



Universidad de la Sierra Sur

Análisis exploratorio del fenómeno del acceso a la información pública gubernamental federal de México, usando métodos de ciencia de datos: periodo 2003 a 2015

TESIS

**Para obtener el título de:
Maestro en Gobierno Electrónico**

Presenta:

Hermelando Cruz Pérez

**Bajo la dirección del
Dr. Sergio Coria Olguín**

Tesis desarrollada por el Ing.Hermelando Cruz Pérez, egresado de la maestría en Gobierno Electrónico, bajo la dirección del comité Tutorial:

Director: Dr. Sergio Rafael Coria Olguín.

Asesor: Dr. Christian Arturo Cruz Meléndez.

Asesor: Dr. Guadalupe Gabriel Durán Férman.

Tesis presentada en Examen Profesional el 24 de septiembre de 2019, ante el siguiente Jurado:

Presidente: Dr. Víctor Alberto Gómez Pérez

Secretario: Mtra. Nina Martínez Cruz

Vocal: Dr. Sergio Rafael Coria Olguín

Suplente:Dr. Diego Soto Hernández

Suplente:Dr. Arisaí Darío Barragán López

AGRADECIMIENTOS

En primer lugar expreso mi agradecimiento al director de tesis, Dr. Sergio Coria Olguín, por el apoyo, paciencia y confianza brindadas en este trabajo, a mi comité tutorial el Dr. Christian Arturo Cruz Meléndez y el Dr. Guadalupe Gabriel Durán Férman por su apoyo y disponibilidad. A mis padres y hermanos quienes a lo largo de toda mi vida han apoyado y motivado mi formación académica, creyeron en mi en todo momento y no dudaron de mis habilidades.

RESUMEN

El acceso a la información pública gubernamental en México es un fenómeno relativamente reciente y poco investigado; quizá, aún menos que el fenómeno de la transparencia, con el cual llega a confundirse. Por ello, el objetivo de esta investigación es analizar y descubrir patrones relevantes en las características de las solicitudes de información pública federal, de la ubicación de los solicitantes y en las dependencias gubernamentales accedidas. El objeto de estudio específico es la base de datos de las solicitudes generadas en el período de 2003 a 2015, dirigidas al gobierno federal mediante la plataforma informática que está a cargo del Instituto Nacional de Acceso a la Información y Protección de Datos Personales (INAI). Esta base contiene aproximadamente veinte atributos, correspondientes a más de un millón 200 mil solicitudes. La metodología de investigación propuesta tiene enfoque cuantitativo, con alcance descriptivo y exploratorio. Las técnicas aplicadas son: estadística descriptiva y aprendizaje automático. De este último, se usa específicamente un algoritmo ya conocido para generar automáticamente árboles clasificadores, con el fin de descubrir y representar las interacciones entre los atributos de las solicitudes. Una aportación sustantiva de esta investigación es el hallazgo de un conjunto de patrones de tipos uni-variable y multi-variable en estos atributos. Algunos de los patrones más generales son: 1) la cantidad anual de solicitudes presenta una tendencia ascendente, y 2) los sectores gubernamentales que han recibido la mayoría de las solicitudes son: seguridad social, hacienda y crédito público, y educación pública. Las entidades federativas que han tenido las mayores tasas de solicitudes por cada cien mil habitantes son: Distrito Federal (CDMX), Morelos y Estado de México. Se presentan también otros patrones más específicos. Además de los patrones hallados, que constituyen una contribución teórica, otra aportación es la propuesta metodológica aplicada para descubrirlos y representarlos. Adicionalmente, el cálculo de tasas de solicitudes por cada cien mil habitantes ofrece una perspectiva complementaria para el entendimiento del fenómeno. En forma práctica, los resultados pueden aprovecharse para revisar y mejorar leyes, políticas, trámites y procesos de acceso a la información pública mediante el uso de las tecnologías de información y comunicación.

Palabras clave: gobierno abierto, gobierno electrónico, minería de datos, árboles clasificadores.

ABSTRACT

Access to federal government public information in Mexico is a relatively recent and understudied phenomenon; perhaps, it has been less investigated than the transparency phenomenon, with which it can be confused. Therefore, the aim of this research is to analyze and discover significant patterns in the features of: federal public information requests, location of requesters, and referred agencies. The specific subject matter of the study is the database of requests generated in the period from 2003 to 2015, addressed to the federal government by means of the computing platform in charge of the Mexican Institute for Information Access and Personal Data Protection (INAI). This database contains, approximately, twenty attributes, corresponding to over one million two hundred thousand requests. The proposed research methodology has a quantitative approach with descriptive and exploratory scope. The applied techniques include: descriptive statistics and machine learning. In regards to the latter a known algorithm for automatically generating classification trees is specifically used to discover and represent interactions among the request attributes. A significant contribution from this research is the finding of a set of uni-variate and multi-variate patterns in these attributes. A number of the most general patterns are: 1) the annual quantity of requests shows a rising trend, and 2) the government sectors receiving most of the requests are: social security, treasury and public credit, and public education. The States having the highest rates of requests per one hundred thousand inhabitants are: Federal District (CDMX), Morelos and State of Mexico. Other more specific patterns are also presented. In addition to the discovered patterns, which are a theoretical contribution, another contribution is the methodological proposal applied for discovering and representing them. Besides, the calculation of rates of requests per one hundred thousand inhabitants offers a complementary perspective for understanding the phenomenon. A practical application of the results can be to use them as input for reviewing and improving laws, policies, procedures and processes of access to public information by means of information and communication technologies.

Key words: open government, electronic government, data mining, classification trees.

ÍNDICE

INTRODUCCIÓN	1
CAPÍTULO 1. PLANTEAMIENTO DEL PROBLEMA	3
1.1 Acerca del problema de investigación	3
1.2 Definición del problema de estudio	5
1.3 Descripción del objeto de estudio	6
1.4 Delimitación espacial y de tiempo.....	9
1.5 Supuestos de investigación	10
1.6 Preguntas de investigación	10
1.6.1 Pregunta principal	10
1.6.2 Preguntas secundarias	10
1.7 Objetivos de investigación	11
1.7.1 Objetivo general.....	11
1.7.2 Objetivos específicos	11
1.8 Posibles patrones a encontrar en la base de datos del IFAI-INAI.....	12
1.9 Justificación de la investigación.....	14
CAPÍTULO 2. ANTECEDENTES Y TRABAJOS RELACIONADOS.....	16
2.1 Antecedentes del acceso a la información.....	16
2.2 Ley General de Transparencia y Acceso a la Información Pública.....	28
2.3 Portal Electrónico de Transparencia	30
2.4. Trabajos de investigación relacionados.....	34
2.4.1 Sobre acceso a la información gubernamental en México.....	34
2.4.2 Sobre existencia y aprovechamiento de datos abiertos.....	37
2.4.3 Sobre estadística y aprendizaje automático en datos gubernamentales	38
2.4.4 Hallazgos rescatables de los trabajos relacionados	41
CAPÍTULO 3. MARCO TEÓRICO-CONCEPTUAL	43

3.1 Gobierno electrónico (GE)	43
3.1.2 Beneficios del gobierno electrónico	48
3.1.3 Clasificación del gobierno electrónico.....	50
3.1.4 Etapas del Gobierno Electrónico	50
3.2 Gobierno Abierto.....	52
3.2.1 Alianza para el Gobierno Abierto	54
3.2.2 Los pilares del gobierno abierto	56
3.2.3 Rendición de cuentas	56
3.2.4 Datos abiertos.....	57
3.2.5 Transparencia y acceso a la información pública	58
3.3 Conceptos y técnicas de informática y computación	60
3.3.1 Base de datos.....	60
3.3.2 Descubrimiento de conocimiento en bases de datos.....	62
3.3.3 Minería de datos	62
3.3.4 Modelos estadísticos	63
3.3.5 Aprendizaje automático con árboles de decisión.....	64
CAPÍTULO 4. METODOLOGÍA.....	65
4.1 Enfoque cuantitativo	65
4.2 Alcance	65
4.3 Metodología de análisis del INAI	67
4.4. Fuentes de datos	68
4.4.1 Base de datos del INAI	68
4.4.2 Datos del INEGI: Censo 2010 y Encuesta Inter-Censal 2015.....	68
4.5 El Método KDD	69
4.6 Método CRISP-DM	71
4.7 Obtención de la Información	72

4.8 Limpieza de los datos	72
4.9 Técnicas de estadística descriptiva uni-variable	73
4.9.1 Desviación estándar	74
4.9.2 Diagrama de barras o de columnas	74
4.9.3 Histograma	76
4.9.4 Análisis de Pareto para variable cuantitativa	79
4.9.5 Análisis de Pareto para variable nominal	80
4.9.6 Análisis de tendencia con regresión lineal simple	82
4.9.7 Cálculo de tasa, razón o proporción	84
4.10 Técnicas de estadística descriptiva bi-variable	85
4.10.1 Matriz de análisis de correlación para variables nominales	85
4.11 Visualización sobre mapas digitales	88
4.12 Análisis multi-variable con aprendizaje automático	90
4.12.1 Estructura e interpretación de árboles	90
4.12.2 Criterios de aceptación de árboles clasificadores	95
CAPÍTULO 5. RESULTADOS	99
5.1 Resultados estadísticos globales del periodo 2003 a 2015	99
5.1.1 Análisis de Pareto de las solicitudes de información	100
5.1.2 Gráfica de tendencia de la cantidad de solicitudes	101
5.1.3 Análisis de Pareto de los sectores gubernamentales	101
5.1.4 Análisis de Pareto de las dependencias gubernamentales	104
5.1.5 Análisis de Pareto por entidad federativa	105
5.1.6 Análisis de Pareto de los municipios de los solicitantes	107
5.1.7 Análisis de Pareto de la cantidad de solicitudes por mes	109
5.1.8 Análisis de Pareto de días de la semana de las solicitudes	110
5.1.9 Análisis de Pareto de las horas en que se generaron solicitudes	110

5.1.10	Análisis de Pareto de los medios de entrega.....	112
5.1.11	Análisis de Pareto de los tipos de respuesta de las solicitudes	113
5.1.12	Fechas con las mayores cantidades de solicitudes de información.....	114
5.2	Resultados estadísticos por año del periodo 2004 a 2015	115
5.2.1	Gráficas de tendencia	115
5.2.2	Tasas de solicitudes por cada 100 mil habitantes	118
5.2.3	Rankings anuales de las tasas de solicitudes en el periodo 2004 a 2015 ..	121
5.2.4	Análisis de Pareto de sectores gubernamentales por año.....	122
5.2.5	Rankings global de porcentajes de los sectores gubernamentales	126
5.3	Resultados de árboles clasificadores producidos con el algoritmo J4.8	128
5.3.1	Desempeño de los modelos de árbol del periodo global 2003 a 2015	129
5.3.1.1	Patrones de la variable target tipo de respuesta de 2003 a 2015.....	130
5.3.2	Desempeño de los árboles clasificadores de solicitudes por cada año	131
5.3.3	Patrones de la variable target tipo de respuesta del año 2015.....	133
5.3.4	Patrones de la variable target tipo de respuesta del año 2012.....	135
5.3.5	Patrones de la variable target entidad federativa del año 2010.....	137
CAPÍTULO 6. CONCLUSIONES Y RECOMENDACIONES.....		139
6.1	Contribución.....	140
6.2	Recomendaciones	141
6.2.1	Al sector público	141
6.2.2	Al sector empresarial	142
6.2.3	Al sector social.....	143
6.3	Trabajo de investigación futuro.....	143
REFERENCIAS.....		145

ÍNDICE DE FIGURAS

Figura 1. Palabras más utilizadas en diversas definiciones de gobierno electrónico.....	47
Figura 2. Representación de un árbol de decisión.	64
Figura 3. El proceso de Descubrimiento de Conocimiento en Bases de Datos (KDD).	70
Figura 4. Fórmula para el cálculo de la desviación estándar.....	74
Figura 5. Ejemplo de diagrama de columnas.	75
Figura 6. Ejemplo de diagrama de barras.....	75
Figura 7. Fórmula para calcular el tamaño óptimo de intervalo de una variable cuantitativa.	76
Figura 8. Ejemplo de sustitución de valores en la fórmula de Sturges.....	77
Figura 9. Ejemplo de histograma de tiempo de respuesta de las solicitudes.	79
Figura 10. Ejemplo de gráfica de análisis de Pareto: tiempo de respuesta de las solicitudes.	80
Figura 11. Ejemplo de diagrama de Pareto: cantidad de solicitudes por mes del año 2005.	82
Figura 12. Fórmula de la regresión lineal simple.....	82
Figura 13. Fórmula del coeficiente de correlación de Pearson.....	83
Figura 14. Ejemplo de regresión lineal de la cantidad de solicitudes diarias en el año 2005.....	84
Figura 15. Ejemplo de visualización de datos sobre mapa digital.	90
Figura 16. La interfaz del software WEKA mostrando un dataset de ejemplo.	92
Figura 17. Ejemplo de un árbol clasificador para la variable target tipo de respuesta.	94
Figura 18. Fórmula para calcular el estadístico Kappa (K).....	96
Figura 19. Gráfica de tendencia, de la cantidad anual de solicitudes en el periodo 2003-2015.....	101
Figura 20. Gráfica de tendencia, de la cantidad de solicitudes diarias en el año 2004.....	116
Figura 21. Gráfica de tendencia, de la cantidad de solicitudes diarias en el año 2009.....	117
Figura 22. Gráfica de tendencia, de la cantidad de solicitudes diarias en el año 2015.....	118

ÍNDICE DE TABLAS

Tabla 1. Atributos en la base de datos de IFAI-INAI.....	6
Tabla 2. Posibles patrones bi-variable a encontrar en la base de datos del IFAI-INAI.....	13
Tabla 3. Atributos de la base de datos de las solicitudes de información del IFAI-INAI.	72
Tabla 4. Variables producidas a partir de la base de datos del IFAI-INAI.	73
Tabla 5. Ejemplo de información para elaborar histograma de tiempo de respuesta.....	78
Tabla 6. Ejemplo de tabla de Pareto para atributo nominal.....	81
Tabla 7. Ejemplo de cálculo de tasa: cantidad de solicitudes por cada cien mil habitantes.....	85
Tabla 8. Ejemplo de matriz de análisis de correlación entre variables nominales.....	86
Tabla 9. Ejemplo de matriz de análisis de correlación entre variables nominales.....	87
Tabla 10. Ejemplo de creación de intervalos de las tasas de solicitudes.....	89
Tabla 11. Análisis de Pareto de la cantidad anual de solicitudes de 2003 a 2015.	100
Tabla 12. Análisis de Pareto de sector gubernamental de 2003 a 2015.	103
Tabla 13. Análisis de Pareto de las dependencias gubernamentales de 2003 a 2015.....	105
Tabla 14. Análisis de Pareto de entidad federativa de las solicitudes de 2003 a 2015.	107
Tabla 15. Análisis de Pareto de la cantidad de solicitudes por municipio de 2003 a 2015.	108
Tabla 16. Análisis de Pareto de la cantidad mensual de solicitudes de 2003 a 2015.....	109
Tabla 17. Análisis de Pareto de la cantidad de solicitudes por día de la semana de 2003 a 2015. ...	110
Tabla 18. Análisis de Pareto de hora del día de generación de solicitud de 2003 a 2015.....	111
Tabla 19. Análisis de Pareto del medio de entrega de la información de 2003 a 2015.....	112
Tabla 20. Tipo de respuesta a las solicitudes en el periodo 2003 a 2015.....	113
Tabla 21. Fechas con las mayores cantidades de solicitudes de información por año.....	115

Tabla 22. Tasas de solicitudes por cada 100 mil habitantes por entidad federativa en 2004.....	119
Tabla 23. Tasas de solicitudes por cada 100 mil habitantes por entidad federativa en 2009.....	120
Tabla 24. Tasas de solicitudes por cada 100 mil habitantes por entidad federativa en 2015. ..	121
Tabla 25. Rankings anuales: tasas de solicitudes por cada 100 mil habitantes de 2004 a 2015.	122
Tabla 26. Análisis de Pareto de los sectores gubernamentales en 2004.	123
Tabla 27. Análisis de Pareto de los sectores gubernamentales que recibieron solicitudes en 2009 ..	124
Tabla 28. Análisis de Pareto de los sectores gubernamentales que recibieron solicitudes en 2015...	125
Tabla 29. Rankings de solicitudes recibidas por sector gubernamental de 2004 a 2015.....	127
Tabla 30. Desempeño general de tres árboles clasificadores del periodo 2003-2015.....	130
Tabla 31. Desempeño general de trece árboles clasificadores del periodo 2003-2015.....	132

LISTA DE ABREVIATURAS

Aguascalientes	Ags
Baja California	BC
Baja California Sur	BCS
Campeche	Camp
Chiapas	Chih
Chihuahua	Chis
Coahuila	Coah
Colima	Col
Ciudad de México	DF
Durango	Dgo
Guanajuato	Gro
Guerrero	Gto
Hidalgo	Hgo
Jalisco	Jal
México	Mex
Michoacán	Mich
Morelos	Mor
Nayarit	Nay
Nuevo León	NL
Oaxaca	Oax
Puebla	Pue
Querétaro	QR
Quintana Roo	Qro
San Luis Potosí	Sin
Sinaloa	SLP
Sonora	Son
Tabasco	Tab
Tamaulipas	Tamps
Tlaxcala	Tlax
Veracruz	Ver
Yucatán	Yuc
Zacatecas	Zac

INTRODUCCIÓN

El presente trabajo tiene como principal objetivo analizar las solicitudes de información pública gubernamental, contenidas en la base de datos del Instituto Nacional de Transparencia Acceso a la Información y Protección de Datos Personales (INAI), correspondientes al periodo 2003 a 2015, aplicando técnicas de minería de datos, que incluyen estadística descriptiva y aprendizaje automático. Junto con ello, se propone una metodología para el análisis de la base de datos.

La importancia de estudiar este tema radica en que el acceso a la información pública federal en México es un tema de interés en el área de conocimiento del gobierno electrónico, específicamente respecto a la relación entre gobierno y ciudadano. El motivo es que recientemente se ha hecho uso de las plataformas digitales denominadas Sistema Informatizado de Solicitudes de Información (SISI) e Infomex para automatizar y agilizar el proceso de acceso a la información pública de las dependencias gubernamentales. Además, la información en formatos abiertos, generada por el gobierno, se ha aprovechado poco debido al desconocimiento del potencial de las técnicas y herramientas de la minería de datos que permiten encontrar patrones y comportamientos relevantes.

Al descubrir patrones estadísticos en la base de datos de solicitudes de información procesadas por el INAI, se puede generar conocimiento innovador acerca del comportamiento de los solicitantes y de los otorgantes de información gubernamental federal. Este conocimiento le sirve al propio gobierno, a la comunidad de investigación, al sector empresarial y a la sociedad en general. Se puede generar un análisis más profundo del fenómeno del acceso a la información y recomendar acciones de política pública para mejorar e incrementar el acceso a la información, contribuyendo la rendición de cuentas.

Esta tesis consta de seis capítulos, como se describe a continuación. En el primer capítulo, se plantea la problemática que dio origen a esta investigación, la descripción del objeto de estudio, las preguntas de investigación, los objetivos y la justificación. El segundo capítulo aborda los antecedentes de la investigación, donde se analizan algunos trabajos previos relacionados con aspectos de metodología. El tercer capítulo se dedica al marco teórico, que incluye principalmente conceptos de gobierno electrónico, gobierno abierto, acceso a la información pública, datos abiertos, minería de datos y modelos estadísticos. El cuarto capítulo describe la metodología, el enfoque de la investigación y las técnicas de análisis de datos. En el quinto capítulo se presentan los resultados, incluyendo los basados en estadística descriptiva y los basados en aprendizaje automático. Finalmente, en el capítulo sexto se presentan las conclusiones, se proponen algunas recomendaciones de política pública y se sugiere trabajo de investigación a futuro.