RESUMEN

Analizar grandes volúmenes de datos se ha convertido en una necesidad recurrente tanto para el sector privado como para el sector público, uno de los casos específicos es el Instituto Nacional de Estadísticas y Censos, ente encargado de realizar los Censos de Población y Vivienda del Ecuador cada diez años. A partir del censo, el instituto se encarga de obtener ciertos productos, entre ellos la estratificación de la población según su nivel socioeconómico mediante análisis de varias características, los estratos son alto, medio y bajo. Actualmente, el proceso de estratificación toma gran cantidad de tiempo y esfuerzo para el instituto, por lo que el propósito principal del estudio fue minimizar el tiempo de desarrollo e implementación de la metodología que se usa para la estratificación, sin alterar la calidad estadística de los estratos. Para esto se realizó una gestión de datos ordenada con el Censo de Población y Vivienda 2010, siguiendo la metodología de minería de datos CRISP-DM y dentro de ella aplicando la técnica de clusterización K-medias. Los resultados obtenidos fueron el disminuir el tiempo de desarrollo e implementación de ocho meses a diez semanas, además, se observó que la aplicación de la metodología propuesta en este documento, mejoró la calidad estadística de los estratos construidos de acuerdo a los análisis de varianza realizados a indicadores socioeconómicos. Por lo que, se recomendó el cambio de metodología para el próximo Censo de Población y Vivienda 2020.

PALABRAS CLAVE

- **CENSO DE POBLACIÓN Y VIVIENDA**
- **CRISP-DM**
- **CLUSTERIZACIÓN**
- **K-MEDIAS**
- **NIVEL SOCIOECONÓMICO**

ABSTRACT

Analyzing Big data has become a recurrent need for both private and public sector, one of the specific cases is the National Institute of Statistics and Censuses, the entity in charge of carrying out the Population and Housing Censuses of Ecuador every ten years. From the census, the institute is in charge of obtaining certain products, among them the stratification of the population according to their socioeconomic level through analysis of several characteristics, the strata are high, medium and low. Actually, the stratification process takes a lot of time and effort for the institute, so the main purpose of the study was to minimize the time of development and implementation of the methodology used for the stratification, without altering the statistical quality of the strata. For this, an ordered data management was carried out with the 2010 Population and Housing Census, following the CRISP-DM data mining methodology and within it applying the K-means clustering technique. The results obtained were to decrease the time of development and implementation from eight months to ten weeks, in addition, it was observed that the application of the methodology proposed in this document, improved the statistical quality of the strata constructed according to the analysis of variance to socioeconomic indicators. Therefore, it was recommended to change the methodology for the next 2020 Population and Housing Census.

KEYWORDS

- **CENSUS OF POPULATION AND HOUSING**
- **CRISP-DM**
- **CLUSTERING**
- **K-MEANS**
- **SOCIOECONOMIC LEVEL**