Fecha de Recepción:5 de diciembre de 2019
Fecha de Aceptación: 22 de enero de 2020

06

Diseño de un Almacén de Datos utilizando metodología HEFESTO. Caso de estudio: “Divorcios del año 2016 en el Ecuador”

Est. Ximena Orellana
Mgs. Leopoldo Pauta

Instituto Superior Tecnológico Particular Sudamericano

Resumen

En esta cambiante era digital en la que se producen millones de datos constantemente es necesario desarrollar herramientas que permitan el análisis de la información y faciliten el acceso a un conocimiento que guíe positivamente la toma de decisiones. La presente investigación muestra cómo pueden ser útiles las nuevas tecnologías de análisis de información con almacenes de datos cuando se precisa información concreta, objetiva y veraz; así pues, el presente artículo se enfoca en Knime y Exaplus como alternativas tecnológicas utilizadas para obtener información. Para ello, el presente trabajo toma como caso de estudio las bases de datos a nivel nacional referentes a Divorcios en el Ecuador (año 2016), publicadas por el INEC, y se concentra en la información relacionada a divorcios de mujeres extranjeras y hombres ecuatorianos, estableciendo de antemano determinados rangos de edad (de 37 a 39 años) y tomando en consideración la edad del matrimonio al momento de darse el divorcio. Finalmente, es importante señalar que los datos que se derivaron de la investigación se plasmaron en herramientas visuales como Tableu.

Palabras Clave:

Base de Datos (BD), Almacenes de Datos, Analítica de datos, Knime, Tableu, Exaplus, Divorcios 2016 Ecuador-INEC

ABSTRACT

In this changing digital age in which millions of data are constantly being produced it is necessary to develop tools to allow the analysis of information and facilitate access to knowledge that positively guides to making correct decisions. This research shows how useful are the new information analysis technologies with data warehouses when it requires concrete, objective and truthful information; this article focuses in Knime and Exaplus as technological alternatives used to obtain information. The present work takes as a study case the databases from the national level of divorces in Ecuador (year 2016), published by the INEC, and focuses on information related to foreign women's divorces and Ecuadorian men, establishing in advance certain ranges of age (from 37 to 39 years) and taking into consideration the age of marriage at the time of divorce. Finally, it is important to mention that the data obtained from this research was applied in visual tools such as Tableu.

Keywords:

Database (bd), Data Warehouses, Data Analytics,
Knime, Tableu, Exaplus, Divorcios 2016 Ecuador-INEC

Introducción

Entre el 2006 y 2016, en el Ecuador, los divorcios se incrementaron en un 83. 45%, al pasar de 13 981 a 25 648, según cifras registradas en la base de datos “Estadísticas Vitales: Matrimonios y Divorcios”, publicada por el Instituto Nacional de Estadística y Censos (INEC, 2017). ¿Cuál es la razón de ese incremento? En la búsqueda de patrones que ayuden a explicar tal desarrollo, la presente investigación utilizó almacenes de datos que permitieran el análisis correspondiente; con su uso, se obtuvieron resultados interesantes, que pueden ser tomados como referencia para estudios posteriores.

Ahora bien, la creación del almacén de datos se dio siguiendo la metodología HEFESTO, que cuenta ya con pasos detallados para el análisis y la construcción de un data warehouse. Asimismo, es importante señalar las facilidades que brindaron distintas herramientas para tal propósito. Así, la herramienta Power Desing, permitió obtener un modelo dimensional adecuado gracias al cual se pudo realizar el proceso de carga de datos ETL. Hecho esto, estos mismos datos alimentaron las tablas de la bodega de datos, a través de la herramienta de integración de datos etl Knime. Finalmente, se utilizó Tableau para generar, tanto un análisis de la información almacenada en la bodega de datos implementada como una visualización de los reportes generados.

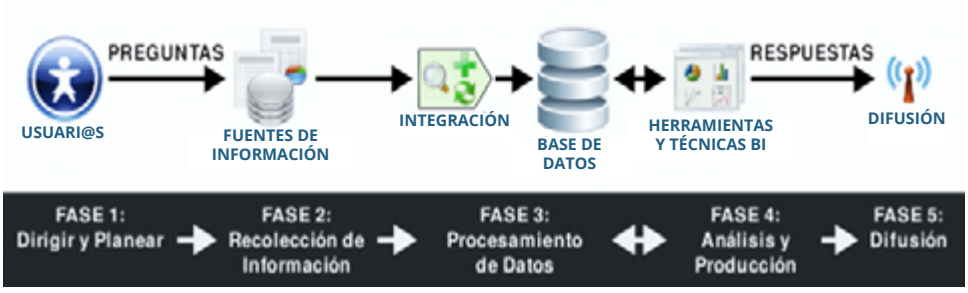
Marco teórico

Ante todo, es necesario señalar algunos aspectos inherentes al BI, aquel conjunto de modelos matemáticos y metodologías de análisis en las que se sustenta la presente investigación, y que se caracteriza por tener en cuenta los datos disponibles de un proceso o negocio para generar conocimiento e información que sirva en la toma de decisiones efectivas para la empresa (Farooqui y Mehra, 2018). A continuación, enlistamos algunos de esos aspectos:

Proceso de BI

El proceso de Business Inteligence (por sus siglas: BI) comprende métodos de recolección y tratamiento de información que posibilitan tomar decisiones oportunas, basadas en datos verídicos y eficaces. Y es que, al poder acceder a este tipo de información, y contrastarla, se puede identificar y corregir posibles soluciones, antes de que estas se efectivicen y se conviertan en problemas.

Figura 1. Proceso BI



Fuente: elaboración de los autores.

El proceso de BI se encuentra dividido en cinco fases (tal y como se puede observar en la figura 1. A continuación la describimos siguiendo a Bernabeu, D. y García M. (2017, pp. 122-163).

- en la **Fase 1** (dirección y planeación) se recolectan los requerimientos de información, se analizan las diversas necesidades y se generen las interrogantes fundamentadas en los datos que ayudarán a alcanzar los objetivos;
- en la **Fase 2** (recolección de información), se trabaja con las distintas fuentes de información y se realiza el proceso de extraer de ellas los datos requeridos;
- en la **Fase 3** (procesamiento de datos) se cargan e integran los datos en crudo en un formato que sea utilizable en el análisis posterior. En esta fase, se agrega la información recabada a una base de datos ya existente o bien se la registra en una nueva;
- en la **Fase 4** (análisis y producción) se emplean herramientas y técnicas propias de la tecnología bi, con las que se obtendrán las respuestas a las preguntas planteadas en nuestro caso de estudio, mediante la creación de reportes, etc.;
- finalmente, en la **Fase 5** (difusión) se presentan los resultados obtenidos a los usuarios, de manera sencilla.

Data Warehouse

Un Data Warehouse (DW) es una base de datos corporativa que se caracteriza por integrar y depurar información de una o más fuentes distintas, para luego procesarla y analizarla desde infinidad de perspectivas y con gran velocidad de respuesta. La creación de un Data Warehouse representa, en la mayoría de las ocasiones, el primer paso, desde el punto de vista técnico, para implantar una solución completa y fiable de Business Inteligencie (Sinnexus, 2014).

Según Savanur, S. y Shreedhara, K.:

> ***un almacén de datos está compuesto por dos fuentes de datos: una interna y una externa; en la primera se encuentran: datos operacionales, de clientes, de manufactura y datos históricos; y en la segunda, se encuentran datos de variables que afectan el desempeño de la compañía pero que no se generan dentro de esta, es decir datos externos (2016).***

Por otra parte, desde el punto de vista de la eficiencia en el uso de la información, un almacén de datos puede estar compuesto de mercados de datos –que son, por definición, aquellos subconjuntos de datos de mayor utilidad en la toma de decisiones (Savanur, S. y Shreedhara, K., 2016)–.

Metodología para la construcción

Por su parte, la metodología HEFESTO, disponible bajo licencia GNU FDL, se fundamenta en una investigación amplia, en la comparación de metodologías existentes y en experiencias propias de procesos de confección de almacenes de datos (Bernabeu, D. y García M., 2017, pp. 122-163). Esta metodología cuenta con las siguientes características:

- los objetivos y resultados esperados en cada fase se distinguen fácilmente y son sencillos de comprender;
- al basarse en los requerimientos del usuario, su estructura es capaz de adaptarse con facilidad y rapidez ante los cambios del negocio;
- ya que involucra al usuario final en cada etapa para que tome decisiones respecto al comportamiento y a las funciones del almacén de datos, reduce su resistencia al cambio;
- utiliza modelos conceptuales y lógicos, sencillos de interpretar y analizar;
- es independiente del tipo de ciclo de vida que se emplee para contener la metodología;
- es independiente, tanto de las herramientas que se utilicen para su implementación, de las estructuras físicas que contenga el almacén de datos, así como de su respectiva distribución;
- con la culminación de casa fase, los resultados obtenidos se convierten en el punto de partida del paso siguiente;
- es aplicable tanto para almacén de datos como para Data Mart (Sinnexus, 2014);
- por lo demás, la ventaja de la aplicación de esta metodología es que, a diferencia de otras, especifica puntualmente los pasos que se deben realizar en cada fase (Brizuela, L. y Castro, Y., 2013).

Desarrollo

Ahora bien, para el análisis de la base de datos Divorcios del Ecuador en el año 2016 se empezó por identificar variables que guiaran la búsqueda de patrones de información; las variables planteadas para este estudio fueron las siguientes:

- divorcio de mujeres extranjeras, considerando determinado rango de edad, y tomando como referencia la edad de su matrimonio;
- divorcio de hombres de nacionalidad ecuatoriana, considerando determinado rango de edad, y tomando como referencia la edad de su matrimonio.

Al realizar una síntesis, los indicadores fueron:

- el número de divorcios;

y las perspectivas de análisis:

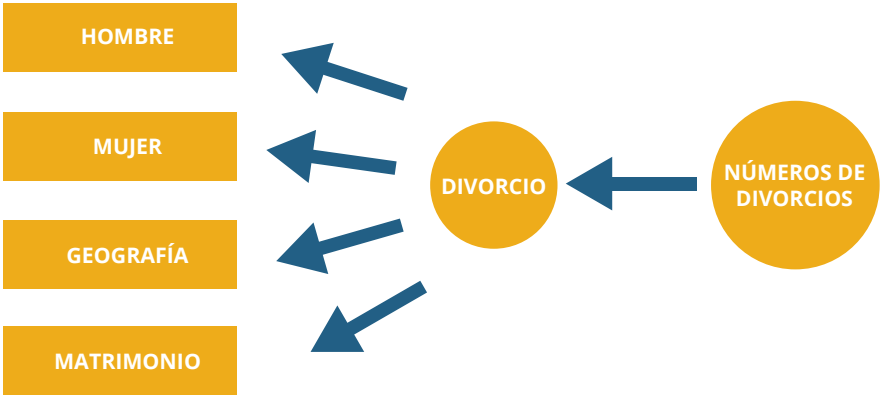
- hombre;
- matrimonio;
- mujer.

En este sentido, es indispensable insistir en que: es a partir de las interrogantes planteadas, a manera de variables, a la base de datos que se encuentra la información requerida, en este caso, información referente tanto al número de divorcios de mujeres extranjeras –considerando determinado rango de edad y tomando como referencia la edad de su matrimonio–; como al número de divorcios de hombres de nacionalidad ecuatoriana –considerando determinado rango de edad y tomando como referencia la edad de su matrimonio–.

Pues bien, a continuación, se señalan los pasos efectuados para la construcción de un almacén de datos HEFESTO.

Modelo Conceptual: a partir de los indicadores, se ha planteado el siguiente modelo conceptual (figura 2):

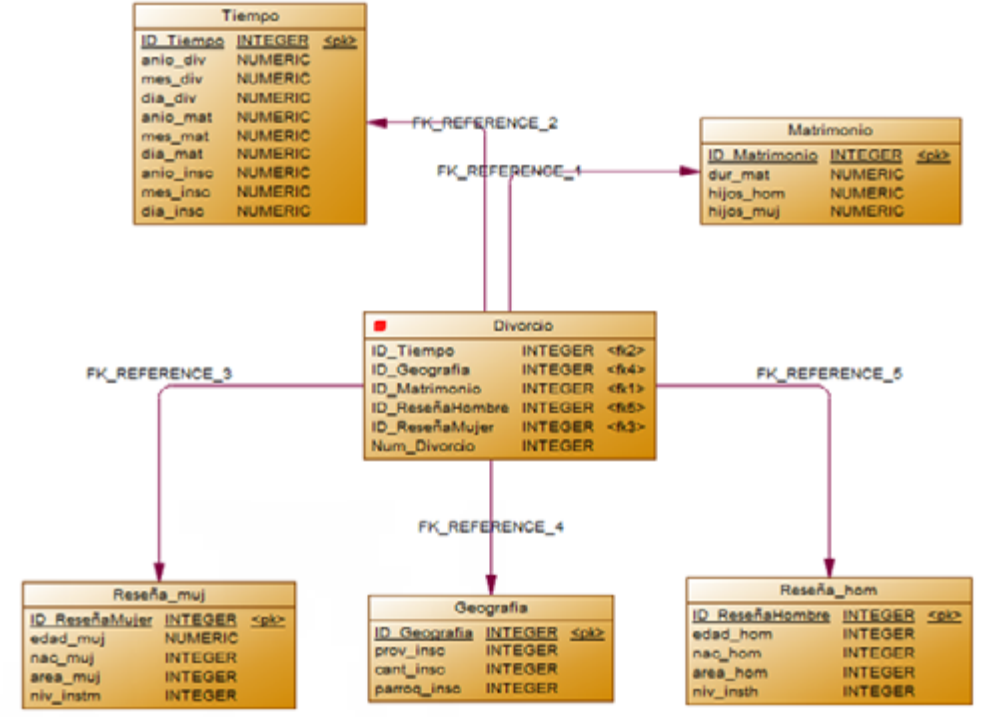
Figura. 2 Modelo Conceptual.



Fuente: elaboración de los autores.

- **Base de datos multidimensional:** ésta se elaboró a partir de las necesidades surgidas al intentar que el programa dé respuesta a las preguntas planteadas. Así pues, en la tabla de hechos se definió las variables que permitirían almacenar los datos (bodegas de datos) para procesos de filtrado (como se puede observar en la figura 3); mientras que las variables que constan dentro de la tabla de dimensiones se definieron en función de la información que se deseaba filtrar. Todo esto, siguiendo el esquema estrella –compuesto por una tabla de hechos y cinco tablas de dimensiones– correspondiente a la base de datos multidimensional propuesta para el caso de estudio.

Figura 3. Base de Datos Multidimensional



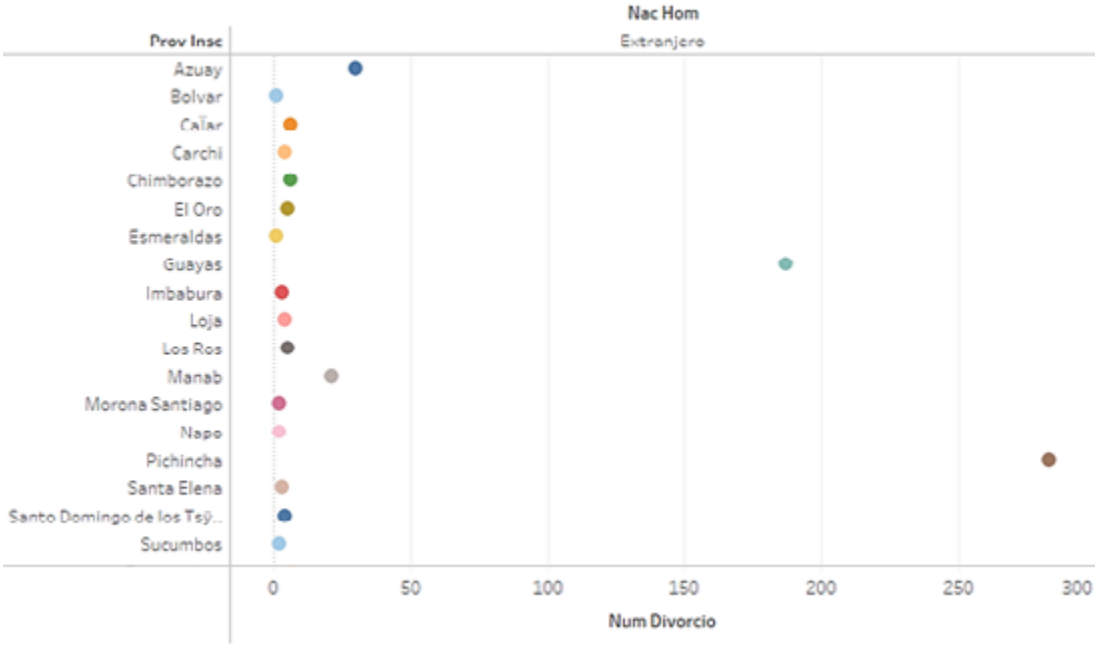
Fuente: elaboración de los autores.

- **Carga de datos:** previo al almacenamiento de los datos en el esquema estrella de la base de datos multidimensional, se procedió a generar la base datos en sí, con SQL. Hecho esto, se procedió a cargar la data con la herramienta Knime, y la apertura con EXAplus, lo cual facilitó la conexión de las herramientas utilizadas y permitió enviar la data ya filtrada a la base.
- **Visualización de los datos:** con los datos ya filtrados y cargados se recurrió a otra herramienta para presentar los datos de forma visual. Con el uso de Tableau se generó las interfaces de muestra de información de una manera atractiva y se logró una rápida conexión con la base alojada en EXAplus.

Ahora bien, ya analizado el aplicativo, se procedió a realizar el estudio de los datos. Se obtuvieron los siguientes resultados:

Al establecer la relación entre provincias y número de divorcios de hombres con nacionalidad extranjera, se evidencia que la provincia con más alto índice de divorcios, en relación con esta variable, es Pichincha, con un total de 283 divorcios (figura 4).

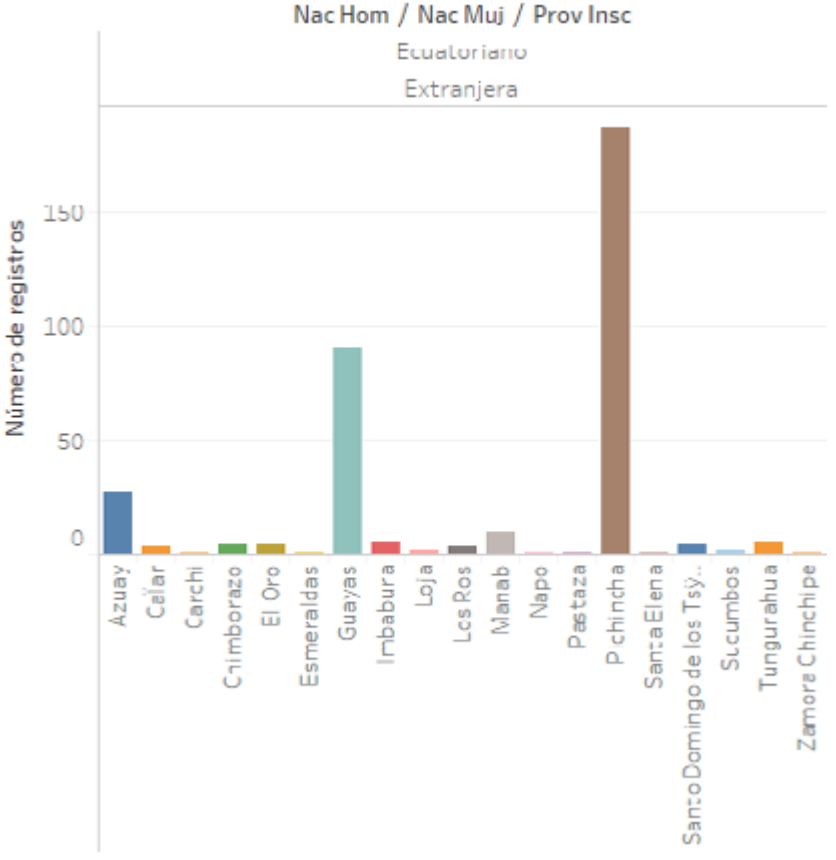
Figura 4. Visualización de datos de hombres de nacionalidad extranjera



Fuente: elaboración de los autores.

Al plantear como variables los divorcios en parejas de mujeres de nacionalidad extranjera y hombres ecuatorianos se obtuvo la cifra de 186 divorcios en la provincia de Pichincha. Le siguen en número Guayas y Azuay (figura 5).

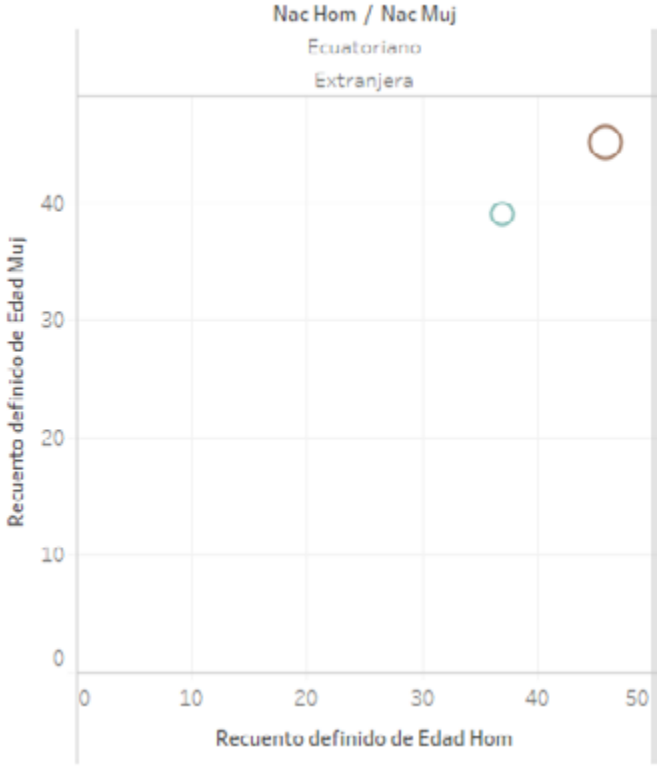
Figura 5. Visualización de datos de hombres ecuatorianos y mujeres extranjeras por provincias



Fuente:elaboración de los autores.

A partir de las variables planteadas, y tomando como referencia las edades de los divorciados (de 37 a 39 años) y la nacionalidad de los mismos (hombres ecuatorianos y mujeres extranjeras), se obtuvieron como valores máximos: Pichincha con 186 divorcios y Guayas con 90 (figura 6).

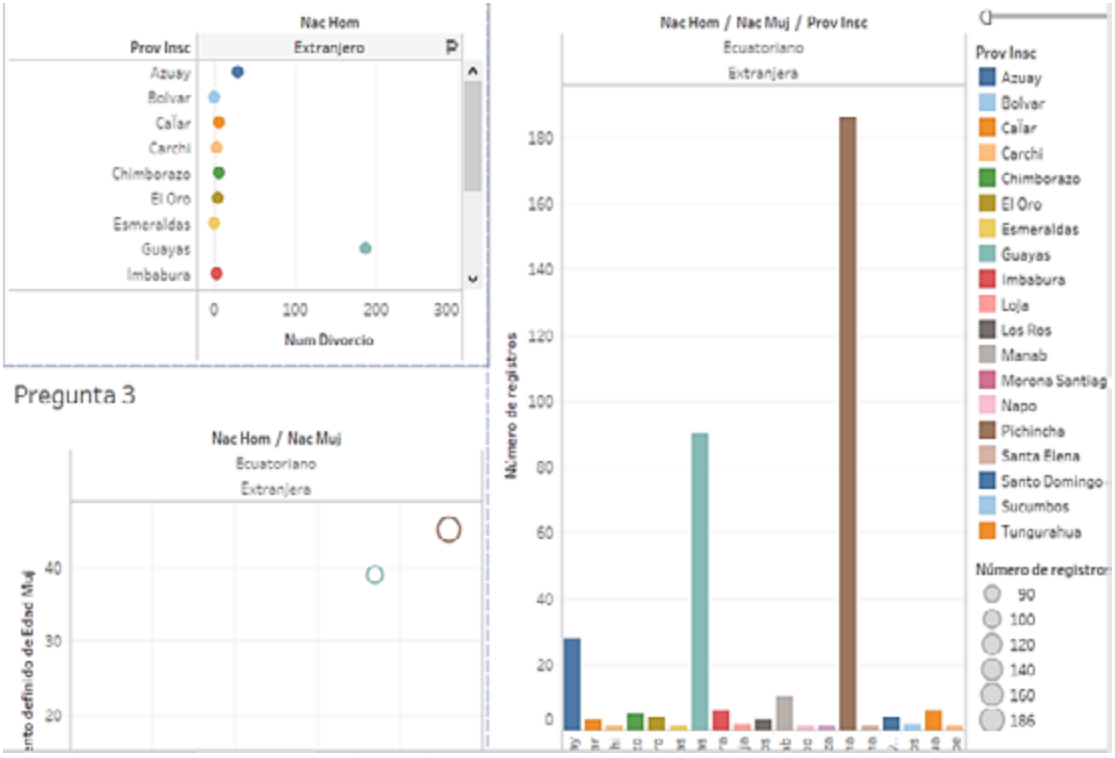
Figura 6. Visualización de datos de hombres ecuatorianos y mujeres extranjeras por edades



Fuente: elaboración de los autores.

En resumen, se obtiene que las provincias donde se suscitan más divorcios de mujeres extranjeras y hombres ecuatorianos es la provincia de Pichincha, tendencia que no varió al aplicar los rangos de edad (figura 7).

Figura 7. Dashboard Divorcios



Fuente: elaboración de los autores.

Finalmente, se lograron datos importantes en relación con los divorcios entre mujeres extranjeras y hombres ecuatorianos, referidas a la población de cada provincia del país (figura 8).

Galápagos	25.124
Zonas no delimitadas	32.384
Pastaza	83.933
Zamora Chinchipe	91.376
Napo	103.697
Orellana	136.396
Morona Santiago	147.940
Carchi	164.524
Sucumbíos	176.472
Bolivar	183.641
Cañar	225.184
Santa Elena	308.693
Santo Domingo de los Tsáchilas	368.013
Imbabura	398.244
Cotopaxi	409.205
Loja	448.966
Chimborazo	458.581
Tungurahua	504.583
Esmeraldas	534.092
El Oro	600.659
Azuay	712.127
Los Ríos	712.127
Manabí	1.369.780
Pichincha	2.576.287
Guayas	3.645.483
Total 14,483,499	

Figura 8. Población ecuatoriana.
Fuente: Instituto Nacional de Estadisticas y Censos.

Resultados

Realizando una comparación de la población por provincia (como se indica en la figura 8), se puede observar los siguientes resultados obtenidos a partir de la base de datos Estadísticas Vitales: Matrimonios y Divorcios"- INEC 2017:

La provincia de Pichincha, con una población de 2 576 287 habitantes, es la provincia con mayor número de divorcios entre mujeres extranjeras y hombres ecuatorianos en el rango de edad de 37 a 39 años.

La provincia del Guayas, pese a ser la más poblada del país con 3 645 483 habitantes, tiene un índice menor de divorcios entre mujeres extranjeras y hombres ecuatorianos en el rango de edad de 37 a 39 años.

Conclusiones

Con el caso expuesto se pretende dar a conocer el aporte que prestan las herramientas de analítica de datos, permitiendo transformar un conglomerado alto de datos, difíciles de entender por sí solo, en información gráfica, en la que se pueden resumir variables complicadas en esquemas que permiten ser analizar de mejor manera los resultados.

Se pudo evidenciar, asimismo, que la aplicación de herramientas de analíticas de datos no es solo para grandes o medianas empresas, sino para todo aquel que desee encontrar información que posibilite la construcción de modelos, mediante los cuales se puedan predecir eventos futuros, y establecer políticas que permitan solucionar problemas a nivel del país, como es el caso del análisis de las bases de datos nacionales provista por el INEC del año 2017.

Así pues, con el uso de herramientas de analítica de datos, se logró detectar que Pichincha es la provincia con mayor índice de divorcios en el Ecuador, incluso mayor a los suscitados en la provincia del Guayas, que cuenta con más habitantes a nivel nacional.

Referencias bibliográficas

Bernabou D. y García M. (2017). Hefesto. Data warehousing. Guía completa de aplicación teórico-práctica; metodología Data Warhouse. Recuperado de https://www.businessintelligence.info/assets/hefesto-v2.pdf

Brizuela, L. y Castro, Y. (2013). Metodologías para desarrollar Almacén de Datos. Revista de Arquitectura e Ingeniería, 7 (3). Recuperado de file:///C:/Users/SALAPROFESORES01/Downloads/Dialnet-MetodologiasParaDesarrollarAlmacenDeDatos-4728463.pdf

Farooqui, N. y Mehra, R. (dicimbre, 2018). Design of A Data Warehouse for Medical Information System Using Data Mining Techniques. Conferencia presentada en la Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC), Solan Himachal Pradesh, India. Conferencia recuperada de https://ieeexplore.ieee.org/document/8745864

Instituto Nacional de Estadisticas y Censos (INEC). (2 de junio de 2017). Los divorcios crecieron 83,45% en diez años en Ecuador. INEC. Recuperado http://www.ecuadorencifras.gob.ec/los-divorcios-crecieron-8345-en-diez-anos-en-ecuador/.

Savanur, S. y Shreedhara, K. (2016) Automated data validation for data warehouse testing. Conferencia presentada en la International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT), Mysuru, India. Conferencia recuperada de https://ieeexplore.ieee.org/document/7955219?section=abstract

Sinnexus (2014). Classora: Classora Knowledge Base. Recuperado de https://www.sinnexus.com/business_intelligence/datawarehouse.aspx.