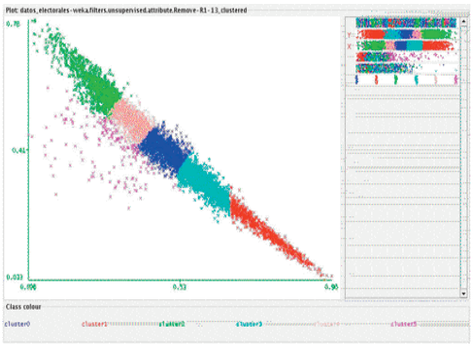

Figura 5. Agrupamiento de 6 conjuntos usando el algoritmo EM (Muestra MIPAIS y PSP).

En la Figura 5 se visualiza el resultado de las elecciones, que corresponde al *cluster* 3 de color azul, así como las muestras más atípicas en diferentes clusters. Utilizando K-medias para el mismo número *clusters*, se reduce el tiempo de ejecución; como resultado se obtuvieron regiones con límites más difusos.

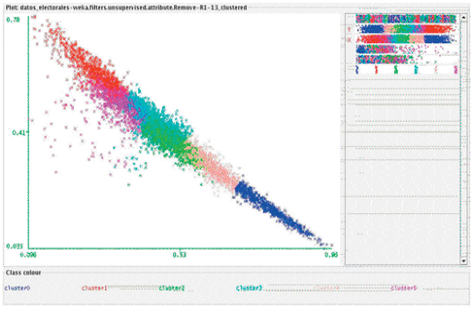

Figura 6. Agrupamiento de 6 conjuntos usando el algoritmo K-Medias (MIPAIS y PSP).

Al aumentar el número de *clusters*, se obtuvieron menos muestras (Figura 6).

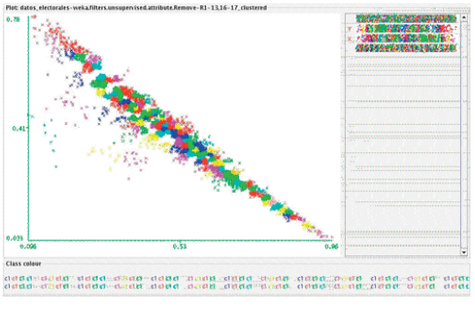

Figura 7. Agrupamiento de 6 conjuntos usando el algoritmo K-Medias (MIPAIS y PSP).

En la Figura 7 existen 115 JRV; en el *cluster* 31 en color rojo y en el segundo caso, el *cluster* 66, existen 193 casos. Usando el algoritmo K-Medias y 200 *clusters*, el proceso de ejecución se realiza en menor tiempo.

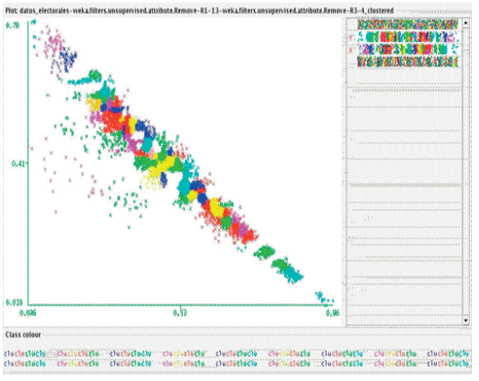

Figura 8. Agrupamiento de 200 conjuntos usando el algoritmo K-Medias (siendo las entradas = MIPAIS y PSP).

Finalmente, el *cluster* 13 fue el seleccionado, con 46 muestras.

4. Distribución del Voto

Para este estudio, la obtención de la JRV más representativa es el punto de partida para realizar la relación entre ganadores de cada JRV con el porcentaje de votos validados (votos validos/censo), así como la opción ganadora en función de la provincia. Para dichos efectos se utilizan dos nuevas variables "ganador" que toman los valores de MIPAIS y PSP y porcentaje_votos, que resulta de la división votos válidos sobre censo.

4.1. Distribución del voto en las provincias con mayor número de electores (Chimborazo y Tungurahua)

En la Figura 9, se muestran las relaciones entre diversas variables con el atributo ganador.

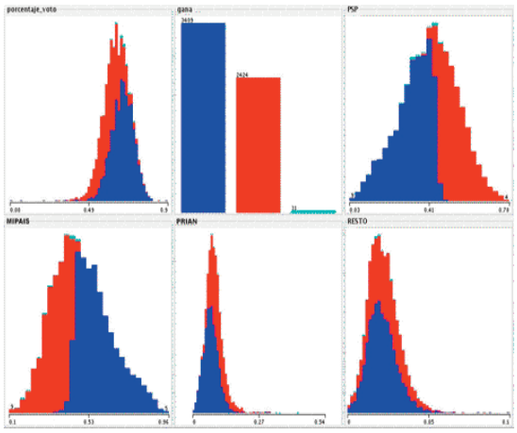

Figura 9. Histogramas de datos y distribución de porcentaje del voto, ganador por JRV y porcentaje de voto por junta para MIPAIS, PSP y PRIAN. El color azul corresponde con las mesas donde gana PSP, y el color rojo con las que gana MIPAIS y el azul claro con las que gana PRIAN.

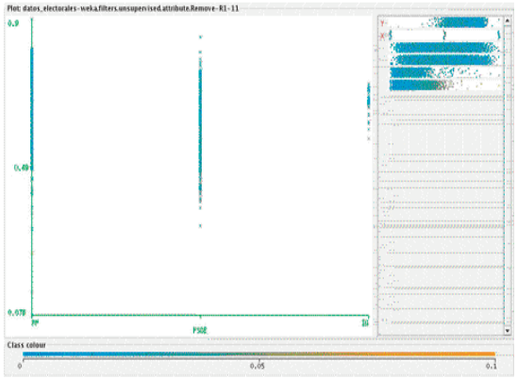

Figura 10. Distribución de las juntas ganadoras en función de porcentaje de votos de cada una.

En la Figura 10 se puede comprobar que MIPAIS necesita un porcentaje algo mayor en una junta para proclamarse ganador en la misma. Además parece que, el que gane una u otra opción política, puede estar relacionado con los ciudadanos que votan en dicha junta. Dicha información puede ser utilizada por el partido político al cual perjudique la abstención.

En vista a la relación entre el porcentaje del voto y el partido político que sale ganador de la JRV, se intentó construir un sistema automático que proporcione la opción ganadora de cada JRV.

Para realizar el Análisis de Distribución del Voto, se utilizó un árbol de decisión proporcionado por Weka, el *DecisionStrump*, obteniendo una sola entrada: el porcentaje de voto. Los arboles de decisión son muy intuitivos y fáciles de comprender.

La cuestión en mención, como era de esperarse, las Juntas Receptoras del Voto más representativas, están centradas en las provincias, cantón, parroquia mostradas en la Tabla 5:

Tabla 5. Provincias con JRV mayormente representativas.

Provincia	Cantón	Parroquia	JRV	Sufrag.
Chimborazo	Riobamba	Lizarzaburo	171	30.659
Chimborazo	Riobamba	Velasco	121	22.437
Tungurahua	Ambato	La Matriz	117	19.907
Cotopaxi	Pujili	Pujili	104	20.884
Pastaza	Pastaza	Puyo	87	20.883

Como se puede observar en la Tabla 5, en una sola provincia, Cotopaxi, la cabecera cantonal (capital) no tiene mayor representación de JRV. Además estos resultados despiertan gran interés de empresas especializadas en sondeos a “boca de urna”, que si bien logra ser una estrategia para conocer resultados parciales, apunta obtener de aquellas JRV mayoritarias de forma específica una buena cantidad de información. Estos resultados permitirán ir afianzando para próximas elecciones.

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 algún movimiento o partido político, determinará efectuar campañas con mayor éxito, sobre este electorado que será decisivo en las elecciones.

Los resultados evidentes demuestran que el Partido Sociedad Patriótica PSP, tiene su feudo y electorado ganado en esta región principalmente en la provincia de Tungurahua con (109.974) votantes por el PSP, seguido por Chimborazo con (93.609) electores y Cotopaxi y Pastaza menos de (12.000) votantes. Sin embargo el partido de gobierno de turno Movimiento Pa-

tria Altiva I Soberana MIPAIS, tiene preferencia la provincia de Cotopaxi (89.628) votantes y Pastaza (11.959) electores para esa tienda política, determinado un crecimiento para futuras elecciones. En cuanto al Partido Renovador Institucional Acción Nacional PRIAN, a pesar de su liderazgo desde pasadas elecciones, no logra superar a los dos partidos políticos antes mencionados.

Los resultados y discusión en mención son los más analizados por movimientos y partidos políticos; así como empresas de marketing político, por cuanto son el termómetro para futuras elecciones. Así como para la toma de decisiones para focalizar en un sitio (provincia, cantón o parroquia), que marque una tendencia electoral para un partido político en concreto.

5. Conclusiones

El análisis realizado para los datos electorales puede ser útil, primero para realizar sondeos en elecciones posteriores, donde se puede tener la certeza de gozar de cierta fiabilidad y por concentración un ahorro en costos. Por otro lado, la relación entre porcentaje de voto y la opción ganadora de cada JRV, puede ser útil para realizar valoraciones a lo largo de la jornada electoral y, sobre todo, justo al término de la misma, cuando el porcentaje del voto puede ser calculado directamente.

Los resultados obtenidos en esta investigación son confiables y para futuras elecciones servirán de base los algoritmos utilizados. Como ya existe una base de datos el cual sirvió para la explotación, solo restaría agregar datos y variables de las elecciones presidenciales futuras.

6. Agradecimientos

La investigación realizada es un esfuerzo en el cual directa o indirectamente han participado varias personas e instituciones como el Consejo Nacional Electoral, la Universidad Indoamérica con su Facultad de Ingeniería en Sistemas y el Instituto de Investigación, Desarrollo e Innovación (IDI-UTI).

Gracias también a mis queridos estudiantes que pusieron su granito de arena en parte de esta investigación. Así mismo se agradece a los revisores anónimos, amigos y compañeros.

7. Referencias

[1] Hernández, J. y Ramírez, M. J. 2004. *Instrucción a la Minería de Datos*. Madrid, España: Prentice Hall.

[2] Alarcón, V. F. 2006. *Desarrollo de sistemas de información: Una Metodología basada en el modelado*. Catalunya, España: Educaciones UPC.

[3] González, Luis. 2009. *Construyendo la democracia, Elecciones Presidenciales* 1979-2010. Quito, Ecuador: Consejo Nacional Electoral.

[4] Antony, S., and Murray, Dennis. 1998. *Data Warehousing in the Real World*. Paris, Francia: Addison-Wesley.

[5] Fraley, W. J., Piatetsky-Shapiro, G. and Matheus, C. J. 1992. *Knowledge Discovery in Databases*. And Overview.

[6] Gill, H. 2009. *Data Warehousing*. Mexico DF, México: Prentice Hall.

[7] Johnson, R. A., Rosenzweing, J. E. 1996. *Teoría, Integración y Administración de Sistemas*. New York, Estados Unidos de América: McGraw Hill.

[8] Jiawei han, M. K. 2006. *Data Mining: Concept and Techniques*. California, Estados Unidos: Morgan Kaufmann.

[9] Lummis, D. 2002. *Democracia radical*. Madrid, España: Editorial Siglo XXI.

[10] Valles, J. y Bosch, A. 2007. *Sistemas electorales y gobiernos representativos*. Madrid, España: Editorial Ariel.

[11] David, I. C., and Gareis, R. 2006. *Global project management handbook*. Nueva York, Estados Unidos: McGraw-Hill Professional.