ESPE

UNIVERSIDAD DE LAS FUERZAS ARMADAS

INNOVACIÓN PARA LA EXCELENCIA

VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA TECNOLÓGICA

CENTRO DE POSGRADOS

MAESTRÍA EN GESTIÓN DE SISTEMAS DE INFORMACIÓN E INTELIGENCIA DE NEGOCIOS

TRABAJO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL TÍTULO DE MAGÍSTER EN: GESTIÓN DE SISTEMAS DE INFORMACIÓN E INTELIGENCIA DE NEGOCIOS

TEMA: "PROPUESTA METODOLÓGICA PARA LA ESTRATIFICACIÓN DEL CENSO POBLACIONAL 2020 CON GESTIÓN DE DATOS"

AUTORES:

BORJA PARREÑO, CRISTINA SALOMÉ

NAVAS OLALLA, JUAN PABLO

DIRECTOR: DR. MOLINA BUSTAMANTE, MARCO EDUARDO

SANGOLQUÍ

2019

UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA TECNOLÓGICA

CENTRO DE POSGRADOS

CERTIFICACIÓN

Certifico que el trabajo de titulación, *"PROPUESTA METODOLÓGICA PARA LA ESTRATIFICACIÓN DEL CENSO POBLACIONAL 2020 CON GESTIÓN DE DATOS"* fue realizado por la señorita ***Borja Parreño, Cristina Salomé*** y el señor ***Navas Olalla, Juan Pablo*** el mismo que ha sido revisado en su totalidad, analizado por la herramienta de verificación de similitud de contenido; por lo tanto cumple con los requisitos teóricos, científicos, técnicos, metodológicos y legales establecidos por la Universidad de Fuerzas Armadas ESPE, razón por la cual me permito acreditar y autorizar para que lo sustente públicamente.

Sangolquí, 16 de diciembre de 2018

Firma:

Dr. Marco Eduardo Molina Bustamante

C.C.: 1705613041-4

UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA TECNOLÓGICA

CENTRO DE POSGRADOS

AUTORÍA DE RESPONSABILIDAD

Nosotros, Borja Parreño, Cristina Salomé con cédula de ciudadanía n°1719538124 y ***Navas Olalla, Juan Pablo*** con cédula de ciudadanía n°1715812846, declaramos que el contenido, ideas y criterios del trabajo de titulación: ***"Propuesta metodológica para la estratificación del Censo Poblacional 2020 con gestión de datos"*** es de nuestra autoría y responsabilidad, cumpliendo con los requisitos teóricos, científicos, técnicos, metodológicos y legales establecidos por la Universidad de Fuerzas Armadas ESPE, respetando los derechos intelectuales de terceros y referenciando las citas bibliográficas. Consecuentemente el contenido de la investigación mencionada es veraz.

Sangolquí, 16 de diciembre de 2018

.....

Cristina Salomé Borja Parreño
C.C.: 1719538124

.....

Juan Pablo Navas Olalla
C.C.: 1715812846

UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA TECNOLÓGICA

CENTRO DE POSGRADOS

AUTORIZACIÓN

Nosotros, ***Borja Parreño, Cristina Salomé*** y ***Navas Olalla, Juan Pablo*** autorizamos a la Universidad de las Fuerzas Armadas ESPE publicar el trabajo de titulación: ***"Propuesta metodológica para la estratificación del Censo Poblacional 2020 con gestión de datos"*** en el Repositorio Institucional, cuyo contenido, ideas y criterios son de nuestra responsabilidad.

Sangolquí, 16 de diciembre de 2018

.....

Cristina Salomé Borja Parreño

C.C.: 1719538124

.....

Juan Pablo Navas Olalla

C.C.: 1715812846

AGRADECIMIENTOS

Agradezco a la Universidad de las Fuerzas Armadas ESPE por los conocimientos adquiridos durante la maestría que hicieron posibles la realización de este trabajo.

Agradezco a mi compañero Juan Pablo Navas Olalla por el esfuerzo y dedicación brindados durante el desarrollo de este trabajo.

Salomé

AGRADECIMIENTOS

Este trabajo se ha llevado a cabo con el esfuerzo de varias personas, el sacrificio de mi esposa María Belén al privarse de varias distracciones por acompañar en este proceso, me es ampliamente valorado y agradezco infinitamente su comprensión y apoyo

A mi compañera Salomé, ya que gracias a su gran capacidad y dedicación hemos concluido con este proceso y nos hemos apoyado al máximo.

A Dios, el universo, la vida por todo y tanto.

Juan Pablo

DEDICATORIA

El presente trabajo está dedicado a mi madre, por enseñarme la importancia de la educación y el profesionalismo y a toda mi familia por su apoyo incondicional.

Salomé

DEDICATORIA

A mi familia por guiarme de una manera poco ortodoxa pero efectiva a la realización de sueños y alcanzar las metas. A mi amada esposa por su incondicional apoyo y sobre todo al bebe que esta en camino, no lo sabía pero todo esto es por ti…

Juan Pablo

ÍNDICE DE CONTENIDOS

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

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**

CAPÍTULO I

PROBLEMA DE INVESTIGACIÓN

1.1. ANTECEDENTES

En nuestro país se está introduciendo la cultura del uso apropiado de la información, hoy en día estamos rodeados de datos que pueden ser utilizados tanto por las empresas privadas como por el sector público en beneficio de la toma de decisiones sustentadas en la realidad.

En el caso del aparato público, la información permite planificar y ejecutar políticas públicas de acuerdo a las necesidades de la sociedad. Considerando que el sector público tiene mayor acceso y custodia de los datos de la población en distintas instituciones como el Registro Civil, el Seguro Social, el Servicio de Rentas Internas y el Instituto Nacional de Estadística y Censos, para que la información pueda ser aprovechada, se necesita gestión adecuada de los datos, debido a que la gran cantidad y a los problemas que se puede encontrar en la calidad de los mismos, se dificulta el procesamiento y análisis. Esta dificultad ha sido sustentada a medida que pasa el tiempo de acuerdo a las crecientes necesidades, "La disponibilidad de grandes volúmenes de información y el uso generalizado de herramientas informáticas ha transformado el análisis de datos orientándolo hacia determinadas técnicas especializadas englobadas bajo el nombre de minería de datos o *Data Mining*" (Pérez López & Santín Gonzáles, 2007, p. 1).

Uno de los casos emblemáticos para el tratamiento de grandes volúmenes de información del sector público, es el Instituto Nacional de Estadística y Censos, este organismo es el encargado de realizar el censo de población y vivienda cada diez años. El censo además de obtener información que permite analizar la situación del país en el momento en que es realizado, permite ser la base de futuros estudios de la población que se realizan de manera continua tales como la Encuesta de

Empleo, Desempleo y Subempleo (ENEMDU)[1], la Encuesta de Ingresos y Gastos de los Hogares Urbanos y Rurales (ENIGHUR)[2], la Encuesta de Condiciones de Vida (ECV)[3], entre otras.

Para la generación de las encuestas es necesario contar con un Marco Muestral estratificado, con el fin de que la muestra logre representar a toda la población objetivo de la encuesta.

Uno de los retos que implica la construcción de un Marco Muestral es el realizar la estratificación, al ser un volumen extenso de información, es complicado evaluar el método adecuado con las herramientas estadísticas que el INEC cuenta en la actualidad.

El producto final de este trabajo es una nueva metodología para la estratificación de la población ecuatoriana, dicha metodología brindará al usuario una guía detallada para su implementación, con lo cual se minimizará problemas de subjetividad por parte de los interventores, además, de una ejecución en menor tiempo que la metodología actual, ya que, el uso de técnicas de minería de datos lleva consigo la búsqueda de patrones poco perceptibles y se lo realiza con herramientas tecnológicas avanzadas.

1.2. CONTEXTO DEL PROBLEMA

El Instituto Nacional de Estadística y Censos, es el ente coordinador de la estadística del país, para la producción de estadística de calidad es necesario el cumplimiento del "Código de Buenas Prácticas Estadísticas" (INEC, 2014), que establece como principio la oportunidad y puntualidad de la producción de información estadística. El Marco de Muestreo es uno de los insumos más importantes al hacer una encuesta, sin embargo, en las estadísticas sociodemográficas y encuestas dirigidas a la población, no se cuenta con un marco actualizado, ya que el marco se construye a partir de los censos que se realizan cada diez años, y además de este inconveniente su construcción

[1] "El propósito de la Encuesta Nacional de Empleo, Desempleo y Subempleo es proporcionar información sobre el Mercado Laboral Ecuatoriano, a través de la recolección de datos, con periodicidad mensual en las principales ciudades y trimestral en el total nacional urbano". (INEC, 2018)

[2] "La Encuesta Nacional de Ingresos y Gastos de Hogares Urbanos y Rurales, proporciona datos sobre el monto, distribución y estructura del ingreso y el gasto de los hogares, a partir de las características demográficas y socioeconómicas de sus miembros" (INEC, 2018)

[3] La Encuesta de Condiciones de Vida tiene como objetivo "Estudiar los impactos económicos y las condiciones de vida en que se encuentra inmersa la población ecuatoriana, desde la perspectiva de las encuestas de hogares, permitiendo contar con una línea de base actualizada y útil para la medición del cumplimiento de los objetivos del Plan Nacional del Buen Vivir (PNVB)" (INEC, 2015, pág. 10).

toma mucho tiempo para cada encuesta, lo que hace que se tenga un marco desfasado en el tiempo y los estratos del marco no correspondan a la realidad del momento de levantamiento de la información.

Uno de los problemas más importantes de la estratificación es que la metodología aplicada actualmente tiene graves problemas para su ejecución, puesto que la descripción de los pasos del proceso involucra extenso análisis y es propenso a subjetividad de los actores, además cuenta con escasa documentación y es realizado sin ningún tipo de sistema de automatización.

1.3. PLANTEAMIENTO DEL PROBLEMA

La metodología de estratificación del Censo de Población y Vivienda 2010 que utilizó el INEC en la generación del Marco Maestro de Muestreo, tomó alrededor de dos años para plantearse la necesidad de cambio del marco y ocho meses en su construcción e implementación. El proceso tomó este tiempo debido a la falta de planificación y a las limitadas herramientas tecnológicas con las que cuenta la institución, que, al ser un proceso complejo el recurso humano tarda gran cantidad de tiempo en su ejecución. En consecuencia, provoca evidente retraso en la entrega de información sumamente importante para la toma de decisiones del gobierno de turno. Adicionalmente, en el año 2020 se presentarán problemas al momento de replicar la metodología de estratificación, pues, esta no se encuentra documentada de forma detallada.

1.4. JUSTIFICACION, IMPORTANCIA Y ALCANCE

Cuando el Marco de Muestreo para la estratificación de la población ecuatoriana toma tanto tiempo y esfuerzo para una entidad pública como el INEC, la pérdida de recursos y el perjuicio para el estado es incalculable; la metodología actual cumple con el objetivo, sin embargo, la propuesta de una nueva metodología, dinámica a través de técnicas de minería de datos para el procesamiento de la información más relevante, llevará a la optimización de tiempo y recursos para la entidad y constituirá la base para aplicación de técnicas de minería de datos en otras fuentes de información que el INEC pueda sacar provecho en beneficio del país.

La investigación pretende generar una nueva metodología de construcción de los estratos socioeconómicos de la población del Ecuador con el Censo de Población y Vivienda que realiza el

INEC en la actualidad, estos son: Estrato Alto, Estrato Medio y Estrato Bajo. Esta no tiene como objetivo realizar una crítica a la metodología actual de estratificación, en su lugar, se propone una metodología alternativa que pueda optimizar los tiempos de construcción de los mismos a través de la minería de datos. Además, la investigación propone comparar y evaluar la metodología de estratificación propuesta.

1.5. HIPOTESIS

Las técnicas de segmentación/agrupación implementadas bajo la gestión de datos organizada permitirán mejorar el tiempo de construcción de los niveles socioeconómicos/estratos del Censo de Población y Vivienda sin afectar la calidad estadística de los clústeres obtenidos.

Señalamiento de Variables:

Variable dependiente: Construcción de los niveles socioeconómicos/estratos.

Variable independiente: Técnicas de segmentación/agrupación.

Para la demostración de esta hipótesis se puede realizar análisis tanto cualitativos como cuantitativos. Para la validación cuantitativa se medirá el tiempo de implementación de la metodología y para la calidad estadística de los clústeres se realizará análisis estadísticos que miden la varianza inter e intra clase de los estratos generados por el modelo.

1.6. OBJETIVOS

1.6.1. Objetivo General

Generar una propuesta metodológica para la estratificación del Censo de Población y Vivienda 2020 en base a gestión de datos que minimicen el tiempo de construcción del Marco Muestral de encuestas dirigidas a hogares del Ecuador.

1.6.2. Objetivos Específicos

OE1: Realizar una revisión de la metodología que usó el INEC en la estratificación del Censo de Población y Vivienda 2010 para la creación del Marco Muestral, estableciendo los puntos críticos de la metodología aplicada.

OE2: Gestionar los datos provenientes del Censo de Población y Vivienda para la preparación de los principales indicadores o variables que intervienen en la estratificación, mediante procesos de migración e integración de datos.

OE3: Implementar un modelo de segmentación para construir los estratos de la población del Censo de Población y Vivienda 2010, a través del uso de técnicas de clusterización.

OE4: Evaluar la metodología propuesta de acuerdo a los resultados obtenidos en la construcción de los estratos en el Censo de Población y Vivienda 2010, comparando el tiempo y los puntos críticos entre la metodología utilizada por el INEC y la propuesta.

1.7. PREGUNTAS DE INVESTIGACIÓN

- **OE1 –RQ1.** ¿Cuáles son los puntos críticos de la metodología que aplica actualmente el INEC para obtener los niveles socioeconómicos de la población?
- **OE2 - RQ2.** ¿Qué metodologías de gestión de datos se pueden usar para la administración de los datos obtenidos desde un censo?
- **OE2 - RQ3**. ¿Cuáles son los principales indicadores o variables que intervienen en la construcción de los niveles socioeconómicos de un país?
- **OE3 - RQ4.** ¿Cuál es el proceso que se debe seguir para implementar un modelo de segmentación?
- **OE4 -RQ5**. ¿Cuáles son los principales aportes de la metodología propuesta frente a la existente?

CAPÍTULO II

ESTADO DEL ARTE Y MARCO TEÓRICO

2.1. ESTADO DEL ARTE

El objetivo del estado del arte es conocer los trabajos existentes relacionados a la investigación que aporten con conocimiento para enriquecer la propuesta metodológica que se busca realizar y respondan a las preguntas de investigación planteadas en el presente documento.

2.1.1. Estudios Primarios

Se realizó un *Systematic Mapping Study(SMS)* mediante el cual se obtuvo cuatro estudios primarios de distintos repositorios que aportarán conocimientos a la investigación.

Tabla 1
Estudios Primarios Seleccionados

Código	Título	Repositorio
EP1	An Effective Algorithm Based on Density Clustering Framework	IEEE
EP2	Organization-Ontology Based Framework for Implementing the Business Understanding Phase of Data Mining Projects	IEEE
EP3	A view on the methodology of analysis and exploration of marketing data	IEEE
EP4	Predicting the need of Neonatal Resuscitation using Data Mining	SCIENCE DIRECT

2.1.2. Extracción de Datos

Basándose en los estudios primarios se puede obtener finalmente los resultados del SMS, para esto es necesario realizar la extracción de datos de los artículos, se extrae las características de los estudios y la información específica objeto del SMS.

Para la etapa de extracción de la información específica objeto de la SMS en primer lugar se analiza los enfoques que tiene cada uno de estos estudios.

En el EP1, “An Effective Algorithm Based on Density Clustering Framework” los autores Jianyun Lu y Qingsheng Zhu enfocan la investigación a proponer un nuevo algoritmo de clustering para disminuir los ruidos que tienen los algoritmos actuales

En el EP2, “Organization-Ontology Based Framework for Implementing the Business Understanding Phase of Data Mining Projects” los autores Sumana Sharma y Kweku-Muata Osei-Bryson de la universidad Virginia Commonwealth University explican cómo debe aplicarse la fase de “comprensión del negocio” en la metodología CRISP-DM de una manera estructurada.

En el EP3, “A view on the methodology of analysis and exploration of marketing data” de los autores Maciej Pondel y Jerzy Korczak de la Universidad Wroclaw de Economía proponen una metodología para el desarrollo de un sistema de apoyo a la toma de decisiones de marketing que utiliza tecnología Big Data y técnicas de minería de datos. El enfoque se inspiró en la metodología CRISP-DM, que no está orientada a proyectos Big Data. Por lo tanto, se ha modificado esta metodología con respecto al propósito y los requisitos tecnológicos del proyecto.

En el EP4, “Predicting the need of Neonatal Resuscitation using Data Mining” los autores Ana Morais, Hugo Peixoto, Cecília Coimbra, António Abelha y José Machado evalúan el método estándar para minería de datos (CRISP-DM) y mediante la consecución de sus fases logran alcanzar resultados de sensibilidad superiores al 90% y los resultados de especificidad y precisión superiores al 98%, que se consideraron satisfactorios respecto de su estudio. Además, utilizaron la herramienta de software WEKA para inducir los modelos de DM.

Después de conocer la metodología es importante analizar a que resultados se llegó en cada uno de los estudios, para poder diferenciarlos y distinguir de qué manera puede aportar al estudio que proponemos realizar y cómo podemos responder nuestras preguntas de investigación. A continuación, se resume los resultados principales a los que llegan los estudios primarios

En el EP1:

- Propone un marco de agrupamiento basado en la densidad (DCF) que puede combinar diferentes algoritmos de agrupamiento para obtener mejores resultados de agrupamiento.
- Explica el cómo se debe realizar el algoritmo paso a paso.

- Desarrolla la construcción del algoritmo mediante fórmulas.
- Evaluar el rendimiento del algoritmo de agrupación propuesto con información real.

En el EP2:

- Detalla paso a paso como realizar la fase de “comprensión del negocio” en la metodología CRISP-DM.
- Propone como desarrollar la fase de manera más simple y para comprensión de todos que permita introducir automatización en el proceso.
- Permite mantener una concordancia con el resto de fases de la metodología CRISP-DM.

En el EP3, se encontró lo siguiente:

- Estudios recientes sobre el uso de metodologías en grandes proyectos de exploración de bases de datos indican que la metodología CRISP-DM, domina (42% de las aplicaciones), seguida por metodologías propias (19%), mientras que la metodología SEMMA propuesta por SAS ocupa el tercer lugar (13%)
- Los estudios han demostrado que a medida que aumentan los flujos de datos de diferentes fuentes, su calidad se deteriora. Por lo tanto, los procesos de recopilación y preparación de datos son extremadamente importantes.

Para el EP4, tenemos la siguiente información:

- La información comprende el año de 2016 y tiene información sobre 3163 recién nacidos, junto con información sobre sus madres y los respectivos episodios de parto.
- Durante el proceso de minería de datos, se siguió el método estándar de industria cruzada para minería de datos (CRISP-DM), que es un modelo de proceso jerárquico que divide el proceso de minería de datos en seis fases: comprensión comercial, comprensión de datos, preparación de datos, modelado, evaluación e implementación.
- Usando datos reales de EHR y registros de admisión del servicio de obstetricia, es viable utilizar modelos de DM para predecir la necesidad de reanimación neonatal dadas algunas condiciones de salud tanto del recién nacido como de la madre

- Fue posible, para algunos modelos de DM, lograr resultados de sensibilidad superiores al 90% y resultados de especificidad y precisión superiores al 98%, que se consideran bastante satisfactorios.

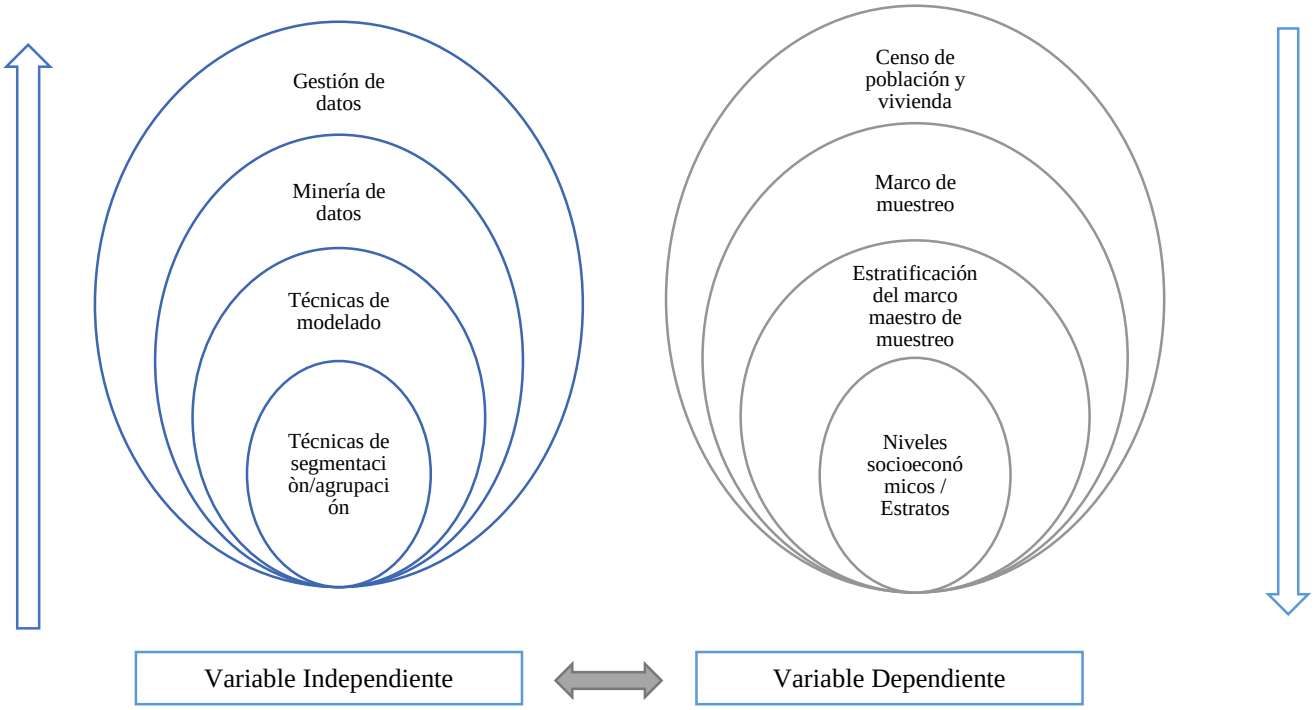
Además de estos estudios se utiliza documentación específica correspondiente a los temas establecidos:

- METODOLOGÍA DEL DISEÑO MUESTRAL DE LA ENCUESTA NACIONAL DE EMPLEO Y DESEMPLEO ENEMDU.- El Instituto Nacional de Estadística y Censos de Ecuador presenta la metodología que utilizó para construir los niveles socioeconómicos o estratos en el documento del año 2013.
- CONSULTORIA PARA LA CONSTRUCCIÓN DEL SISTEMA DE DIFUSIÓN DE LOS CENSOS NACIONALES DE POBLACIÓN Y VIVIENDA 2007.- El Instituto Nacional de Estadística e Informática de Perú en su documento hace un acercamiento a la construcción de un Data warehouse para el censo de un país.
- HACIA UN SISTEMA INTEGRADO DE ENCUESTAS DE HOGARES EN LOS PAÍSES DE AMÉRICA LATINA.-Este documento realizado por la CEPAL busca guiar a los países en las consideraciones que deben tener al momento de implementar un sistema integrado de encuestas, entre las que se encuentran el cómo realizar un Marco de muestreo.

En conclusión, se logró obtener artículos de apoyo a nuestro estudio en los temas de aplicación de la metodología de gestión de datos y aplicación de un algoritmo de agrupación. En algunos de estos artículos se aplica los temas propuestos, lo cual puede servir como un ejemplo a seguir en las distintas fases del desarrollo de este estudio.

2.2. MARCO TEÓRICO

El marco teórico permite mantener una base teórica orientada a la investigación de la hipótesis planteada. Para esto es recomendable realizar una red de categorías ordenadas jerárquicamente que explican las variables dependiente e independiente.



2.2.1. Fundamentación de la variable independiente

- **Gestión de datos**

La definición oficial suministrada por la Data Management Association (DAMA) es "La Gestión de Datos es el desarrollo y ejecución de arquitecturas, políticas, prácticas y procedimientos que gestionan apropiadamente las necesidades del ciclo de vida completo de los datos de un estudio". (Dynamic, 2018)

- **Minería de datos**

Es un paso en el proceso de KDD (Knowledge Discovery in Databases) que consiste en aplicar algoritmos de análisis y descubrimiento de datos que producen una enumeración particular de patrones (o modelos) sobre los datos. El término minería de datos ha sido utilizado principalmente por estadísticos, analistas de datos y comunidades de sistemas de información gerencial (MIS). (Piatetsky-Shapiro, 1996, p. 39)

- **Técnicas de modelado**

Las técnicas de modelado buscan responder un tipo de problema y se clasifican en:

  - Descripción y resumen de datos.
  - Segmentación.
  - Descripción del concepto.
  - Clasificación.
  - Predicción.
  - Análisis de dependencias. (Wirth & Hipp, 2000)

- **Técnicas de segmentación/agrupación**

Las técnicas de segmentación responden al tipo de problema de minería de datos que tiene como objetivo la separación de los datos en subgrupos o clases interesantes y significativos. Todos los miembros de un subgrupo comparten características comunes.

La segmentación se puede realizar manualmente o (semi) automáticamente. El analista puede hipotetizar ciertos subgrupos como relevantes para la pregunta de negocios basada en conocimiento o basado en el resultado de la descripción y resumen de datos. Sin embargo, también hay técnicas de agrupación automática que pueden detectar estructuras previamente ocultas e insospechadas en los datos que permiten la segmentación. (Chapman, y otros, 1999)

Existen tres tipos de algoritmos de segmentación:

  - Agrupamiento por particiones: dentro de estos los más conocidos son *K-means* y CLARANS
  - Métodos basados en densidad como DBSCAN
  - *Clustering* jerárquico: BIRCH, ROCK, CHAMELEON (Berzal, 2018)

El algoritmo de agrupamiento por particiones *K-means* esta definido de la siguiente manera:

> El algoritmo *K-means*, creado por MacQueen en 1967 es el algoritmo de clustering más conocido y utilizado ya que es de muy simple aplicación y eficaz. Sigue un procedimiento simple de clasificación de un conjunto de objetos en un determinado número K de clústeres, K determinado a priori.
> El nombre de *K-means* viene porque representa cada uno de los clústeres por la media (o media ponderada) de sus puntos, es decir, por su centroide. La representación mediante centroides tiene la ventaja de que tiene un significado gráfico y estadístico inmediato. Cada clúster por tanto es caracterizado por su centro o centroide que se encuentra en el centro o el medio de los elementos que componen el clúster. *K-means* es traducido como K-medias.
> O un conjunto de objetos Dn = (x1, x2..., xn), para todo el i, xi reales y k, v1, los centros de los K clúster. El algoritmo del *K-means* se realiza en 4 etapas:
>
> Etapa 1: Elegir aleatoriamente K objetos que forman así los K clusters iniciales. Para cada clúster k, el valor inicial del centro es = xi, con los xi únicos objetos de Dn pertenecientes al clúster.
> Etapa 2: Reasigna los objetos del clúster. Para cada objeto x, el prototipo que se le asigna es el que es más próximo al objeto, según una medida de distancia, (habitualmente la medida euclidiana).
> Etapa 3: Una vez que todos los objetos son colocados, recalcular los centros de K clúster. (los baricentros).
> Etapa 4: Repetir las etapas 2 y 3 hasta que no se hagan más reasignaciones. Aunque el algoritmo termina siempre, no se garantiza el obtener la solución óptima. En efecto, el algoritmo es muy sensible a la elección aleatoria de los K centros iniciales. Esta es la razón por la que, se utiliza el algoritmo del *K-means* numerosas veces sobre un mismo conjunto de datos para intentar minimizar este efecto, sabiendo que a centros iniciales lo más espaciados posibles dan mejores resultados. (García Cambronero & Gómez Moreno, 2012, págs. 7 - 8)

2.2.2. Fundamentación de la variable dependiente

- **Censo de población y vivienda**

Naciones Unidas define a los censos de población como “un conjunto de operaciones que consiste en reunir, elaborar y publicar datos demográficos y también económicos y sociales, correspondientes a todos los habitantes de un país o territorio definido y referido a un momento determinado o a ciertos periodos de tiempo dados”. (INEC, 2012, p. 8)

- **Marco de muestreo**

El marco de muestreo se encuentra definido por el INEC de la siguiente manera:

> Es una lista completa, organizada en forma de base de datos que contiene a todos y cada uno de los elementos de la población de interés que participaran en cada una de las fases de selección de la muestra. El Marco también está formado por todos los mapas y planos a diferentes escalas que nos permiten identificar en forma precisa y clara los límites físicos que tienen las diferentes unidades de selección. (INEC, 2013, p. 3)

- **Estratificación del marco maestro de muestreo**

La definición de estratificación del marco maestro de muestreo de acuerdo al diseño muestral planteado por el INEC para la encuesta ENEMDU:

> Consiste en agrupar de acuerdo a ciertas similitudes, las Unidades Primarias de Muestreo (UPM) creadas previamente en base a la Información del Censo de Población y Vivienda 2010. Formalmente, la estratificación se refiere a la subdivisión de una población determinada en subconjuntos con características propias. Esta acción se lleva a cabo como una etapa previa a la selección de la muestra y las variables utilizadas para ello contienen información acerca de todas las unidades de la población. (INEC, 2013, p. 6)

- **Niveles socioeconómicos / Estratos**

Los estratos o niveles socioeconómicos que son generados en el Marco Maestro de Muestreo responden a un fin estadístico y se forman de acuerdo a las características de la población:

> El contar con una clasificación previa de las UPM del país de acuerdo a la similitud de los fenómenos descritos anteriormente permite generar muestreos más eficientes. Estas variables permiten agrupar las UPM de acuerdo a cierto grado de marginación o bienestar de la población. Desde un punto de vista de "bienestar", mientras más existan estas condiciones en las UPM, tenderán a ser consideradas de estrato "alto", por el contrario, mientras las UPM tiendan a la ausencia de estas características (en conjunto) serán consideradas de estrato "bajo". (INEC, 2013, pp. 6 - 7)

2.3. ANÁLISIS CONCEPTUAL

La estratificación que se realiza de los niveles socioeconómicos del país, en este estudio, tiene como único fin la de garantizar que en los estudios basados en encuestas que parten del marco muestral, se represente la realidad de los distintos estratos sociales para los fenómenos que se investigan. "La estratificación social tanto en Ecuador como otros países, se ha constituido como una variable de importancia, generalmente los modelos utilizados elevan discusiones amplias y los diseñan en función de objetivos para los que fueron concebidos". (Salas L., 2018, pág. 1)

En el caso de esta investigación, se cuenta con las variables que son investigadas en el Censo de Población y Vivienda 2010, estas variables se dividen en tres grupos, características de las viviendas, de los hogares y de la población.

> Existen dos clases de indicadores para los niveles socioeconómicos, multidimensionales y unidimensionales, el primero trata con varias variables tales como el ingreso, el nivel de

> educación, el empleo, las características de la vivienda, servicios del hogar y el acceso a la tecnología; mientras que el segundo se basa normalmente en los ingresos o gastos del hogar, recibiendo varias críticas por no reflejar completamente la calidad de vida de los hogares, los mismo que pueden depender de más factores (Salvador, Larrea, Belmont, & Baroja, 2014). (Salas L., 2018, pág. 1)

En el Censo de Población y Vivienda no se cuenta con variables como ingresos o gastos, por lo que es importante considerar un modelo multidimensional que contemple las características investigadas por este instrumento.

Dentro de la Minería de Datos se encuentran las técnicas de segmentación o agrupamiento, las cuales facilitan la agrupación de los individuos de acuerdo a las características que presentan de forma multidimensional. Para este estudio es importante considerar que el número de clústeres a formar ya están predefinidos, por lo que el tipo de algoritmos a utilizar deberán ser de tipo "Agrupamiento por particiones", de esta forma se obtiene los niveles socioeconómicos alto, medio y bajo.

2.4. METODOLOGÍA PROPUESTA

De acuerdo a la extensa cantidad información con la que cuenta el INEC, y se evidencia que se trata de un problema de clusterización o agrupación, es necesario alinearse a una metodología de Minería de Datos, entre estas tenemos: KDD (Knowledge Discovery in Databases), SEMMA (Sample, Explore, Modify, Model, and Assess) y CRISP-DM (Cross Industry Standard Process for Data Mining). "SEMMA y CRISP-DM han sido elegidos, ya que son considerados los más populares. Aunque no es científica, existe esta percepción, porque SEMMA y CRISP-DM se presentan en muchas de las publicaciones del área y se utilizan realmente en la práctica." (Azevedo & Santos, 2008).

KDD: Usa métodos de minería de datos para extraer conocimiento de acuerdo con la especificación de medidas, utilizando una base de datos junto con cualquier procesamiento previo, submuestreo y transformación requeridos de la base de datos. Consta de 5 fases ilustradas en la siguiente figura:

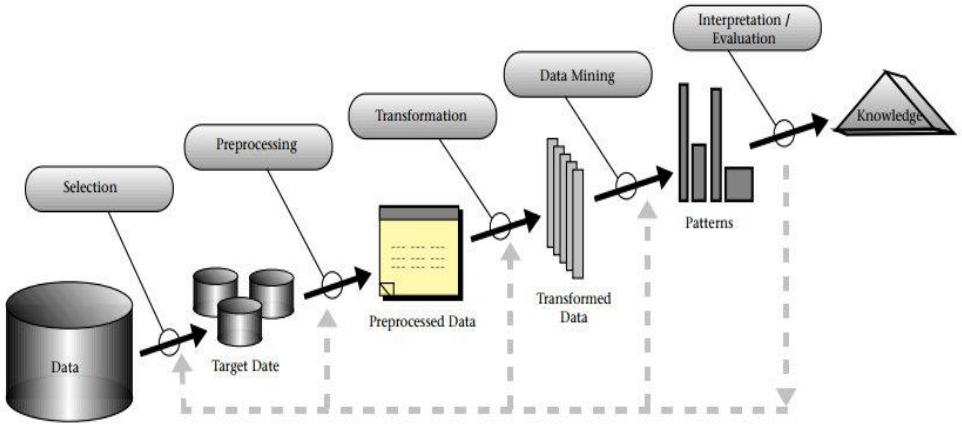


Figura 1. Proceso KDD
Fuente: (Mayo, 2016)

1. Selección: consiste en crear un conjunto de datos objetivo, o centrarse en un subconjunto de variables o muestras de datos, en el que se realizará el descubrimiento
2. Preprocesamiento: consiste en la limpieza de datos objetivo y el preprocesamiento para obtener datos consistentes
3. Transformación: esta etapa consiste en la transformación de los datos utilizando métodos de reducción de dimensionalidad o transformación
4. Minería de datos: esta etapa consiste en la búsqueda de patrones de interés en una forma de representación particular, dependiendo del objetivo de DM (generalmente, predicción)
5. Interpretación / Evaluación: esta etapa consiste en la interpretación y evaluación de los patrones minados.

SEMMA: El acrónimo SEMMA significa Muestra, Explorar, Modificar, Modelar, Evaluar para el proceso de minería de datos. A continuación, se detalla cada una de las fases

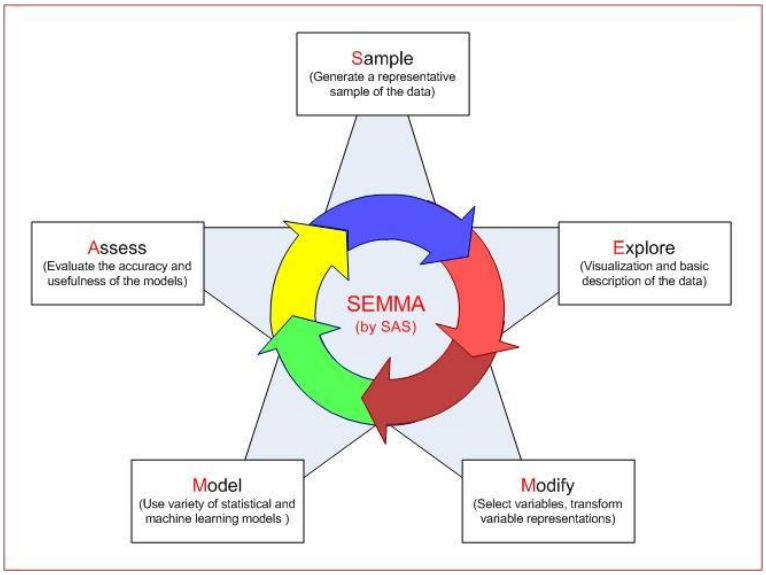


Figura 2. Proceso SEMMA
Fuente: (BINUS University, 2014)

1. Muestra: consiste en la extracción de una parte de un conjunto de datos grandes, lo suficientemente grande para contener la información significativa, y lo suficientemente pequeño para manejarlo rápidamente.
2. Explorar: consiste en la búsqueda de tendencias y anomalías para obtener comprensión de los datos.
3. Modificar: consiste en la modificación de los datos, de tal forma que las variables se alineen al modelo
4. Modelo: consiste en permitir que el software busque automáticamente una combinación de datos que predice de manera confiable el resultado.
5. Evaluación: consiste en la evaluación de la utilidad y confiabilidad de los hallazgos y estimar qué tan bien se desempeñan.

CRISP-DM: (Proceso estándar entre-industria para la minería de datos) propone un modelo de proceso integral para llevar a cabo proyectos de minería de datos. El modelo de proceso es independiente tanto del sector industrial como de la tecnología utilizada. (Wirth & Hipp, 2000, pp. 1 - 2). Consiste en un ciclo de los datos representado en la siguiente figura:

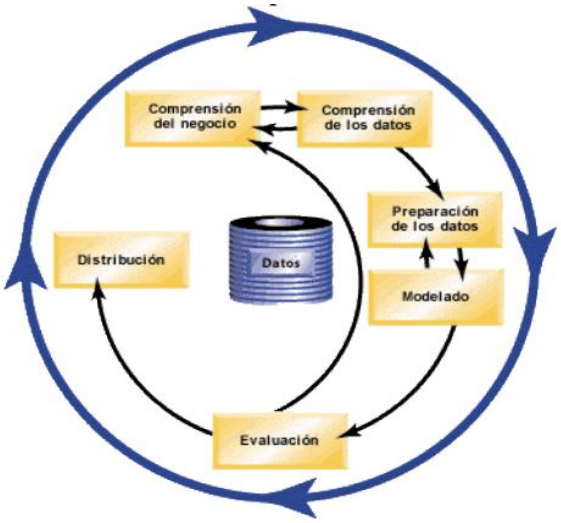


Figura 3. Metodología CRISP-DM
Fuente: IBM (IBM SPSS Inc., 2011)

1. Entendimiento del negocio: Se enfoca en comprender todos los aspectos relevantes del negocio, y luego convertir este conocimiento en una definición de problema de minería de datos.
2. Comprensión de datos: Se realiza una recopilación de datos inicial, con las actividades para familiarizarse con los datos e identificar problemas de calidad de los datos.
3. Preparación de datos: Cubre todas las actividades para construir el conjunto de datos final a partir de los datos sin procesar iniciales.
4. Modelado: Se seleccionan y aplican diversas técnicas de modelado y sus parámetros se calibran a valores óptimos.
5. Evaluación: el modelo (o los modelos) obtenidos se evalúa más a fondo y los pasos ejecutados para construir el modelo se revisan para asegurarse de que logre adecuadamente los objetivos comerciales.
6. Distribución creación del modelo generalmente no es el final del proyecto. Incluso si el propósito del modelo es aumentar el conocimiento de los datos, el conocimiento adquirido deberá organizarse y presentarse de manera que el cliente pueda utilizarlo.

CRISP-DM es extremadamente completo y documentado. Todas sus etapas están debidamente organizadas, estructuradas y definidas, lo que permite que el proyecto pueda entenderse o revisarse fácilmente. (Azevedo & Santos, 2008)

Comparando los tres estándares desde el punto de vista de los pasos a seguir para cada uno se tiene la siguiente tabla:

Tabla 2

Resumen de correspondencia en el proceso de KDD, SEMMA y CRISP-DM

KDD	SEMMA	CRISP-DM
Pre KDD	----------	Comprensión del Negocio
Selección	Muestra	Comprensión de los datos
Pre procesamiento	Exploración	
Transformación	Modificar	Preparación de los datos
Minería de datos	Modelo	Modelamiento
Interpretación/Evaluación	Valoración/Evaluación	Evaluación
Post KDD	----------	Distribución

Fuente: Traducido de KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW (Azevedo & Santos, 2008)

Al revisar estos estándares, primero es necesario establecer que tanto CRISP-DM como SEMMA son la implementación del proceso KDD, sin embargo, CRISP-DM tiene una visión más completa por lo que se apega con mayor flexibilidad a la realidad actual y tiene la ventaja de tener un modelo circular en donde las etapas pueden retroalimentarse.

El modelo de proceso CRISP-DM tiene como objetivo hacer grandes proyectos de minería de datos, menos costosos, más confiables, más repetibles, más manejables y más rápidos.

La metodología CRISP-DM se describe en términos de un modelo de proceso jerárquico, que comprende cuatro niveles de abstracción (de general a específico): fases, tareas genéricas, tareas especializadas e instancias de procesos. (Wirth & Hipp,2000, p. 3)

Después de seleccionar la metodología estándar a utilizar, CRISP-DM, se realizará la implementación de la metodología para estratificar el Censo de Población y Vivienda 2010.

A continuación, en el tercer capítulo se desarrolla la primera fase de la metodología CRISP-DM conocida como la comprensión del negocio, en esta parte se conoce acerca de la institución, los objetivos y el recurso humano involucrado en la ejecución del proyecto, además se expone cómo fue efectuada la estratificación en el año 2010.

En el capítulo cuatro, se pone en marcha la fase de comprensión de los datos con cada una de las tareas que esta implica, tales como recopilación, descripción, exploración y calidad de los datos. En el capítulo cinco, se efectúan las fases de preparación de los datos con la ayuda de un software ETL llamado *Data Integration* de Pentaho y la fase de modelamiento.

En el capítulo seis, se realiza la fase de evaluación y distribución del modelo, donde se obtiene los resultados y se compara su efectividad frente a la metodología usada por el INEC. Para finalizar, en el capítulo siete se expone las conclusiones obtenidas del proyecto y se proporciona recomendaciones para el uso de la metodología propuesta e investigaciones futuras.

CAPÍTULO III

COMPRENSIÓN DEL NEGOCIO

Para iniciar es preciso y necesario dedicar tiempo a explorar las expectativas de la institución con respecto a la minería de datos en el proyecto de estratificación, es por ello, que se debe implicar a la mayor cantidad de personas que sea posible en las discusiones, para luego documentar los resultados con esto se asegura la asignación de recursos en cada paso del proyecto. En la comprensión del negocio y la adecuada documentación del mismo se encuentra la base para el éxito del proyecto

3.1. DETERMINACIÓN DE LOS OBJETIVOS COMERCIALES

El primer objetivo a alcanzar es obtener la mayor cantidad de información posible de los objetivos comerciales de la minería de datos. Probablemente esta no sea una tarea fácil como parece, pero puede reducir el riesgo aclarando problemas, objetivos y recursos.

3.2. COMPILACIÓN DE LA INFORMACIÓN DE LA EMPRESA

La comprensión de la situación de la institución ayudará a conocer en términos de:

- Recursos disponibles (personal y material)
- Problemas
- Objetivos

Se debe explorar la situación para hallar soluciones a situaciones que puedan afectar al proyecto de minería de datos. A continuación, información relevante respecto de la institución.

3.2.1. Estructura de la organización

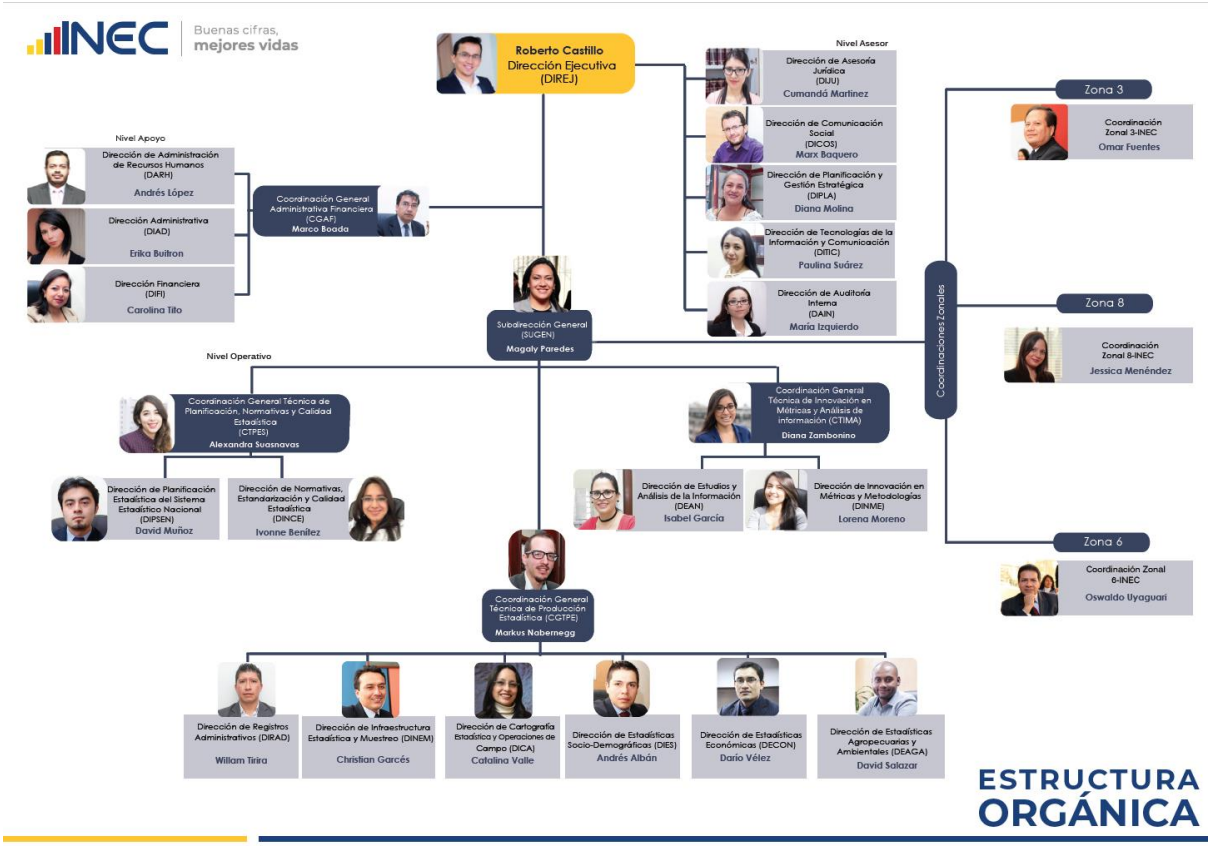


Figura 4. Organigrama del Instituto Nacional de Estadística y Censos
Fuente: (INEC, 2018)

Personal más importante para la consecución del proyecto:

- Markus Nabernegg – Coordinador General Técnico de Producción Estadística
  Encargado de la aprobación de los proyectos
- Christian Garcés – Dirección de Infraestructura Estadística y Muestreo
  Encargado de la implementación del proyecto de mejora del marco muestral

Unidades directamente afectadas e involucradas en el proyecto: Dirección de Infraestructura Estadística y Muestreo.

El INEC busca obtener los mejores resultados estadísticos en cuanto a la situación de la población, y así puedan actuar otras instituciones estatales en atacar los mayores problemas que aquejan

El Marco de Muestreo es uno de los insumos más importantes al hacer una encuesta, sin embargo, en las estadísticas sociodemográficas y encuestas dirigidas a la población, no se cuenta con un marco actualizado, ya que el marco se construye a partir de los censos que se realizan cada diez años.

Solución actual: usar la metodología creada en el año 2013

Ventajas: cumple con la función otorgar estratos en el marco de muestreo.

Desventajas: el tiempo de ejecución.

Metodología no replicable.

Se desconoce todas las medidas tomadas para su construcción.

3.2.2. Metodología de estratificación actual

En esta sección se da a conocer la metodología utilizada por el INEC en el año 2013 para la construcción del marco muestral maestro, específicamente la estratificación que se hizo a la población para segmentar en los estratos alto, medio y bajo. Para esto, nos basaremos en la documentación existente de la metodología utilizada, la cual se encuentra en el sitio web del INEC, además de la explicación recibida por parte del equipo de diseño muestral de la Institución. De esta manera, estableceremos los procesos que fueron llevados a cabo y la forma en cómo se realizó cada uno de la manera más detallada posible con la información que se cuenta en la actualidad.

Se ha identificado que para conocer mejor la metodología existen cuatro puntos esenciales a tratar: la finalidad del marco de muestreo y su estratificación, la gestión de la fuente de información, la aplicación de algoritmos estadísticos para la estratificación, y, los resultados obtenidos.

La finalidad del marco de muestreo y su estratificación

Para empezar, es necesario conocer la definición de lo que es un marco de muestreo, de acuerdo a la Comisión Económica para América Latina y el Caribe (CEPAL):

> Un marco de muestreo se define como una lista exhaustiva - organizada en forma de base de datos- que contiene todos y cada uno de los elementos de una población de interés que participarán en las distintas fases de selección de la muestra. A su vez, el marco está formado por un conjunto de mapas y planos a diferentes escalas, que permiten la delimitación física de las diversas unidades de selección. Asimismo, se considera parte fundamental de éste los registros físicos de las unidades de viviendas, así como los listados en que se detallan las referencias que permiten a los encuestadores la plena identificación de las viviendas seleccionadas, así como de aquellas que sin ser parte de la muestra pertenecen a la población objeto de estudio. (CEPAL, 2001).

Es decir, al hablar de un marco de muestreo se hace referencia a un listado de unidades, en la que se procura que se encuentren todas las unidades de la población con sus características de identificación. Dicho esto, el marco de muestreo al que se hace referencia en este documento es el marco de muestreo de viviendas obtenido a partir del Censo de Población y Vivienda 2010.

El marco de viviendas es utilizado para las encuestas dirigidas a hogares, entre estas encuestas se encuentran:

- Encuesta Nacional de Empleo, Desempleo y Subempleo (ENEMDU).
- Encuesta Nacional de Ingresos y Gastos de los Hogares Urbanos y Rurales (ENIGHUR).
- Encuesta de Condiciones de Vida (ECV).
- Encuesta de Victimización y Percepción de Inseguridad.
- Encuesta de Uso del Tiempo.
- Encuesta de Violencia de Género.

Todas estas encuestas tienen una misma naturaleza, son encuestas sociodemográficas, es decir, investigan fenómenos sociales de la población.

Teniendo presente que las encuestas se realizan con una muestra que represente a la población total, es importante garantizar para encuestas sociodemográficas, que todos los estratos sociales intervengan en la investigación. Para esto la muestra debe estar distribuida en cada uno de estos

estratos, para esto es necesario contar en el marco de muestreo con una variable que identifique el estrato que tienen las unidades a ser seleccionadas.

> Los proyectos que principalmente hacen uso del marco, de manera continua, periódica o especial, tienen por objetivo generar información estadística de interés nacional acerca de diversos temas que acontecen en nuestro país: empleo, ocupación, gastos, ingresos, trabajo infantil, género, salud y nutrición, uso del tiempo, educación, inserción laboral, cultura, política, seguridad ciudadana, percepción de diversos temas, características de viviendas y hogares, ciencia y tecnología, acceso a programas sociales, etc. El contar con una clasificación previa de las UPM del país de acuerdo a la similitud de los fenómenos descritos anteriormente permite generar muestreos más eficientes. (INEC, 2013, pág. 7)

De esta manera el INEC tiene una variable con los estratos “Alto”, “Medio” y “Bajo”, para su construcción se realizan algunos procedimientos que se verán en las siguientes secciones.

Gestión de la fuente de información

Como se mencionó anteriormente la fuente para la construcción del marco muestral es el “Censo de Población y Vivienda 2010”, como se puede observar en el cuestionario censal (ver ANEXO 1), existen cuatro secciones de preguntas:

- Datos de la vivienda
- Datos del hogar
- Remesas y emigración
- Datos de población

En las que la unidad de investigación son las viviendas, hogares y personas. Todos estos datos convergen para la creación de variables que intervienen en la estratificación.

Dimensiones	Variables
Características de la vivienda	Materiales del techo
	Materiales de paredes exteriores
	Materiales del piso
	Procedencia del agua
Acceso a Servicios Básicos	Recibe el agua
	Servicio higiénico de la vivienda
	Servicio de luz de la vivienda
	Eliminación de basura de la vivienda
Hacinamiento	Hacinamiento
Educación	Principal combustible del hogar
	Telefono convencional del hogar
	Telefono celular del hogar
	Servicio de internet del hogar
	Computadora del hogar
	Servicio de tv cable del hogar
	Espacio para cocinar del hogar
	Servicio higiénico del hogar
	Ducha del hogar
Patrimonio de los Hogares	Ocupados
	Población en edad de trabajar
	Población económicamente activa
	Tasa de dependencia del hogar
	Alfabetismo
Capacidad de generación de ingreso de los hogares	Escolaridad
	Escolaridad del jefe del hogar

Figura 5. Dimensiones y Variables que intervienen en la estratificación
Fuente: Metodología del Diseño Muestral de la ENEMDU (INEC, 2013).

En la metodología publicada al parecer hay un error en el orden de las tres últimas dimensiones, ya que no corresponden conceptualmente a las variables de la derecha. Sin embargo, se entiende las características que influirán en la asignación del estrato de la Unidad Primaria de muestreo (UPM).

La UPM es la unidad a la que se le asigna el estrato, de acuerdo al INEC, las UPM son los sectores censales, una subdivisión estadística relativamente permanente de un lugar delimitada por un grupo local de usuarios de datos censales con el propósito de presentar datos. Los límites de sectores censales coinciden con rasgos visibles, pero, en ciertas ocasiones, pueden concordar con límites de unidades gubernamentales y otros rasgos no visibles.

En el Ecuador existen dos tipos de sectores censales, el sector censal disperso que se encuentra en el área rural y que “Es una extensión razonable de territorio con límites perfectamente definidos, identificada por un nombre y un número. Un sector censal disperso está constituido por un promedio de 80 viviendas” (INEC, 2017, p. 22), y el sector amanzanado, referente al área urbana “que es una superficie perfectamente delimitada y continua geográficamente, constituida por una o más manzanas”. (INEC, 2017, p. 22).

Esta clasificación es considerada al momento de estratificar la población. Además, la estratificación respeta los límites geográficos de las provincias del Ecuador, diferenciando las reglas de decisión dentro de cada una, tanto para la zona urbana como para la rural.

Es decir, que con las variables mencionadas anteriormente se realiza la estratificación a nivel de sector, considerando que dentro de un sector existe un número determinado de viviendas, hogares y personas. Para esto es necesario construir indicadores a nivel de sector, por ejemplo:

> Las variables se transforman en indicadores con sentido de acceso o tenencia. Si la vivienda posee la característica de interés asume el valor 1 y cero en otro caso. Posteriormente, se suman los valores y el valor obtenido se divide entre el total de viviendas del sector censal. De esta manera se obtiene, por ejemplo, el porcentaje de viviendas en un sector determinado con acceso a agua potable o con baño para uso exclusivo del hogar. (INEC, 2013, pág. 9)

Para el tratamiento de las tres bases de datos (Vivienda, hogar y población), que intervienen en la construcción de los indicadores, se utiliza el software estadístico *Statistical Package for the Social Sciences* (SPSS). En este software se realiza cada paso por comandos dados en la barra de menú o por sintaxis o código. La desventaja es que si se realiza por comandos y no se guarda el código de programación no se puede replicar el ejercicio y por lo tanto no se puede validar el proceso realizado para obtener los resultados, se desconoce las decisiones que tomó el elaborador para la validación de la información, por ejemplo, el tratamiento de valores faltantes y atípicos. Al momento el INEC no cuenta con la sintaxis de cómo fue realizada la estratificación.

La aplicación de algoritmos estadísticos para la estratificación

Para realizar la estratificación se ha identificado tres pasos a seguir, el primero es la construcción de indicadores a nivel de UPM, el segundo identificar las variables que intervienen en la estratificación y el tercero aplicación de técnicas multivariadas.

Construcción de indicadores a nivel de UPM

Una vez construidos los indicadores a nivel de persona, hogar o vivienda se procede a crear indicadores a nivel de sector censal o UPM, estos corresponden a puntajes que van de 0 a 100 de acuerdo a la proporción de las viviendas que tienen una característica dentro de una UPM:

> Los indicadores expresados en forma de porcentajes permiten conocer la cobertura alcanzada en las distintas dimensiones del bienestar: porcentaje de viviendas con acceso a agua entubada, con vivienda digna, con acceso a luz, baño propio, etc. En la medida que se acerca a 100, estarían indicando adecuada cobertura y en caso contrario permiten identificar los retos que prevalecen en materia de calidad de vida. (INEC, 2013, pág. 9)

Es importante identificar los indicadores que influyen en el nivel socioeconómico, ya que la estratificación tiene como fin diferenciar las características que existe entre estratos, de esta manera, indicadores en los que se tiene que cerca del 100% de los hogares cuentan con una característica, esta no ayudaría a diferenciar los estratos.

> Niveles cercanos al 100% estarían señalando que el indicador analizado no permite generar diferencias entre hogares. Por ejemplo, si el porcentaje de hogares con luz eléctrica es 98%, esta variable no es útil para la estratificación, debido a que la mayoría de los hogares no presenta situación de privación en esta variable. (INEC, 2013, pág. 9)

Además de indicadores de tipo porcentual se tiene "variables continuas: escolaridad del jefe del hogar, por ejemplo. Existen, a su vez, variables continuas que representan proporciones, como tasas de dependencia". (INEC, 2013, pág. 9)

Identificar las variables a intervenir en la estratificación

Para identificar las variables que pueden intervenir en la estratificación es necesario realizar una matriz de correlaciones parciales, "Un primer criterio, es elegir variables que estén altamente correlacionadas entre sí, y que además tengan relación con el fenómeno de estudio (matriz de correlaciones parciales)". (INEC, 2013, pág. 9) . Además, es importante garantizar que la relación lineal entre variables no sea alta, ya que se puede estar redundando en variables que reflejan un mismo fenómeno. "Para medir el grado de asociación lineal entre dos variables se utilizó el concepto de Correlación de Pearson (ρ) y fueron identificadas aquellas variables que obtuvieron un valor $\rho > 0.5$ como fuertemente relacionados, este procedimiento se realizó de manera

independiente para cada uno de los 46 dominios de estratificación establecidos". (INEC, 2013, pág. 10).

En este punto, es necesario aclarar lo que es un dominio para el INEC, "Los dominios de estratificación del MMM considerados son cada una de las 23 provincias continentales, divididas en sus componentes urbanas y rurales" (INEC, 2013, pág. 7).

Aplicar técnicas multivariadas

Para este paso, el documento de diseño muestral lo resume todo en un párrafo:

> El proceso de estratificación consiste en la aplicación de técnicas multivariadas que permiten asignar cada UPM del país en una de las categorías llamadas "estratos". Estos estratos agrupan a su interior UPM "similares" con respecto a los indicadores elegidos previamente. El proceso de estratificación se sustenta en el algoritmo de las k-medias para variables continuas. Una vez definido el número óptimo de estratos para cada uno de los dominios establecidos, se ha caracterizado los estratos utilizando para ello las variables utilizadas en el proceso. El propósito es asociar las características de los estratos a un nivel determinado de bienestar en una escala categórica (bajo, medio, y alto). (INEC, 2013, pág. 10)

Por lo que lo único que se conoce con respecto, es que se aplicó el algoritmo de K-medias, y no se conocen las decisiones que se tomaron a lo largo de la ejecución del algoritmo.

Resultados obtenidos

En resumen, la metodología seguida por el INEC para la estratificación del Censo de Población y Vivienda 2010 es la siguiente:

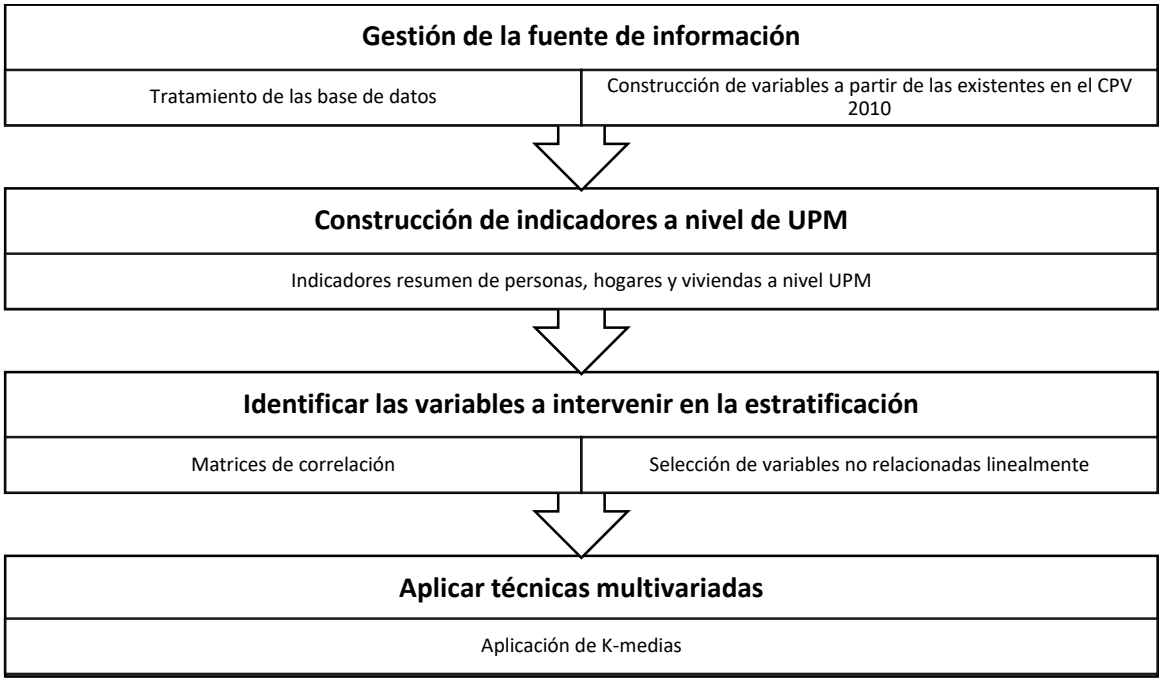


Figura 6. Metodología seguida por el INEC para la estratificación

Finalmente, el producto obtenido es el marco de muestreo a nivel de UPM con el estrato que le corresponde de acuerdo al dominio al que pertenece, es decir, de acuerdo a la provincia en la que se ubica la UPM y el área, sea esta urbana o rural.

3.3. OBJETIVO ESTRATÉGICO

El Instituto Nacional de Estadística y Censos como institución responsable de la estadística oficial, es la entidad encargada de planificar, normar y certificar la producción del Sistema Estadístico Nacional, además de producir información estadística pertinente, oportuna, confiable y de calidad; e, innovar en metodologías, métricas y análisis de información estadística necesaria para el diseño, implementación y evaluación de la planificación nacional.

3.4. POLÍTICAS

Operar como centro oficial general con información de datos estadísticos del país.

Coordinar y supervisar la ejecución de los programas y planes de trabajo que deben realizar las instituciones del Sistema Estadístico nacional (SEN).

Hacer inventarios estadísticos y mantener un archivo centralizado de las metodologías y el instrumental de investigación que utiliza el SEN.

Preparar y actualizar la cartografía estadística necesaria para la ejecución de las investigaciones que realizan las dependencias que conforman el SEN.

Realizar los censos de población y vivienda, agropecuarios, económicos y otros, publicar y distribuir sus resultados.

Difundir la información estadística en forma oportuna, a través de medios impresos y magnéticos a personas o entidades públicas y privadas a nivel nacional o internacional. (INEC, 2018)

3.5. DEFINICIÓN DE LOS OBJETIVOS COMERCIALES

Se debe especificar las soluciones planteadas en reuniones con la gente involucrada y la investigación realizada a fin de encontrar un objetivo principal concreto el cual fue acordado por las unidades comerciales y patrocinadores que tendrán directa afectación por los resultados del proyecto.

3.6. VALORACIÓN DE LA SITUACIÓN

Teniendo claro el objetivo comercial, se puede realizar la valoración de la situación actual analizando cuestiones como:

- Tipos de datos disponibles
- Personal necesario para completar el proyecto
- Factores de riesgo y planes de contingencia

3.7. DETERMINACIÓN DE LOS OBJETIVOS DE MINERÍA DE DATOS

En este paso con el objetivo comercial se llevará a cabo la interpretación en términos de minería de datos

La tecnología y la empresa deben ir de la mano para que el Proyecto de minería de datos sea efectivo.

Objetivos de minería de datos

Se debe definir una solución técnica a través de los analistas comerciales y de datos de la siguiente manera:

- Describir el problema de minería de datos como conglomerado, pronóstico o clasificación
- Documente objetivos técnicos utilizando unidades específicas de tiempo, como predicciones con una validez de tres meses.
- Si es posible, proporcione datos reales para resultados deseados, como producir resultados de abandono para el 80 % de los clientes actuales.

3.8. PRODUCCIÓN DEL PLAN DE PROYECTO

El plan de proyecto es el documento principal del trabajo de minería de datos. Si se elabora correctamente, puede informar a todos los usuarios relacionados con los objetivos, recursos, riesgos del proyecto y programar todas las fases de minería de datos. Es posible que desee publicar el plan, así como la documentación recopilada en esta fase en la intranet de la empresa.

Tabla 3
Plan de Proyecto

Fases	Tiempos	Recursos	Riesgos
Comprensión del negocio	1 semana	Todos los analistas	Cambios en estructuras de mando, difícil acceso a información
Comprensión de los datos	2 semanas	Todos los analistas	Problemas de datos, problemas tecnológicos
Preparación de los datos	2 semanas	Asesor de minería de datos, tiempo de análisis de base de datos. Desarrollador de sistemas.	Problemas de datos, problemas tecnológicos

CONTINÚA=>

Modelamiento	**3 semanas**	**Asesor de minería de datos, tiempo de análisis de base de datos**	**Problemas de tecnología, incapacidad para encontrar un modelo adecuado**
Evaluación	1 semana	Todos los analistas	Cambio económico, incapacidad para implementar resultados
Distribución	1 semana	Asesor de minería de datos, tiempo de análisis de base de datos	Cambio económico, incapacidad para implementar resultados

Para la implementación del proyecto se recomienda contar con un equipo conformado por:

- Asesor de minería de datos
- Desarrollador de sistemas.
- Especialista en Diseño Muestral
- Especialista en Censos y Demografía

3.9. Valoración de herramientas y técnicas

A lo largo del desarrollo de la metodología propuesta se utilizará los siguientes programas:

- SPSS: Formato inicial en el que se encuentran las bases de datos del INEC, además facilita el análisis estadístico de la calidad de los datos.
- *Pentaho Data Integration*: Con este software se realiza el pre procesamiento de los datos, se integra las distintas bases de datos y se construye los indicadores que conformarán el Dataset.
- *Rapid Miner*: Esta herramienta, permite aplicar las distintas técnicas de modelado y evaluar los resultados obtenidos
- R: *Software* libre que permite realizar análisis estadístico mediante comandos, así como administrar las bases de datos.
- Q-GIS: Sistema de Información Geográfica que en este caso permitirá la visualización de la localización de los estratos.

Para el funcionamiento correcto de estas herramientas es necesario contar con al menos 16 GB de memoria RAM en la computadora en la que se procese la información.

Con respecto a las técnicas de modelado, se evaluará las posibles opciones de algoritmos de agrupación por particiones en el programa Rapid Miner.

CAPÍTULO IV

COMPRENSIÓN DE LOS DATOS

En la fase de comprensión de datos se busca el aprendizaje de los datos que serán usados en la minería. Es extremadamente importante que el estudio de los datos este claro pues se evadirá problemas posteriores y permitirá la exploración a través de tablas y gráficos.

4.1. RECOPILACIÓN DE DATOS INICIALES

A partir de este punto se debe admitir el uso de los datos. Existen diferentes origines para los datos y se los encasilla de la siguiente manera:

- Datos existentes. Son los datos transaccionales, datos de encuesta, registros Web, etc. Comprobar si con los datos existentes es suficiente para el proyecto
- Datos adquiridos. Considerar datos que se puedan incluir si la institución no maneja los suficientes para el proyecto.
- Datos adicionales. Si resulta necesario otras fuentes de datos como encuestas estas resultarían como complemento a los almacenes de datos disponibles.

4.1.1. Informe de recopilación de datos

Las fuentes de información provienen del Censo de Población y Vivienda, para obtener esta información es necesario pasar por la etapa de empadronamiento en la que se aplica el formulario censal a toda la población. Los datos son posteriormente digitalizados en el sistema y se ponen a disposición de los analistas y público en general las bases de datos cuidando la seguridad de datos sensibles como nombres y direcciones de los informantes.

En este caso las bases de datos a utilizarse como personas externas al INEC son de información pública y extraídas de la página web del Instituto Nacional de Estadística y Censos bajo el link http://www.ecuadorencifras.gob.ec/base-de-datos-censo-de-poblacion-y-vivienda-2010-a-nivel-de-manzana/ .

Base de Datos existentes. Se cuenta con las bases de datos de Encuestas de Vivienda, Población y Hogar extraídas de la página oficial del INEC, en su versión para SPSS las cuales son:

- **CPV2010_Spss_Vivienda.sav.-** Contiene la información de todas las viviendas que son registradas en territorio nacional al momento en el que se realizó el Censo.
- **CPV2010_Spss_Hogar.sav.-** Se considera hogar al conjunto de personas que duermen en una misma vivienda y cocinan sus alimentos en forma conjunta y comparten un mismo gasto para la comida. (INEC, 2010) Bajo este concepto dentro de una vivienda puede existir más de un hogar. Los registros de la base de datos corresponden a las características propias de cada hogar.
- **CPV2010_Spss_Poblacion.sav.-** En esta base consta la información de toda la población registrada en el Censo 2010 a nivel de persona.

Adicionalmente para la estratificación no se utiliza datos adicionales, ni se adquiere ningún otro tipo de información.

4.2. DESCRIPCIÓN DE LOS DATOS

Para describir la mayoría de los datos se centra en cantidad y calidad de los datos, cantidad de datos disponible y estado de los mismos. Algunas características importantes para describir los datos son las siguientes:

- Cantidad de datos. En las técnicas de modelamiento, los tamaños de datos son relacionados, es decir, mientras el conjunto de datos es más grande puede producir un resultado más preciso, sin embargo, puede aumentar el tiempo de procesamiento, por lo tanto, es necesario considerar un subconjunto de datos, para lo cual se torna importante tomar nota de los tamaños de los conjuntos de datos, el número de registros como los campos cuando se escriba el informe final.
- Tipos de valores. Los datos pueden tener gran diversidad de formatos como tipo cadena, numéricos, boleanos, etc. Es necesario poner atención al tipo de valor para evitar posteriores problemas en la fase de modelamiento

- Esquemas de codificación. En ocasiones, se incluye representaciones de atributos como el género representado en algunos casos con M y F pero en otros con 1 y 2, por tanto, es necesario registrar los esquemas discordantes en el informe de datos. (IBM, 2012)

4.2.1. Informe de descripción de datos

Es necesario describir los datos de cada una de las bases de datos para comprender mejor su contenido y formatos de los datos, para esto se considera como parámetros el número de registros, el número de variables, los nombres y descripciones de cada variable y el formato de las mismas. Además, se incluye la descripción de los valores que puede tomar una variable.

En la base de viviendas, constan un total de 4’654.309 de viviendas registradas en el Censo 2010 para las que se investigaron 35 variables tanto nominales como escalares. (Ver ANEXO 2).

En el caso de la base de hogares, se registró un total de 3’815.527 hogares y se investigaron 34 variables tanto nominales como escalares. (Ver ANEXO 3).

En la base de datos de población, hay un total de 14’483.499 personas y se investigaron 100 variables tanto nominales como escalares. (Ver ANEXO 4).

4.3. EXPLORACIÓN DE DATOS

CRISP-DM se vale de esta fase para explorar los datos con las tablas, gráficos y otras herramientas de visualización disponibles. Esto facilita la descripción de los objetivos de minería de datos generados durante la fase de comprensión comercial. También puede apoyar en la formulación de la hipótesis y ayuda a configurar las tareas de transformación de datos que se realizan durante la preparación de los datos.

4.3.1. Informe exploración de datos

Para comenzar es importante comprender la función de las variables en las bases de datos, las variables de identificación que serán parte de las claves que permitirán anclar entre las bases de datos son las variables de localización geográfica, estas variables son las que tienen en su nomenclatura como letra inicial la "I":

Tabla 4
Variables de identificación

I01	PROVINCIA
I02	CANTON
I03	PARROQUIA
I04	ZONA
I05	SECTOR
I09	No. DE VIVIENDA
I10	No. DE HOGAR

En la base de datos de vivienda las variables se identifican por empezar la nomenclatura con la letra "V". Es necesario identificar las variables que aportarán a futuros análisis, para esto en primer lugar se considerarán las variables utilizadas de acuerdo a la metodología del INEC, además se adicionará posibles variables que puedan influir en la estratificación, y las variables que no aporten en la investigación serán excluidas. Existen trece variables que son utilizadas en la estratificación realizadas por el INEC, de manera directa o en forma de un indicador construido, se propone evaluar la introducción de cuatro variables que podrían aportar al estudio y diez variables serán excluidas de análisis ya que teóricamente no serían factores de influencia en el estrato que será asignado.

Tabla 5
Variables seleccionadas de la base de viviendas

Nombre	Descripción	INEC	Propuesta	Excluida
VTV	TIPO DE VIVIENDA		X	
VAP	VIA DE ACCESO PRINCIPAL A LA VIVIENDA		X	
VCO	CONDICIÓN DE OCUPACIÓN DE LA VIVIENDA			X
V123	Materiales de la vivienda			X
V01	Material predominante del techo o cubierta de la vivienda	X		
V03	Material predominante de las paredes exteriores de la vivienda	X		
V05	Material predominante del piso de la vivienda	X		
V02	Estado del techo de la vivienda			X
V04	Estado de las paredes de la vivienda			X

CONTINÚA=>

Nombre	Descripción	INEC	Propuesta	Excluida
V06	Estado del piso de la vivienda			X
V07	De donde proviene principalmente el agua que recibe la vivienda	X		
V08	El agua que recibe la vivienda es	X		
V09	El servicio higiénico o escusado de la vivienda es	X		
V10	El servicio de luz (energía) eléctrica de la vivienda proviene principalmente	X		
V11	Dispone la vivienda de medidor de energía eléctrica			X
V12A	Cuantos focos ahorradores tiene su vivienda		X	
V12B	Cuantos focos convencionales tiene su vivienda		X	
V13	Principalmente como elimina la basura	X		
V14	Sin contar la cocina, el baño y cuartos de negocio. Cuántos cuartos tiene la vivienda	X		
V15	Todas las personas que duermen en esta vivienda, cocinan sus alimentos en forma conjunta y comparten un mismo gasto para			X
V16	Cuantos grupos de personas(hogares) duermen en su vivienda y cocinan los alimentos por separado incluya su hogar			X
TOTPER	Total de personas de la vivienda	X		
TOTDOR	Total de dormitorios de la vivienda	X		
TOTEMI	Total de emigrantes			X
PERCUA	Número de personas por cuarto	X		
PERDOR	Número de personas por dormitorio	X		
VIVREM	Viviendas con remesas			X

En la base de datos de hogares las variables se identifican por empezar la nomenclatura con la letra “H”. Diez variables son utilizadas por el INEC para la estratificación. Se propone el análisis de siete variables que se pueden introducir y se excluye nueve variables de la base.

Tabla 6

Variables seleccionadas de la base de hogares

Nombre	Descripción	INEC	Propuesta	Excluida
H00	Secuencial de Hogar	X		
H01	Del total de cuartos de este hogar, Cuantos son exclusivos para dormir		X	
H01N	Ningún Dormitorio			X
H02	Tiene este hogar cuarto o espacio exclusivo para cocinar	X		
H03	El servicio higiénico o escusado que dispone el hogar es	X		
H04	Dispone este hogar de espacio con instalaciones y/o ducha para bañarse	X		
H05	Cuál es el principal combustible o energía que utiliza este hogar para cocinar	X		
H06	Principalmente, el agua que toman los miembros del hogar		X	
H07	Dispone este hogar de servicio de teléfono convencional	X		
H08	Algún miembro de este hogar dispone de servicio de teléfono celular	X		
H09	Dispone este hogar de servicio de internet	X		
H10	Dispone este hogar de computadora	X		
H11	Dispone este hogar de servicio de televisión por cable	X		

CONTINÚA=>

Nombre	Descripción	INEC	Propuesta	Excluida
H12	Cuanto fue el último pago que realizó el hogar por concepto de luz eléctrica		X	
H12NP	No paga la luz eléctrica		X	
H13A	Algún miembro de este hogar se traslada fuera de esta ciudad o parroquia rural para trabajar			X
H13B	Cuantos se trasladan fuera de esta ciudad o parroquia para trabajar			X
H14A	Algún miembro de este hogar se traslada fuera de esta ciudad o parroquia rural para estudiar			X
H14B	Cuantos se trasladan fuera de la ciudad o parroquia rural para estudiar			X
H15	La vivienda que ocupa este hogar es	X		
M1	Durante el año 2010, Alguna persona de este hogar recibió dinero por parte de familiares o amigos que viven en el exterior			X
M2A	A partir del último censo de población y vivienda(Noviembre 2001) una o más personas que vivían en este hogar viajaron a			X
M2B	Cuantos personas viajaron al Exterior			X
TP1	Total Personas 1		X	
TH1	Total Hombres 1		X	
TM1	Total Mujeres 1		X	

En la base de datos de personas las variables se identifican por empezar la nomenclatura con la letra “P”. En la estratificación del INEC se utilizan nueve variables de personas. Se propone introducir dieciocho variables al análisis, algunas de las variables que se propone son para inclusión en el marco de muestreo, para conocer las características de la población en cada UPM para las encuestas dirigidas a segmentos de la población como por ejemplo mujeres de una edad determinada de un grupo étnico. Sesenta y una variables son excluidas del análisis.

Tabla 7

Variables seleccionadas de la base de población

Nombre	Descripción	INEC	Propuesta	Excluida
P01	Cuál es el Sexo		X	
P02	Que parentesco o relación tiene con el/la jefe/a del hogar			X
P03	Cuantos años cumplidos tiene			X
P04M	Cuál es el mes en que nació			X
P04A	Cuál es el año en que nació			X
P05	Tiene cédula de ciudadanía ecuatoriana		X	
P06	Está inscrito en el Registro Civil			X
P07	Tiene seguro de salud privado		X	
P08	Tiene discapacidad permanente por más de un año			X
P091	Discapacidad intelectual			X
P092	Discapacidad Físico-Motora			X
P093	Discapacidad Visual			X
P094	Discapacidad Auditiva			X
P095	Discapacidad Mental			X

CONTINÚA=>

Nombre	Descripción	INEC	Propuesta	Excluida
P10	Asiste actualmente a un establecimiento de educación especial para personas con discapacidad			X
P11L	En dónde nació			X
P11A	En qué año llegó al Ecuador			X
P11	Provincia/País Cantón Parroquia de Nacimiento			X
P11P	Provincia de nacimiento			X
P11C	Cantón de nacimiento			X
P11Q	Parroquia de nacimiento			X
P12L	En qué lugar vive habitualmente			X
P12	Provincia/País Cantón Parroquia Que Vive Habitualmente			X
P12P	Provincia de residencia habitual			X
P12C	Cantón de residencia habitual			X
P12Q	Parroquia de residencia habitual			X
P13L	Hace 5 años(Noviembre 2005) En qué lugar vivía habitualmente			X
P13	Provincia/País Cantón Parroquia Que Vivía Hace 5 años			X
P13P	Provincia donde vivía hace 5 años			X
P13C	Cantón donde vivía hace 5 años			X
P13Q	Parroquia donde vivía hace 5 años			X
P141P	Habla Lengua Indígena el Papá			X
P141M	Habla Lengua Indígena la Mamá			X
P142P	Habla Lengua Castellano/Español el Papá			X
P142M	Habla Lengua Castellano/Español la Mamá			X
P143P	Habla Lengua Extranjera el Papá			X
P143M	Habla Lengua Extranjera la Mamá			X
P144P	No habla el Papá			X
P144M	No habla la Mamá			X
P151	Habla Lengua Indígena			X
P151C	Cual lengua indígena habla			X
P152	Habla Lengua Castellana			X
P153	Habla Lengua Extranjera			X
P154	No habla Ningún Idioma			X
P16	Como se identifica según su cultura y costumbres			X
P17	Cuál es la Nacionalidad o Pueblo indígena al que pertenece			X
P181	El/la niño/a participa en el Programa del INFA			X
P182	El/la niño/a participa en el Programa del Ministerio de Educación			X
P183	El/la niño/a participa en el Programa del Centro Infantil Privado			X
P184	El/la niño/a participa en el Programa del Centro Infantil Público			X
P185	El/la niño/a participa en Otro Programa			X
P186	El/la niño/a Le cuida la madre, el padre, familiares o conocidos gratis			X
P187	El/la niño/a paga a familiares o conocidos por el cuidado			X
P19	Sabe leer y escribir	X		
P20T	En los últimos 6 meses ha utilizado Teléfono Celular		X	
P20I	En los últimos 6 meses ha utilizado Internet		X	
P20C	En los últimos 6 meses ha utilizado Computadora		X	
P21	Asiste actualmente a un establecimiento de enseñanza regular		X	
P22	El establecimiento de enseñanza regular al que asiste es		X	
P23	Cuál es el nivel de instrucción más alto al que asiste o asistió	X		
P24	Cuál es el grado, curso o año más alto que aprobó	X		

CONTINÚA=>

Nombre	Descripción	INEC	Propuesta	Excluida
P26	Que título tiene			X
P27	Qué hizo la semana pasada	X		
P28	Si NO ha trabajado	X		
P29	A qué se dedica o que hace el negocio o empresa en la que trabaja o trabajo			X
P291	Rama de actividad (1 Nivel)			X
P30	Que hace o que es en donde trabaja o trabajo			X
P31	En el lugar indicado trabaja o trabajo como			X
P32	Cuantas horas trabajo la semana pasada o la última semana que trabajó	X		
P33	El trabajo que realiza o realizó es o fue			X
P34	Estado conyugal			X
P35	Seguridad Social aporta o es afiliado		X	
P36	Cuantos hijos e hijas nacidos vivos ha tenido durante toda su vida		X	
P36H	Cuantos hijos nacidos vivos ha tenido durante toda su vida			X
P36M	Cuantas hijas nacidas vivas ha tenido durante toda su vida			X
P36NSN	No sabe cuántos hijo/as nacidos vivos ha tenido durante toda su vida			X
P37	Total de Hijos Vivos Actualmente			X
P37NS	No sabe cuántos hijos están vivos actualmente			X
P38	A qué edad tuvo su primer hijo o hija			X
P38NS	No sabe a qué edad tuvo su primer hijo			X
P39A	Año que tuvo su último hijo nacido vivo			X
P39M	Mes que tuvo su último hijo nacido vivo			X
P39NS	No sabe cuándo tuvo su último hijo nacido vivo			X
P40	Está vivo el último hijo o hija nacido vivo			X
TIPOACT	Tipo de actividad	X		
RAMCT	Rama de actividad recodificada			X
GRUOCU	Grupo de ocupación recodificado		X	
GRAESC	Grados de escolaridad	X		
GRAESCSA	Grados de escolaridad (Sistema anterior)	X		

4.4. CALIDAD DE DATOS

Se hace imprescindible el análisis de la calidad de los datos ya que en muchos casos contienen errores, existen valores faltantes etc. Dicha verificación busca mejorar la veracidad del modelado. En la revisión de calidad podrían presentarse los siguientes problemas.

Datos faltantes Valores vacíos o nulos

Errores de datos errores tipográficos al ingresar los datos.

Errores de mediciones basados en mediciones incorrectas

Incoherencias de codificación uso de no estándares para medidas

Metadatos erróneos errores en el significado del valor de un campo

4.4.1. Informe de Calidad de datos

Al revisar las bases de datos es necesario comprobar que las variables de identificación no contengan datos faltantes, errores de datos ni errores de mediciones. En este caso las variables de identificación guardan la calidad necesaria para poder utilizarlas como variables clave para anclar la información entre bases.

En la base de viviendas, como se puede observar en la siguiente tabla, existen valores faltantes para las distintas variables, y en el análisis de mínimos y máximos que tienen las variables se puede verificar que no existe errores de digitación, ya que los valores están entre los ya preestablecidos como respuestas de categorías válidas.

Tabla 8
Calidad de datos de las variables de vivienda

		N			
Variable	**Etiqueta**	**Válido**	**Faltantes**	**Mínimo**	**Máximo**
VTV	TIPO DE VIVIENDA	4654309	0	1	17
VAP	VIA DE ACCESO PRINCIPAL A LA VIVIENDA	4649330	4979	1	6
VCO	Condición DE Ocupación DE LA VIVIENDA	4649330	4979	1	4
V01	Material predominante del techo o cubierta de la vivienda	3748919	905390	1	6
V03	Material predominante de las paredes exteriores de la vivienda	3748919	905390	1	7
V05	Material predominante del piso de la vivienda	3748919	905390	1	7
V02	Estado del techo de la vivienda	3748919	905390	1	3
V04	Estado de las paredes de la vivienda	3748919	905390	1	3
V06	Estado del piso de la vivienda	3748919	905390	1	3
V07	De donde proviene principalmente el agua que recibe la vivienda	3748919	905390	1	5
V08	El agua que recibe la vivienda es	3748919	905390	1	4
V09	El servicio higiénico o escusado de la vivienda es	3748919	905390	1	6
V10	El servicio de luz (energía) eléctrica de la vivienda proviene principalmente	3748919	905390	1	5
V11	Dispone la vivienda de medidor de energía eléctrica	3493549	1160760	1	3
V13	Principalmente como elimina la basura	3748919	905390	1	6
V14	Sin contar la cocina, el baño y cuartos de negocio. Cuantos cuartos tiene la vivienda	3748919	905390	1	20

CONTINÚA=>

Variable	Etiqueta	N Válido	N Faltantes	Mínimo	Máximo
V15	Todas las personas que duermen en esta vivienda, cocinan sus alimentos en forma conjunta y comparten un mismo gasto para	3748919	905390	1	2
V16	Cuantos grupos de personas(hogares) duermen en su vivienda y cocinan los alimentos por separado incluya su hogar	3748919	905390	1	6
TOTDOR	Total de dormitorios de la vivienda	3748919	905390	0	20
TOTEMI	Total de emigrantes	3748919	905390	0	14
PERCUA	Número de personas por cuarto	3748919	905390	1	4
PERDOR	Número de personas por dormitorio	3748919	905390	1	5
VIVREM	Viviendas con remesas	3748919	905390	1	2

En las variables VAP y VCO existen 4.979 valores faltantes, sin embargo, los datos faltantes corresponden a los tipos de vivienda “Hotel, Pensión, Residencial u Hostal”, “Cuartel Militar o de Policía/Bomberos”, “Centro de rehabilitación social/Cárcel”, “Centro de acogida y protección para niños y niñas, mujeres e indigentes”, “Hospital, Clínica, etc.”, “Convento o Institución Religiosa”, “Asilo de Ancianos u orfanato”, “Otra vivienda colectiva” y “Sin Vivienda”. Por lo que los datos faltantes no son falta de información, sino que estas preguntas no están dirigidas a ese tipo de viviendas. Además, este tipo de viviendas no forman parte del marco muestral, por lo que deben ser descartadas del análisis.

En el caso de las variables que tienen 905.390 valores faltantes corresponden a las viviendas mencionadas anteriormente y viviendas con condición de ocupación “Ocupada con personas ausentes”, “Desocupada” y “En construcción”, viviendas que también deben excluirse del análisis.

Para el caso de la variable “Dispone la vivienda de medidor de energía eléctrica” existen 1´160.760 valores faltantes, de los cuales 905.390 corresponden a valores faltantes ya explicados en el párrafo precedente y 255.370 viviendas que no obtienen luz eléctrica de la “Red de empresa eléctrica de servicio público”. Por lo que no les corresponde responder a la pregunta.

En la base de datos de hogares, se presentan al menos 4.979 valores faltantes por variable, este valor corresponde a las viviendas que por su tipo no contienen hogares dentro, por lo que deben ser excluidos del análisis.

Tabla 9

Calidad de datos de las variables de hogar

Variable	Etiqueta	N Válido	N Faltantes	Mínimo	Máximo
H00	Secuencial de Hogar	3810548	4979	1	6
H01	Del total de cuartos de este hogar, Cuantos son exclusivos para dormir	3810548	4979	0	20
H01N	Ningún Dormitorio	194689	3620838	0	0
H02	Tiene este hogar cuarto o espacio exclusivo para cocinar	3810548	4979	1	2
H03	El servicio higiénico o escusado que dispone el hogar es	3810548	4979	1	3
H04	Dispone este hogar de espacio con instalaciones y/o ducha para bañarse	3810548	4979	1	3
H05	Cuál es el principal combustible o energía que utiliza este hogar para cocinar	3810548	4979	1	7
H06	Principalmente, el agua que toman los miembros del hogar	3810548	4979	1	5
H07	Dispone este hogar de servicio de teléfono convencional	3810548	4979	1	2
H08	Algún miembro de este hogar dispone de servicio de teléfono celular	3810548	4979	1	2
H09	Dispone este hogar de servicio de internet	3810548	4979	1	2
H10	Dispone este hogar de computadora	3810548	4979	1	2
H11	Dispone este hogar de servicio de televisión por cable	3810548	4979	1	2
H12	Cuanto fue el último pago que realizó el hogar por concepto de luz eléctrica	3084448	731079	1	9999
H12NP	No paga la luz eléctrica	468962	3346565	1	2
H13A	Algún miembro de este hogar se traslada fuera de esta ciudad o parroquia rural para trabajar	3810548	4979	1	2
H13B	Cuantos se trasladan fuera de esta ciudad o parroquia para trabajar	832101	2983426	1	25
H14A	Algún miembro de este hogar se traslada fuera de esta ciudad o parroquia rural para estudiar	3810548	4979	1	2
H14B	Cuantos se trasladan fuera de la ciudad o parroquia rural para estudiar	372987	3442540	1	10
H15	La vivienda que ocupa este hogar es	3810548	4979	1	7
M1	Durante el año 2010, Alguna persona de este hogar recibió dinero por parte de familiares o amigos que viven en el exterior	3810548	4979	1	2
M2A	A partir del último censo de población y vivienda(Noviembre 2001) una o más personas que vivían en este hogar viajaron al Exterior	3810548	4979	1	2
M2B	Cuantos personas viajaron al Exterior	186508	3629019	1	7

En la variable “Ningún dormitorio”, se presentan 3’620.838 valores faltantes, la variable contiene solo registros de 0 para los hogares que no tienen dormitorio, sin embargo, esta información no es necesaria ya que en la variable H01 se tiene ya la información por lo que se puede prescindir de esta variable.

La calidad de la variable “Cuanto fue el último pago que realizó el hogar por concepto de luz eléctrica” presenta varios problemas, tiene 731.079 valores faltantes, 97.216 con el código 9999 que no se especifica que significa y algunos valores atípicos tanto altos como bajos que se necesitaría corroborar. Esta variable se deberá analizar con cuidado si se utiliza para el análisis.

Las otras variables con más de 4.979 valores faltantes se deben a que son variables con filtros anteriores, es decir, variables que debían ser respondidas sólo por hogares que cumplan una característica.

Para finalizar con la base de hogares, las variables no tienen problemas de medición ni de datos como se puede verificar con mínimos y máximos a excepción de las ya mencionadas.

Tabla 10
Calidad de datos de las variables de población

		N			
Variable	**Etiqueta**	**Válido**	**Faltantes**	**Mínimo**	**Máximo**
P01	Cuál es el Sexo	14483499	0	1	2
P03	Cuantos años cumplidos tiene	14483499	0	0	120
P05	Tiene cédula de ciudadanía ecuatoriana	14483499	0	1	2
P07	Tiene seguro de salud privado	14483499	0	1	9
P12L	En qué lugar vive habitualmente	14483499	0	1	3
P16	Como se identifica según su cultura y costumbres	14483499	0	0	8
P17	Cuál es la Nacionalidad o Pueblo indígena al que pertenece	1018176	13465323	1	99
P19	Sabe leer y escribir	13021222	1462277	1	2
P20T	En los últimos 6 meses ha utilizado Teléfono Celular	13021222	1462277	1	9
P20I	En los últimos 6 meses ha utilizado Internet	13021222	1462277	1	9
P20C	En los últimos 6 meses ha utilizado Computadora	13021222	1462277	1	9
P21	Asiste actualmente a un establecimiento de enseñanza regular	13021222	1462277	1	2
P22	El establecimiento de enseñanza regular al que asiste es	4795641	9687858	1	4
P23	Cuál es el nivel de instrucción más alto al que asiste o asistió	13021222	1462277	1	99
P24	Cuál es el grado, curso o año más alto que aprobó	12366540	2116959	1	99
P27	Qué hizo la semana pasada	13021222	1462277	1	7
P28	Si NO ha trabajado	7176877	7306622	1	7
P32	Cuantas horas trabajo la semana pasada o la última semana que trabajó	5844345	8639154	1	999
P34	Estado conyugal	10839693	3643806	1	6
P35	Seguridad Social aporta o es afiliado	10839693	3643806	1	9
P36	Cuantos hijos e hijas nacidos vivos ha tenido durante toda su vida	5513467	8970032	0	99
P38	A qué edad tuvo su primer hijo o hija	3652893	10830606	12	99

En el caso de la base de datos de personas, existen demasiadas variables a revisar, por lo que se restringirá el análisis a las variables que en la fase de exploración se determinaron como variables a utilizar.

Existen variables con datos faltantes, en el caso de las variables P17, P22, P24, P28, P32, P36, y P38 son variables que tienen preguntas filtro y no están dirigidas a toda la población. Las variables P19, P20T, P20I, P20C, P21, P22, P23 y P27 son preguntas dirigidas a personas mayores a 5 años. Las variables P35 y P36 excluyen a los menores de 12 años.

En algunas variables se codifica con 9, 99 o 999 como se ignora, es decir, que el informante no sabe la respuesta. No se observa problemas en cuanto a la medición además no existen datos atípicos o códigos no predefinidos.

CAPÍTULO V

PREPARACIÓN DE LOS DATOS Y MODELAMIENTO

5.1. PREPARACIÓN DE LOS DATOS

De acuerdo a la metodología CRISP-DM “La preparación de datos es uno de los aspectos más importantes y con frecuencia que más tiempo exigen en la minería de datos. De hecho, se estima que la preparación de datos suele llevar el 50-70 % del tiempo y esfuerzo de un proyecto” (IBM, 2012, pág. 21).

En esta sección del capítulo se tratará las bases de datos para construir un *Dataset*, que contenga los indicadores creados a partir de las variables mencionadas en la fase de exploración de datos integrando las tres bases de datos con las que se cuenta, de manera que con el producto resultante se facilite el modelado de datos.

Para la selección de datos es importante considerar tanto la definición de que atributos o características se utilizarán, así como los elementos o individuos que son parte de la investigación, esto fue definido en la fase de comprensión de los datos, principalmente en las secciones de exploración de datos y calidad de datos.

En el caso de la selección de elementos, para las bases de población y hogares se utiliza la información de todos los elementos, y en el caso de la base de viviendas, se restringe la selección a viviendas particulares ocupadas con personas presentes.

En la selección de atributos se considera los seleccionados en el informe de exploración de datos, y posteriormente fueron transformadas para cargar el *Dataset*.

A continuación, se muestra las transformaciones realizadas en el software de D*ata Integration* de Pentaho. Para la base de viviendas fue necesario hacer dos transformaciones, la primera para la creación de identificadores y la transformación de variables nominales a variables binarias, con unos y ceros dependiendo de si cumplen una condición o no.

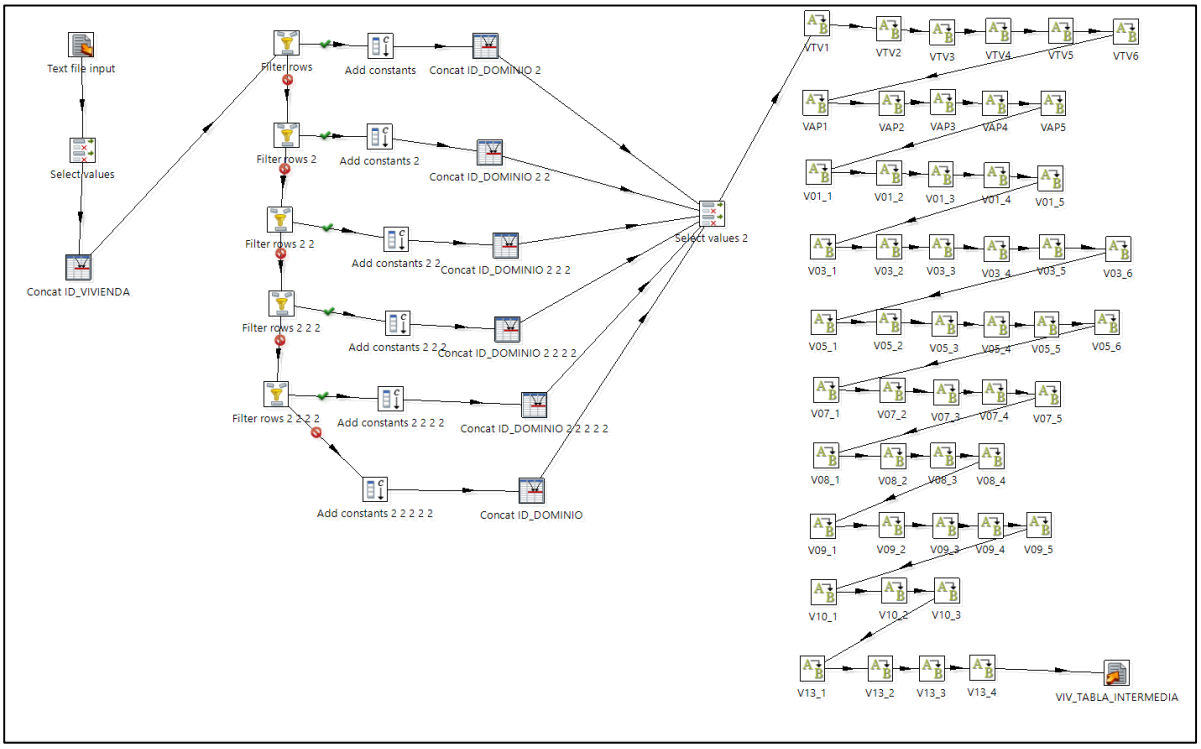


Figura 7. Transformación de la base de viviendas

La segunda transformación cumple la función de calcular indicadores y de agrupar las variables a nivel de la UPM o Sector censal.

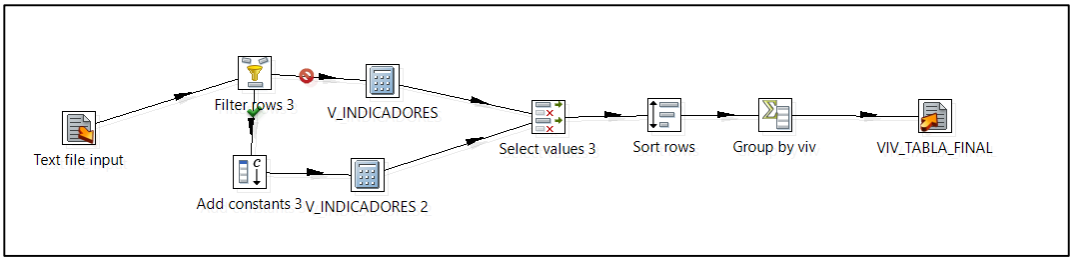


Figura 8. Transformación II de la base de viviendas

En la base de hogares, se efectuó una transformación para obtener las variables binarias.

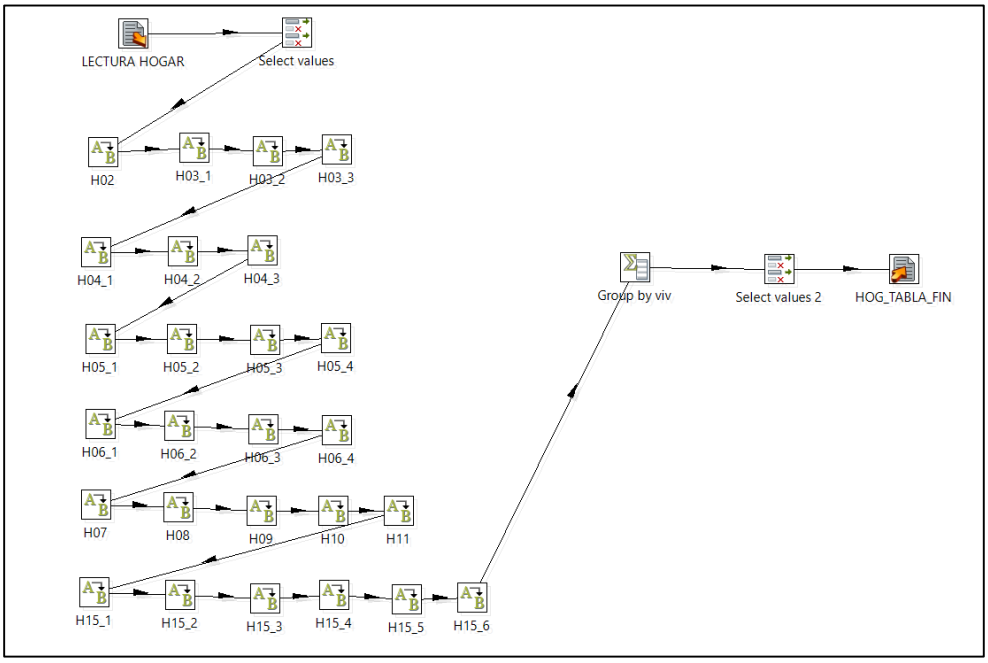


Figura 9. Transformación de la base de hogar

Para la base de población se realizó dos transformaciones, la primera para obtener las variables binarias que son parte de la investigación.

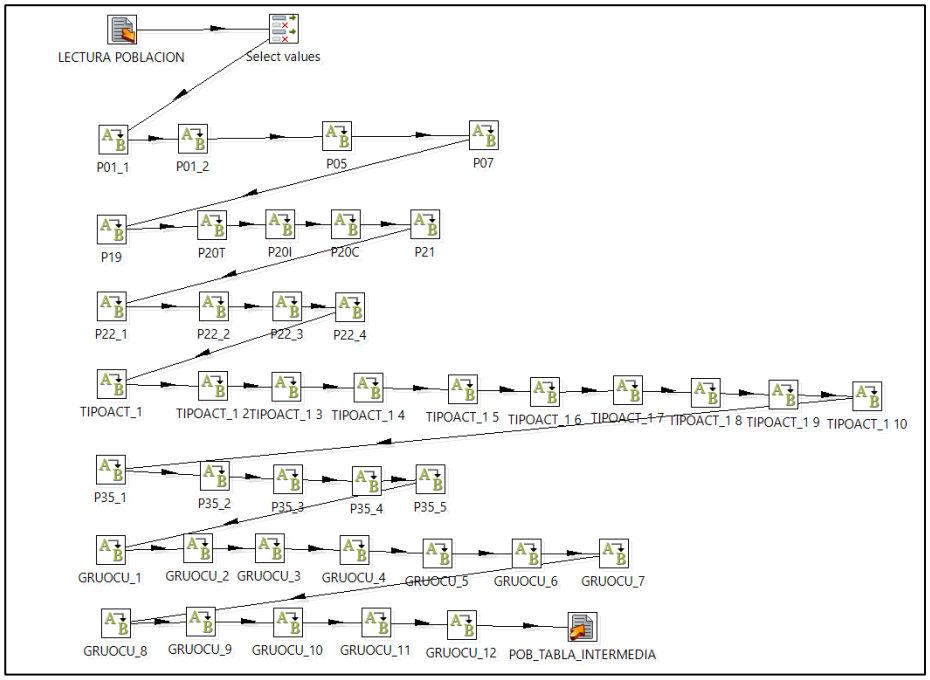


Figura 10. Transformación de la base de población

La transformación fue dividida por motivos de memoria de la computadora, la segunda transformación es la encargada de crear los indicadores a nivel de Sector.

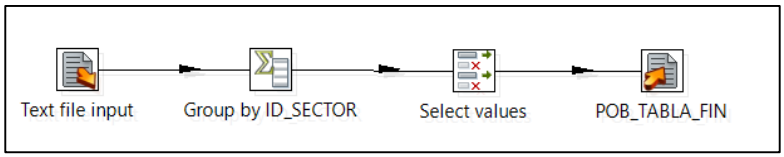


Figura 11. Transformación II de la base de población

Para la integración de los tres archivos planos se realizó una última transformación la cual consolida las variables mediante el identificador único ID_SECTOR.

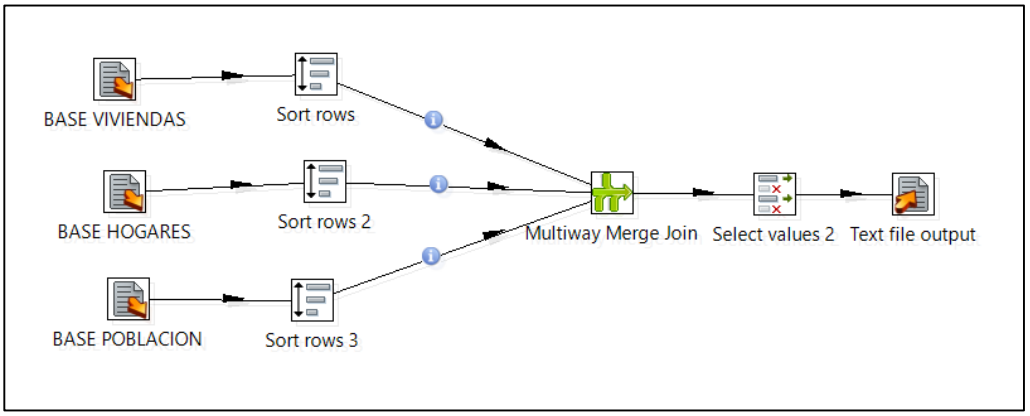


Figura 12. Transformación de integración de las bases

Posterior a las transformaciones se obtiene la base de trabajo para modelamiento, la cual contiene las siguientes variables:

Tabla 11
Variables finales del Dataset

NOMBRE	TIPO	NOMBRE	TIPO
PROVINCIA	String	H04_DUCHA_EXCLUSIVO	Number
CANTON	String	H04_DUCHA_COMPARTIDO	Number
PARROQUIA	String	H04_DUCHA_NO_TIENE	Number
ID_SECTOR	String	H05_GAS	Number
AREA	String	H05_LEÑA_CARBON	Number
TOTPER	Number	H05_ELECTTRICIDAD	Number
ID_SECTOR_1	String	H05_OTRO	Number
ID_DOMINIO	String	H06_AGUA_COMO_LLEGA	Number

CONTINÚA=>

NOMBRE	TIPO
VTV_CASA/VILLA	Number
VTV_DEPARTAMENTO	Number
VTV_CUARTO	Number
VTV_MEDIAGUA	Number
VTV_RANCHO	Number
VTV_OTRO	Number
VAP_ADOQ_PAVIM_CONCR	Number
VAP_EMPEDRADA	Number
VAP_LASTRADA_TIERRA	Number
VAP_CAMINO_SENDERO_CHAQUINAN	Number
VAP_OTRO	Number
V01_HORMIGON	Number
V01_ASBESTO	Number
V01_ZINC	Number
V01_TEJA	Number
V01_OTRO	Number
V03_HORMIGON	Number
V03_LADRILLO_BLOQUE	Number
V03_ADOBE_TAPIA	Number
V03_MADERA	Number
V03_CAÑA_NO_REVESTIDA	Number
V03_OTRO	Number
V05_DUELA_PARQUET_TABLON	Number
V05_TABLA_SIN_TRATAR	Number
V05_CERAMICA_BALDOSA_VINIL	Number
V05_TIERRA	Number
V05_OTRO	Number
V07_RED_PUBLICA	Number
V07_POZO	Number
V07_RIO_VERTIENTE_ACEQUIA	Number
V07_CARRO_REPARTIDOR	Number
V07_OTRO	Number
V08_TUBERIA_DENTRO_VIVIENDA	Number
V08_TUBERIA_DENTRO_EDIFICIO	Number
V08_TUBERIA_FUERA_EDIFICIO_LOTE	Number
V08_OTROS_MEDIOS	Number
V09_ALCANTARILLADO_PUBLICO	Number
V09_POZO_SEPTICO	Number
V09_POZO_CIEGO	Number
V09_NO_TIENE	Number
V10_RED_PUBLICA	Number
V10_NO_TIENE	Number
V10_OTRO	Number
V13_CARRO_RECOLECTOR	Number
V13_QUEMAN	Number
V13_ARROJA_TERRENO_QUEBR	Number

NOMBRE	TIPO
H06_AGUA_HERVIDA	Number
H06_AGUA_COMPRADA	Number
H06_OTRO	Number
H07_TELEFONO_CONVENCIONAL	Number
H08_CELULAR	Number
H09_INTERNET	Number
H10_COMPUTADORA	Number
H11_CABLE	Number
H15_PROPIA_PAGADA	Number
H15_PROPIA_PAGANDO	Number
H15_PROPIA_REGALADA	Number
H15_PRESTADA	Number
H15_ARRENDADA	Number
H15_OTRAS	Number
P05_CEDULA	Number
P07_SEGURO	Number
P19_ANALFABETISMO	Number
P20T_USO_CELULAR	Number
P20I_USO_INTERNET	Number
P20C_USO_COMPUTADORA	Number
P21_ASISTE_ESTABLECIMIENTO	Number
P22_FISCAL	Number
P22_PARTICULAR	Number
P22_FISCOMISIONAL	Number
P22_MUNICIPAL	Number
TIPOACT_AL_MENOS_UNA_HORA	Number
TIPOACT_NO_TRABAJO_PERO_SI	Number
TIPOACT_SERVICIOS_PRODUCTOS	Number
TIPOACT_NEGOCIO_FAMILIAR	Number
TIPOACT_LABORES_AGRICOLAS	Number
TIPOACT_CESANTE	Number
TIPOACT_ESTUDIANTE	Number
TIPOACT_QUEHACERES_DOMESTICOS	Number
TIPOACT_AL_MENOS_UNA_HORA_1	Number
TIPOACT_AL_MENOS_UNA_HORA_2	Number
P35_IESS	Number
P35_JUBILADO	Number
P35_NO_APORTA	Number
P35_SE_IGNORA	Number
P35_OTRO	Number
GRUOCU_DIRECTORES_GERENTES	Number
GRUOCU_PROFES_CIENTIFICOS	Number
GRUOCU_TECNICOS	Number
GRUOCU_PERSONAL_DE_APOYO_ADMIN	Number
GRUOCU_SERVICIOS_VENDEDORES	Number
GRUOCU_AGRICULTORES	Number

CONTINÚA=>

NOMBRE	TIPO
V13_OTRO	Number
V_HACINAMIENTO	Number
V12_FOCOS	Number
V12_FOCOS_POR_PERSONA	Number
H02_COCINA	Number
H03_SSHH_EXCLUSIVO	Number
H03_SSHH_COMPARTIDO	Number
H03_SSHH_NO_TIENE	Number

NOMBRE	TIPO
GRUOCU_OFIC_OPERARIOS_ARTESANOS	Number
GRUOCU_OPERARIOS_DE_INST_Y_MAQ	Number
GRUOCU_OCUPACIONES_ELEMENTALES	Number
GRUOCU_NO_DECLARADO	Number
GRUOCU_MILITARES	Number
GRUOCU_TRABAJADORES_NUEVOS	Number
GRAESC_GRADO_ESCOLARIDAD	Number

5.2. MODELAMIENTO

5.2.1. Selección de técnicas de modelado

De acuerdo a la metodología CRISP-DM, es importante considerar tres puntos para una selección adecuada de la técnica de modelado:

- Los tipos de datos disponibles para la minería
- Los objetivos de minería de datos
- Requisitos específicos de modelado. (IBM, 2012)

Para el presente trabajo, los datos disponibles son de tipo numérico continuo, el objetivo de minería de datos es consolidar estratos que respondan a los niveles socioeconómicos representados por las distintas variables del *Dataset* y finalmente el requisito específico es obtener los niveles Alto, Medio y Bajo de la población.

En consecuencia, las técnicas de modelado que responden a estas características son técnicas de clusterización, específicamente algoritmos divisivos, dentro de estos algoritmos se encuentra el denominado K-medias.

> El método de las k-medias [24, 25], es hasta ahora el más utilizado en aplicaciones científicas e industriales. El nombre le viene porque representa cada uno de los *clusters* por la media (o media ponderada) de sus puntos, es decir, por su centroide. Este método únicamente se puede aplicar a atributos numéricos, y los *outliers* le pueden afectar muy negativamente. Sin embargo, la representación mediante centroides tiene la ventaja de que tiene un significado gráfico y estadístico inmediato. (Garre, Cuadrado, & Sicilia, 2005, pág. 5)

5.2.2. Generación de los modelos

Configuración de parámetros

Para la aplicación del algoritmo de K-medias, se utilizó el programa para minería de datos "*RapidMiner Studio Comunity* 7.0", para aplicar K-medias en el programa, es necesario definir tres parámetros. El primero, establecer el número de clústeres a formar. Para este caso el número ya está predeterminado, se necesita tres clústeres.

El segundo parámetro es especificar la forma en la que se medirán las distancias inter e intra clústeres, para esto existen diferentes posibles métricas de distancias:

- Euclidean Distance
- Chebychev Distance
- Correlation Similarity
- Cosine Similarity
- Dice Similarity
- Jaccard Similarity
- Kernel Euclidean Distance
- Manhattan Distance
- Overlap Similarity.

Para definir cuál de las medidas de distancia será usada, se experimentó con cada una de ellas, obteniendo las métricas para evaluar la efectividad para construir los clústeres. Entre los valores que se evalúan se encuentra el promedio dentro de la distancia al centroide tanto dentro de cada uno de los clústeres como entre ellos, y el índice de Davies Bouldin.

	Avg. within centroid distance	Avg. within centroid distance_cluster_0	Avg. within centroid distance_cluster_1	Avg. within centroid distance_cluster_2	Davies Bouldin
EuclideanDistance	-2.004	-1.21	-1.8	-3.196	-1.826
ChebychevDistance	-2.104	-1.235	-2.06	-3.194	-2.002
Correlation Similarity	-2.007	-1.272	-1.914	-3.22	-1.884
Cosine Similarity	-2.007	-1.269	-1.911	-3.22	-1.883
Dice Similarity	-2.174	-1.918	-2.31	-2.983	-1.862
JaccardSimilarity	-2.174	-1.918	-2.31	-3.028	-1.862
KernelEucledianDistance	-2.006	-1.246	-1.905	-3.22	-1.861
ManhattanDistance	-2.028	-1.329	-2.026	-3.231	-1.95
OverlapSimilarity	-2.112	-1.751	-2.259	-2.986	-1.933

Figura 13. Comparación de la medida de distancias

El índice Davies Bouldin está definido de la siguiente manera:

$$DB = \frac{1}{K} \sum_{i=1, i \neq j}^{k} max \left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \right)$$

> Donde k es el número de clústeres, σi es la distancia promedio entre cada punto en el clúster i y el centroide del clúster, σj es la distancia promedio entre cada punto del clúster j y el centroide del clúster, y $d\ (ci, cj)$ es la distancia entre los centroides de los 2 clústeres.
> Valores pequeños para el índice DB indica clústeres compactos, y cuyos centros están bien separados los unos de los otros. Consecuentemente el número de clústeres que minimiza el índice DB se toma como el óptimo. (León Guzmán, 2016, pág. 13)

El índice Davies Bouldin indica, en este caso que, la mejor métrica es “Euclidean Distance”, pues contiene el índice más bajo, la distancia más alta entre clústeres y la menor distancia dentro de los clústeres, por lo tanto, es escogida para el desarrollo de la metodología.

El tercer parámetro es establecer el número de veces que se va a ejecutar el modelo, para lo cual se utilizó el parámetro por defecto, es decir, 10 ejecuciones, a continuación, se realizó para 100 ejecuciones y se obtuvo el mismo resultado, por lo tanto, no es necesario cambiar el parámetro por defecto.

Ejecución del modelo

Para la implementación del modelo, en primer lugar, es necesario leer el *Dataset* construido especificando la *meta data* que contiene el archivo.

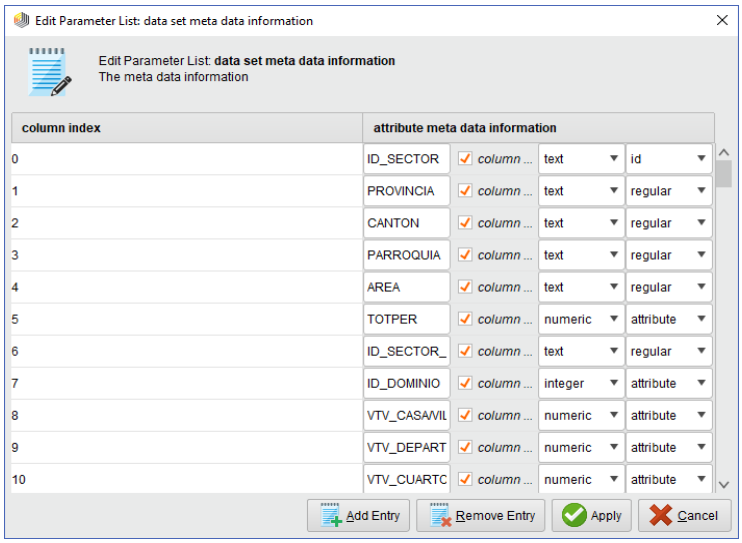


Figura 14. Meta data de lectura del Dataset

Posteriormente, es necesario tomar una decisión en cuanto a los valores faltantes, encontrados en varios atributos. De acuerdo al análisis realizado el mejor tratamiento para los faltantes es reemplazarlos por el valor de cero, ya que, significa la ausencia de una característica en el sector censal.

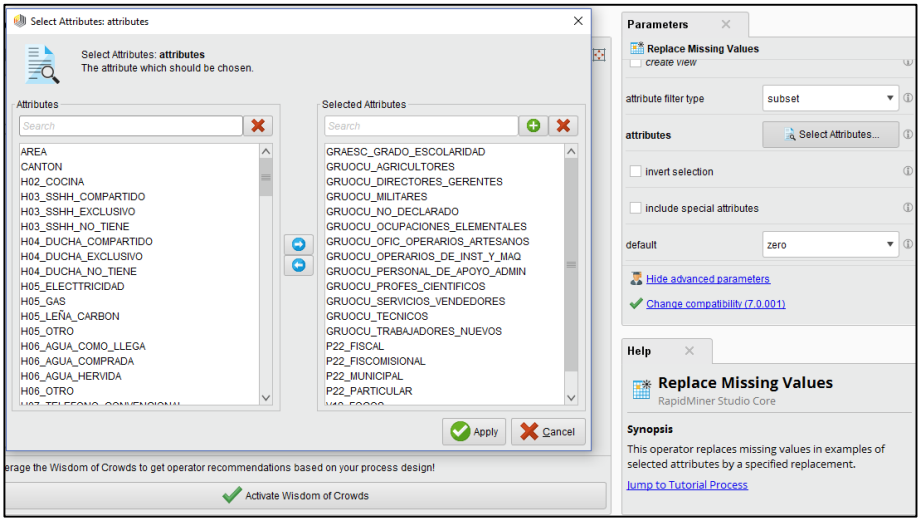


Figura 15. Tratamiento de valores faltantes

Posteriormente se selecciona los atributos que intervienen en el modelo excluyendo variables de identificación como Provincia, Cantón, Parroquia, entre otros.

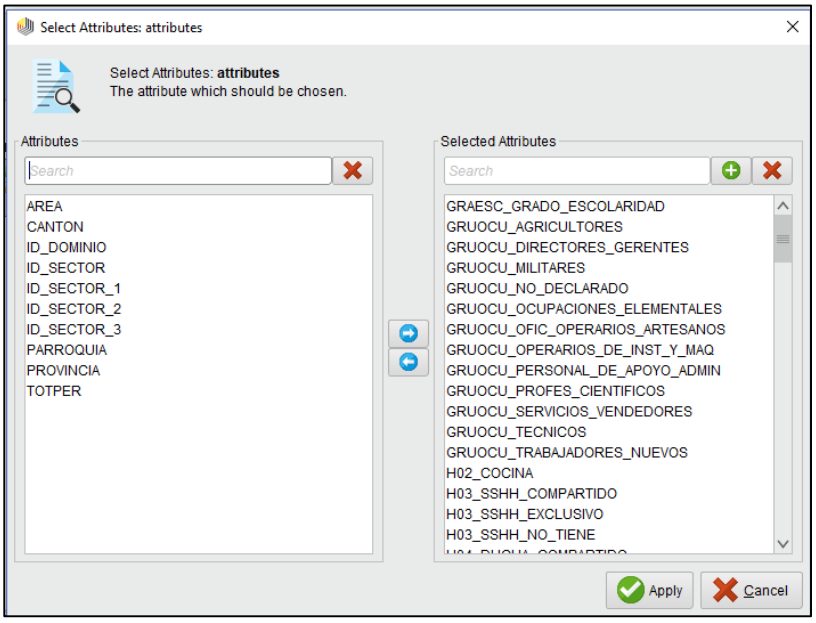


Figura 16. Selección de los atributos para el modelo

A continuación, con el fin de mantener una misma medida en las variables, para no otorgar un mayor peso a uno u otro atributo es necesarios normalizar porque no se encuentran en un rango de 0 a 1.

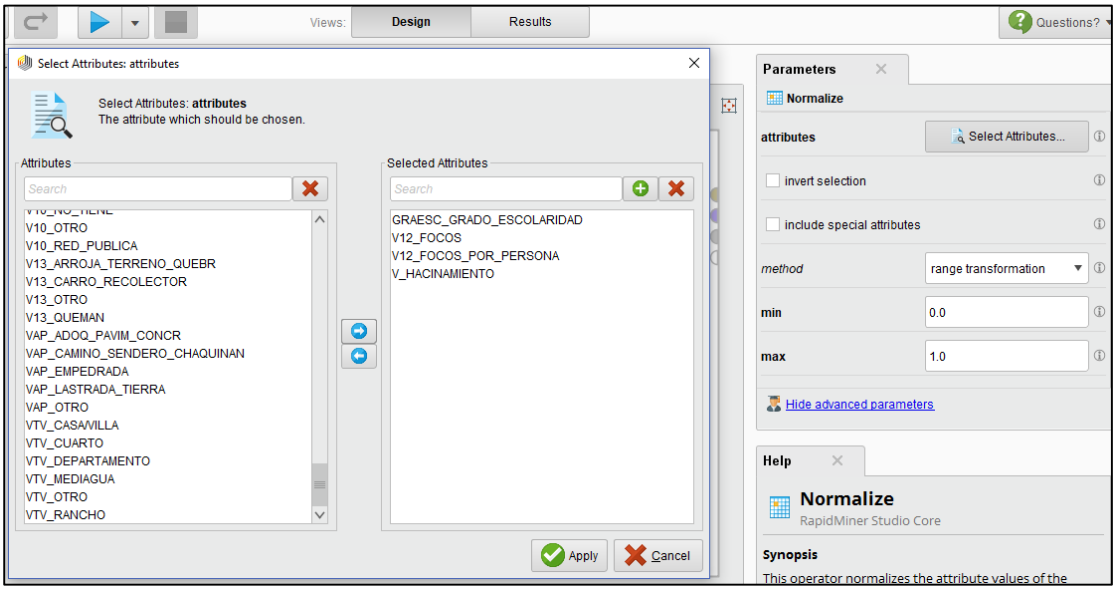


Figura 17. Normalización de las variables

Después de la preparación es posible ejecutar el modelo con los parámetros analizados en la sección anterior.

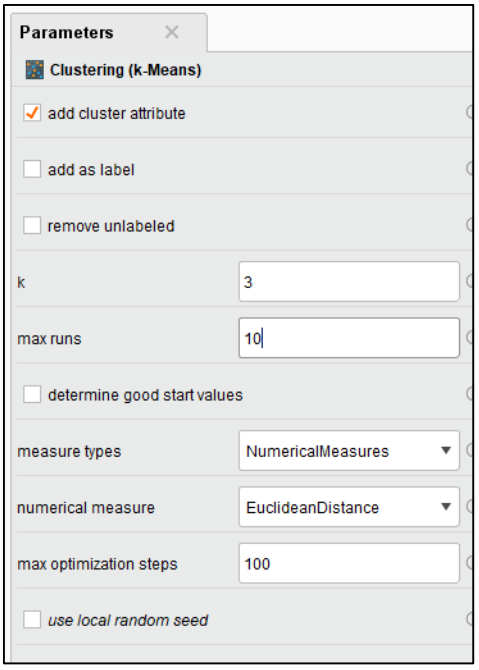


Figura 18. Parámetros para Clusterización con K-medias

Finalmente, se utiliza el operador *Cluster Distance Performance* para la evaluación del modelo y se guarda en formato *CSV* el *Dataset* con la información de a que clúster pertenece cada sector censal obteniendo en conjunto el siguiente proceso.

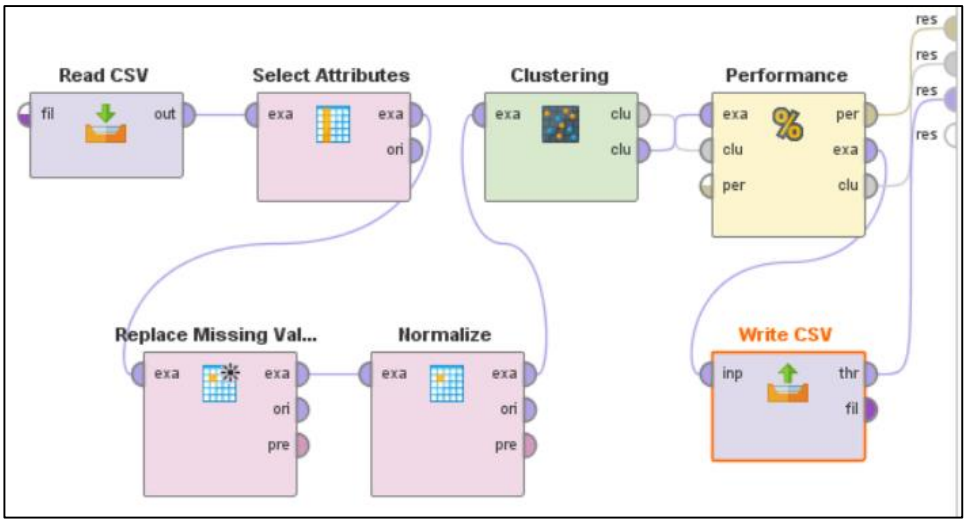


Figura 19. Proceso de implementación del modelo

Evaluación del modelo

Una vez ejecutado el modelo se obtiene los tres clústeres planteados y con lo cual se procede al análisis mediante los centroides de los atributos.

Se toma atributos en los que claramente se puede identificar los estratos socioeconómicos alto, medio o bajo, por ejemplo, hacinamiento, este atributo dice que mientras mayor sea el número de personas por dormitorio, menor es la clase social a la que pertenece, por lo tanto se puede notar que con la ejecución del modelo se cumple, el número mayor se encuentra en el clúster 1 que vendría a ser el estrato bajo, después le sigue el clúster 0 que viene a ser el estrato medio y por último el clúster 2 que es el estrato alto. Asimismo, el grado de escolaridad es menor en el estrato bajo identificado en el clúster 1, continua el clúster 0 con el estrato medio y el clúster 2 con el mayor grado de escolaridad que viene a ser el estrato alto

Tabla 12

Centroides de las variables de estratificación Nacional

ATRIBUTO	CLUSTER 0	CLUSTER 1	CLUSTER 2
V_HACINAMIENTO	0.34914009	0.37246155	0.28510411
V12_FOCOS_POR_PERSONA	0.1161547	0.0891603	0.19844881
GRAESC_GRADO_ESCOLARIDAD	0.41362213	0.33997412	0.5532412
VTV_CASA.VILLA	0.83444768	0.7122306	0.65427085
VTV_DEPARTAMENTO	0.02972711	0.00547089	0.2370026
VTV_CUARTO	0.02695459	0.0042301	0.08334589
VTV_MEDIAGUA	0.06107335	0.06391038	0.02144187
VTV_RANCHO	0.03758696	0.15958585	0.00212173
VTV_OTRO	0.01059526	0.05452166	0.00222648
VAP_ADOQ_PAVIM_CONCR	0.25666788	0.07696761	0.81192318
VAP_EMPEDRADA	0.2672071	0.1453864	0.08095303
VAP_LASTRADA_TIERRA	0.3515922	0.35635716	0.0854327
VAP_CAMINO_SENDERO_CHAQUINAN	0.12181415	0.40331152	0.01894495
VAP_OTRO	0.00281473	0.01804688	0.00297207
V01_HORMIGON	0.17152962	0.04110163	0.53136862
V01_ASBESTO	0.17662567	0.11786383	0.17806997
V01_ZINC	0.50688837	0.64552224	0.20685266
V01_TEJA	0.13894047	0.13093929	0.08222237
V01_OTRO	0.00637826	0.06465253	0.00174517
V03_HORMIGON	0.05165624	0.01972004	0.16297344
V03_LADRILLO_BLOQUE	0.73209213	0.40623871	0.77480624
V03_ADOBE_TAPIA	0.07824407	0.12451338	0.0409633

CONTINÚA=>

ATRIBUTO	CLUSTER 0	CLUSTER 1	CLUSTER 2
V03_MADERA	0.05913986	0.21833513	0.01334725
V03_CAÑA_NO_REVESTIDA	0.03908165	0.14221983	0.00221347
V03_OTRO	0.04001528	0.08891576	0.00623237
V05_DUELA_PARQUET_TABLON	0.04475986	0.01510726	0.20855676
V05_TABLA_SIN_TRATAR	0.16581793	0.41640769	0.07546488
V05_CERAMICA_BALDOSA_VINIL	0.16163368	0.03147685	0.44273381
V05_TIERRA	0.09625018	0.22186532	0.00944475
V05_OTRO	0.01387207	0.04224302	0.00747501
V07_RED_PUBLICA	0.73486465	0.15646898	0.9673148
V07_POZO	0.10014627	0.31590408	0.01288375
V07_RIO_VERTIENTE_ACEQUIA	0.10419517	0.389479	0.00953238
V07_CARRO_REPARTIDOR	0.03695823	0.09792512	0.00638368
V07_OTRO	0.02428977	0.04025926	0.0042106
V08_TUBERIA_DENTRO_VIVIENDA	0.53585068	0.11723267	0.87315008
V08_TUBERIA_DENTRO_EDIFICIO	0.31491777	0.28155305	0.10645488
V08_TUBERIA_FUERA_EDIFICIO_LOTE	0.0431713	0.09398907	0.00880734
V08_OTROS_MEDIOS	0.10623999	0.50726166	0.01195331
V09_ALCANTARILLADO_PUBLICO	0.32733008	0.01780253	0.93416336
V09_POZO_SEPTICO	0.4219364	0.29382341	0.04381419
V09_POZO_CIEGO	0.12367487	0.26651785	0.00597905
V09_NO_TIENE	0.0735155	0.30385654	0.00481514
V10_RED_PUBLICA	0.94391864	0.77834921	0.99308229
V10_NO_TIENE	0.04350822	0.18613352	0.00502807
V10_OTRO	0.01271576	0.03552472	0.00200671
H02_COCINA	0.80092417	0.74929181	0.89480693
H03_SSHH_EXCLUSIVO	0.80147577	0.64696679	0.89562714
H03_SSHH_COMPARTIDO	0.13595838	0.06723681	0.10119677
H03_SSHH_NO_TIENE	0.06264809	0.28579226	0.00329043
H04_DUCHA_EXCLUSIVO	0.53681051	0.19594384	0.84013282
H04_DUCHA_COMPARTIDO	0.07231044	0.01502775	0.08186704
H04_DUCHA_NO_TIENE	0.39082302	0.78902758	0.07808093
H05_GAS	0.92362174	0.67074298	0.97179515
H05_LEÑA_CARBON	0.05748654	0.3119556	0.00347528
H05_ELECTTRICIDAD	0.00149469	0.00114387	0.00813159
H06_AGUA_COMO_LLEGA	0.3944615	0.53579806	0.27105231
H06_AGUA_HERVIDA	0.37428467	0.3195171	0.43198754
H06_AGUA_COMPRADA	0.19632004	0.07737596	0.26685883
H06_OTRO	0.03491268	0.06722935	0.03023073
H07_TELEFONO_CONVENCIONAL	0.21738684	0.03835915	0.56550664
H08_CELULAR	0.71735774	0.55842873	0.86844311
H09_INTERNET	0.04770121	0.0117858	0.24426195
H10_COMPUTADORA	0.14366395	0.02831773	0.46427769
H11_CABLE	0.106566	0.02780005	0.28799055
H15_PROPIA_PAGADA	0.5323279	0.59974654	0.36328084
H15_PROPIA_PAGANDO	0.0566264	0.03552887	0.08196289

CONTINÚA=>

ATRIBUTO	CLUSTER 0	CLUSTER 1	CLUSTER 2
H15_PROPIA_REGALADA	0.10872071	0.146411	0.08124675
H15_PRESTADA	0.15344491	0.15849168	0.09863618
H15_ARRENDADA	0.12389608	0.02224882	0.36287622
H15_OTRAS	0.02501965	0.03756316	0.0122792
P05_CEDULA	0.78516591	0.75659322	0.84298781
P07_SEGURO	0.05526998	0.0298857	0.15774065
P19_ANALFABETISMO	0.08980934	0.16430382	0.03910995
P20T_USO_CELULAR	0.49111338	0.35818521	0.68465562
P20I_USO_INTERNET	0.15593946	0.05265054	0.43478776
P20C_USO_COMPUTADORA	0.2243247	0.09631657	0.50928523
P21_ASISTE_ESTABLECIMIENTO	0.36698224	0.35871035	0.36733808
P22_FISCAL	0.78773104	0.85918827	0.52894975
P22_PARTICULAR	0.16848858	0.08615754	0.41668561
P22_FISCOMISIONAL	0.03301921	0.04315829	0.04118855
P22_MUNICIPAL	0.0106884	0.00913692	0.01329796
TIPOACT_AL_MENOS_UNA_HORA	0.34668171	0.26894475	0.44112077
TIPOACT_NO_TRABAJO_PERO_SI	0.01308179	0.00870289	0.01445091
TIPOACT_SERVICIOS_PRODUCTOS	0.01291151	0.00757972	0.01435848
TIPOACT_NEGOCIO_FAMILIAR	0.01167807	0.0094028	0.01477133
TIPOACT_LABORES_AGRICOLAS	0.04181633	0.12139733	0.00601191
TIPOACT_CESANTE	0.0039463	0.00291228	0.00647131
TIPOACT_ESTUDIANTE	0.0188517	0.00973909	0.02180748
TIPOACT_QUEHACERES_DOMESTICOS	4.26E-04	2.02E-04	0.00119745
TIPOACT_AL_MENOS_UNA_HORA_1	0.34668171	0.26894475	0.44112077
TIPOACT_AL_MENOS_UNA_HORA_2	0.34668171	0.26894475	0.44112077
P35_IESS	0.15429559	0.13550899	0.25704368
P35_JUBILADO	0.00740212	0.00303239	0.02845406
P35_NO_APORTA	0.77483336	0.78820177	0.6617397
P35_SE_IGNORA	0.05840125	0.07172285	0.04021293
P35_OTRO	0.00461941	0.00143626	0.01231275
GRUOCU_DIRECTORES_GERENTES	0.00985446	0.00416301	0.04209982
GRUOCU_PROFES_CIENTIFICOS	0.03863193	0.01392197	0.12896686
GRUOCU_TECNICOS	0.01741377	0.0045871	0.06230385
GRUOCU_PERSONAL_DE_APOYO_ADMIN	0.03882986	0.01428063	0.09472066
GRUOCU_SERVICIOS_VENDEDORES	0.1532055	0.05627764	0.22269615
GRUOCU_AGRICULTORES	0.15133459	0.36052762	0.02317746
GRUOCU_OFIC_OPERARIOS_ARTESANOS	0.153124	0.06301416	0.12664042
GRUOCU_OPERARIOS_DE_INST_Y_MAQ	0.07584194	0.02865651	0.0753745
GRUOCU_OCUPACIONES_ELEMENTALES	0.23793771	0.3341862	0.11431946
GRUOCU_NO_DECLARADO	0.07885679	0.09448439	0.06141312
GRUOCU_MILITARES	0.0020863	9.13E-04	0.00551417
GRUOCU_TRABAJADORES_NUEVOS	0.04321423	0.02539303	0.04272559

Se obtuvo los tres clústeres deseados, sin embargo, para conocer el nivel socioeconómico al que pertenecen, ya sea alto, medio o bajo se debe recurrir a la caracterización del clúster dada por los centroides.

Este proceso fue realizado a nivel Nacional y se repite para cada uno de los dominios preestablecidos como las distintas Provincias del país tanto para el área urbana como rural, para esto es necesario cambiar la fuente de datos por cada uno de los archivos por dominio.

Para dividir el *Dataset* original en las 58 bases de datos por dominio se utilizó el software estadístico R, para el cuál el código de lectura, división y guardado de los *Dataset* es el siguiente:

```
rm(list = ls())
library(dplyr)
library(stringr)


#Lectura del Data set Nacional
censo<-read.table("C:/Users/Salome/Documents/TESIS/ETL/tabla_censo3.txt",
                  header = T,
                  sep = ";",
                  dec = ".")

#Creación de vectores para ser recorridos por el for de acuerdo al i

dom<-unique(censo$ID_DOMINIO)

dom_name<-paste0("dom_",dom)

#i es el número de dominio que va a recorrer

#Filtro por dominio y guardado
for (i in 1:length(dom)){
  dom_i<-filter(censo,ID_DOMINIO==dom[i])
  write.table(dom_i,paste0("C:/Users/Salome/Documents/TESIS/ETL/dominio/",dom_name[i]), sep = ";",dec = ".",row.names = F)
  print(i)
}
```

Figura 20. Código en R para división del Dataset

Después de ejecutar el modelo para cada uno de los *Dataset*, utilizando el R consolidamos en un solo archivo que será el Marco de Muestreo.

```r
rm(list = ls())
library(dplyr)
library(stringr)

#Lectura del Data set Nacional
censo<-read.table("C:/Users/Salome/Documents/TESIS/ETL/tabla_censo3.txt",
                  header = T,
                  sep = ";",
                  dec = ".")

#Creación de vectores para ser recorridos por el for de acuerdo al i

dom<-unique(censo$ID_DOMINIO)

dom_name<-paste0("dom_",dom)
rm(censo)

for (i in 1:length(dom)){
  cluster_dom<-read.table(paste0("C:/Users/Salome/Documents/TESIS/CLUSTER/",dom_name[i],".csv"), sep = ";",dec = ".",header = T)
  print(i)
  ifelse(
    i==1,
    cluster_domi<-cluster_dom,
    cluster_domi<-rbind(cluster_domi,cluster_dom)
  )
}
class(cluster_domi$ID_SECTOR)
cluster_domi<-mutate(cluster_domi,ID_SECTOR=str_pad(as.character(ID_SECTOR),12,"left","0"))
write.table(cluster_domi,"C:/Users/Salome/Documents/TESIS/CLUSTER/dominios_consolidados.txt", sep = ";",dec = ".",row.names = F)
```

Figura 21. Código en R para consolidar los resultados de Rapid Miner

Para el análisis de centroides de cada dominio se realizó una tabla adicional en excel en la que se especifica a qué nivel socioeconómico pertenece cada clúster, ya que el orden de los estratos, alto, medio y bajo pueden cambiar por dominio por lo que es necesario analizar las características de cada uno.

	A	B	C	D
1	Dominio	Cluster_0	Cluster_1	Cluster_2
2	dom_0101	Alto	Bajo	Medio
3	dom_0102	Alto	Bajo	Medio
4	dom_0201	Medio	Bajo	Alto
5	dom_0202	Alto	Medio	Bajo
6	dom_0301	Medio	Bajo	Alto
7	dom_0302	Medio	Bajo	Alto
8	dom_0401	Bajo	Medio	Alto
9	dom_0402	Bajo	Medio	Alto
10	dom_0501	Medio	Alto	Bajo
11	dom_0502	Bajo	Medio	Alto
12	dom_0601	Bajo	Alto	Medio
13	dom_0602	Medio	Bajo	Alto
14	dom_0701	Medio	Alto	Bajo
15	dom_0702	Medio	Bajo	Alto
16	dom_0801	Alto	Bajo	Medio
17	dom_0802	Bajo	Alto	Medio
18	dom_0901	Medio	Alto	Bajo
19	dom_0902	Alto	Medio	Bajo
20	dom_1001	Alto	Bajo	Medio
21	dom_1002	Alto	Medio	Bajo

Figura 22. Tabla de correspondencia entre clústeres y estratos

Así, en el caso del dominio 0101, correspondiente a la provincia 01 Azuay y al área 01 Urbana, el Cluster_0 corresponde al estrato alto, el Cluster_1 al bajo y el Cluster_2 al medio. Es importante

mencionar que para los dominios de la provincia de Galápagos, se utiliza un estrato único, es decir, no se realiza la clasificación de acuerdo a requerimientos del INEC.

Esta tabla se la utiliza para asignar el estrato a cada sector de la tabla consolidada, para esto el código utilizado en R es:

```r
rm(list = ls())
library(dplyr)
library(stringr)
library(xlsx)
#Lectura del Data set Nacional
censo<-read.table("C:/Users/Salome/Documents/TESIS/ETL/tabla_censo3.txt",
                  header = T,
                  sep = ";",
                  dec = ".")


for (i in 1:7){
  censo[,i]<-as.character(censo[,i])
}

censo<-mutate(censo,ID_SECTOR=str_pad(ID_SECTOR,12,"left","0"),
              PROVINCIA=str_pad(PROVINCIA,2,"left","0"),
              CANTON=str_pad(CANTON,2,"left","0"),
              PARROQUIA=str_pad(PARROQUIA,2,"left","0"),
              ID_DOMINIO=str_pad(ID_DOMINIO,4,"left","0")
)


censo1<-select(censo,1:7,123,124)
rm(censo)
```

Figura 23. Código en R para crear identificadores únicos

```r

#Lectura de dataset clusterizado por dominios consolidado
clusteres<-read.table("C:/Users/Salome/Documents/TESIS//CLUSTER/dominios_consolidados.txt",
                  header = T,
                  sep = ";",
                  dec = ".")
clusteres<-select(clusteres,115,116)
clusteres$cluster<-as.character(clusteres$cluster)

clusteres<-mutate(clusteres,ID_SECTOR=str_pad(as.character(ID_SECTOR),12,"left","0"))

clusteres2<-merge(clusteres,censo1,
                  by.x="ID_SECTOR",
                  by.y="ID_SECTOR")

#LECTURA DE CLASIFICACION DE CLUSTERES
library(xlsx)
clasif_est<-read.xlsx("C:/Users/Salome/Documents/TESIS/CLUSTER/CLASIF_EST.xlsx",
                      1,
                      header = T)

clasif_est<-mutate(clasif_est,ID_DOMINIO=substr(Dominio,5,9),
                   cluster_0=as.character(Cluster_0),
                   cluster_1=as.character(Cluster_1),
                   cluster_2=as.character(Cluster_2))
summary(clasif_est)
clusteres3<-merge(clusteres2,clasif_est,
                  by="ID_DOMINIO")

clusteres3<-mutate(clusteres3,estrato_c=ifelse(cluster=="cluster_0",Cluster_0,ifelse(cluster=="cluster_1",Cluster_1,Cluster_2)))
write.table(clusteres3,"./clusteres.txt",sep = ";",dec = ".",row.names = F)

```

Figura 24. Código en R para asignar estratos por clúster

La tabla resultante es la que se muestra a continuación, la cual cumple la función del Marco de Muestreo ya que contiene variables de identificación de los sectores censales y el estrato al que corresponden. En este caso se obtiene además las variables “ESTRATO” y “Estrato_ant” hacen

referencia al estrato utilizado en la actualidad por el INEC y el “estrato_c” corresponde al estrato construido con metodología propuesta con el fin de realizar las respectivas comparaciones.

	ID_DOMINIO	ID_SECTOR	cluster	PROVINCIA	CANTON	PARROQUIA	AREA	TOTPER	ESTRATO	Estrato_ant	Dominio	estrato_c
1	0101	010550001008	cluster_0	01	05	50	1	516	Bajo	1	dom_0101	Alto
2	0101	010250001009	cluster_2	01	02	50	1	159	Alto	3	dom_0101	Medio
3	0101	010250001007	cluster_2	01	02	50	1	543	Alto	3	dom_0101	Medio
4	0101	010250001008	cluster_2	01	02	50	1	536	Alto	3	dom_0101	Medio
5	0101	011250001001	cluster_2	01	12	50	1	234	Alto	3	dom_0101	Medio
6	0101	010350003001	cluster_1	01	03	50	1	296	Bajo	1	dom_0101	Bajo
7	0101	010550001012	cluster_0	01	05	50	1	233	Bajo	1	dom_0101	Alto
8	0101	010550001010	cluster_0	01	05	50	1	302	Bajo	1	dom_0101	Alto
9	0101	010650001001	cluster_2	01	06	50	1	416	Alto	3	dom_0101	Medio
10	0101	010550001001	cluster_1	01	05	50	1	373	Bajo	1	dom_0101	Bajo
11	0101	010550001011	cluster_0	01	05	50	1	319	Bajo	1	dom_0101	Alto
12	0101	010350002007	cluster_0	01	03	50	1	395	Bajo	1	dom_0101	Alto
13	0101	010350002008	cluster_0	01	03	50	1	457	Bajo	1	dom_0101	Alto
14	0101	010350002009	cluster_0	01	03	50	1	236	Bajo	1	dom_0101	Alto
15	0101	010250001005	cluster_2	01	02	50	1	434	Alto	3	dom_0101	Medio
16	0101	010250001006	cluster_2	01	02	50	1	427	Alto	3	dom_0101	Medio
17	0101	010350002010	cluster_0	01	03	50	1	400	Bajo	1	dom_0101	Alto
18	0101	010350004003	cluster_1	01	03	50	1	329	Bajo	1	dom_0101	Bajo
19	0101	011250001002	cluster_2	01	12	50	1	251	Alto	3	dom_0101	Medio
20	0101	010550001009	cluster_0	01	05	50	1	235	Bajo	1	dom_0101	Alto

Figura 25. Marco de Muestreo a nivel de sector censal

CAPÍTULO VI

EVALUACIÓN Y DISTRIBUCIÓN

6.1. EVALUACIÓN

Después de consolidado el modelo a utilizar, “antes de continuar, debe evaluar los resultados de sus esfuerzos utilizando los criterios de rendimiento comercial establecidos en el inicio del proyecto. Es la clave para asegurar que su organización pueda utilizar los resultados que ha obtenido”. (IBM, 2012, pág. 42).

Es el momento de responder a la hipótesis planteada: “Las técnicas de segmentación/agrupación implementadas bajo la gestión de datos organizada permitirán mejorar el tiempo de construcción de los niveles socioeconómicos/estratos del Censo de Población y Vivienda sin afectar la calidad estadística de los clústeres obtenidos”. Para esto es necesario comparar la metodología actual utilizada por el INEC y la metodología propuesta en términos tanto cuantitativos como cualitativos.

6.1.1. Evaluación Cuantitativa

Para la evaluación del modelo propuesto, se comparó los estratos obtenidos para los 40.421 sectores censales contra los designados por el INEC, alcanzando un 66% de coincidencia, en principio se pensaría que es un porcentaje bajo la expectativa, por lo que se procedió a hacer análisis de consistencia de los clústeres verificando su calidad estadística utilizando análisis de varianzas ANOVA de un factor, con las variables más representativas a medir en las encuestas que utilizan el marco de muestreo llamadas Variables de Control.

Variables de Control

- Tasa de Participación Bruta (TPB)

> Mide el tamaño de la oferta laboral o fuerza de trabajo en relación con la población total. Es decir, la cantidad de personas de cierta edad --en este caso, de 15 años y más-- que están en capacidad y disponibilidad de ejercer actividades económicas productivas.
> Se trata de un indicador particularmente útil para analizar la incorporación al mercado laboral (SIISE, 2013)

Población Económicamente Activa (PEA) /Población Total (PT)

Donde,

Población Económicamente Activa (PEA): Personas de 15 años y más que trabajaron al menos 1 hora en la semana de referencia o, aunque no trabajaron, tuvieron trabajo (empleados); y personas que no tenían empleo, pero estaban disponibles para trabajar y buscan empleo (desempleados). (INEC, 2016, p. 7)
Población en edad de trabajar (PET): Comprende a todas las personas de 15 años y más. (INEC, 2016, p. 7)

- Tasa de Participación Global (TPG)

Refleja la oferta de fuerza de trabajo de una sociedad, es decir, mide la proporción de la población en edad de trabajar que desea participar activamente en el mercado laboral. Indica la proporción de personas económicamente activas con relación a la PET. Se trata de una medida más ajustada y clara que la tasa bruta de participación laboral ya que establece la relación entre el número de personas económicamente activas y la población en edad de trabajar. (SIISE, 2013)

Población Económicamente Activa (PEA) /Población en Edad de Trabajar (PET)

- Acceso a Servicios Básicos

La dimensión considera las condiciones sanitarias de la vivienda. El hogar es pobre si: i) la vivienda no tiene servicio higiénico o si lo tiene es por pozo ciego o letrina o, ii) si el agua que obtiene la vivienda no es por red pública o por otra fuente de tubería (INEC, 2015)

Análisis de Varianzas

Para el análisis y evaluación del modelo se mide la situación del país en cuanto al empleo utilizando el paquete estadístico SPSS.

El primer análisis se realizó un ANOVA para las variables Tasa de Participación Bruta y Tasa de Participación Global a nivel nacional entre la metodología actual aplicada por el INEC y la metodología nueva propuesta

Tabla 13
ANOVA Metodología Actual para las variables TPB y TPG

	Varianza	Suma de cuadrados	gl	Media cuadrática	F	Sig.
TPB	Entre grupos	2676,285	3	892,095	208941,104	0,000
	Dentro de grupos	61509,164	14406307	0,004		
	Total	64185,449	14406310			
TPG	Entre grupos	547,027	3	182,342	31231,250	0,000
	Dentro de grupos	84110,635	14406307	0,006		
	Total	84657,662	14406310			

Tabla 14
ANOVA Metodología Propuesta para las variables TPB y TPG

	Varianza	Suma de cuadrados	gl	Media cuadrática	F	Sig.
TPB	Entre grupos	4307,342	3	1435,781	345440,099	0,000
	Dentro de grupos	59878,106	14406307	0,004		
	Total	64185,449	14406310			
TPG	Entre grupos	887,148	3	295,716	50855,327	0,000
	Dentro de grupos	83770,514	14406307	0,006		
	Total	84657,662	14406310			

En la comparación entre grupos por suma de cuadrados, la cual representa la distancia entre las medias de los agrupamientos para las variables planteadas, se puede apreciar que con la metodología propuesta la distancia es 60.94% más amplia que la existente en la variable TPB y 62.17% en la variable TPG, y dentro de los grupos es menor, lo cual implica que la metodología propuesta proporciona una mejor estimación de los agrupamientos.

El siguiente análisis se realizó para saber si existe diferencia significativa entre las medias de los grupos:

Tabla 15
Contraste de Scheffe para la variable TPB de Metodología Actual

Scheffe[a,b]		TPB			
Estrato anterior	**N**	**Subconjunto para alfa = 0,05**			
		1	2	3	4
1	3658254	0,3744			
2	5827905		0,3928		
3	4896087			0,4083	
4	24065				0,4998
Sig.		1,000	1,000	1,000	1,000

Tabla 16

Contraste de Scheffe para la variable TPB de Metodología Propuesta

Scheffe[a,b]		TPB			
Estrato construido	**N**	**Subconjunto para alfa = 0,05**			
		1	2	3	4
1	3111171	0,3720			
2	6497133		0,3872		
3	4773942			0,4157	
4	24065				0,4998
Sig.		1,000	1,000	1,000	1,000

En ambos análisis los grupos tienen diferencias significativas en la Tasa de Participación Bruta a un nivel del 0.05.

Tabla 17

Contraste de Scheffe para la variable TPG de Metodología Actual

Scheffe[a,b]		TPG			
Estrato anterior	**N**	**Subconjunto para alfa = 0,05**			
		1	2	3	4
1	5827905	0,5677			
2	3658254		0,5729		
3	4896087			0,5744	
4	24065				0,7027
Sig.		1,000	1,000	1,000	1,000

Tabla 18

Contraste de Scheffe para la variable TPG de Metodología Propuesta

Scheffe[a,b]		TPG			
Estrato construido	**N**	**Subconjunto para alfa = 0,05**			
		1	2	3	4
1	6497133	0,5655			
2	3111171		0,5723		
3	4773942			0,5785	
4	24065				0,7027
Sig.		1,000	1,000	1,000	1,000

En ambos análisis los grupos tienen diferencias significativas en la Tasa de Participación Global a un nivel del 0.05.

A continuación, el análisis basado en gráficas de la metodología actual frente a la propuesta.

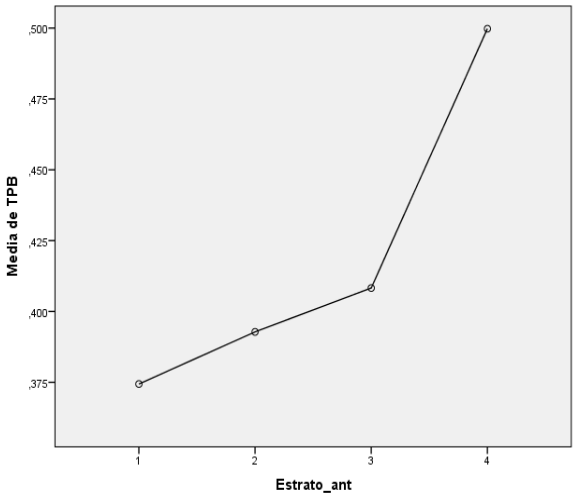


Figura 26. Gráfica de medias de la Tasa de Participación Bruta con estratos actuales

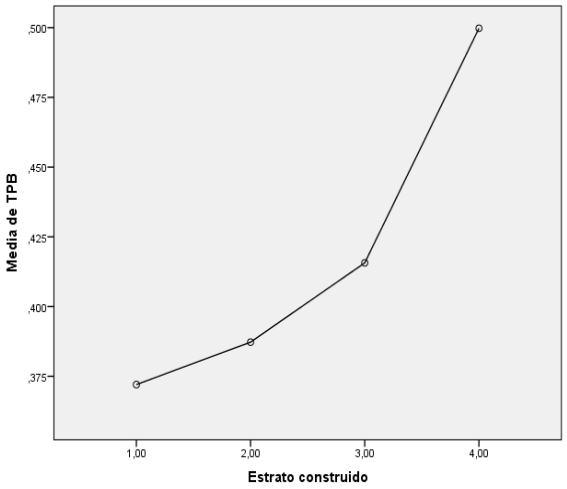


Figura 27. Gráfica de medias de la Tasa de Participación Bruta con estratos propuestos

La gráfica de medias indica que la Tasa de Participación Bruta en los estratos actuales frente a los propuestos existe una diferencia en las medias, en donde la media es más baja tanto para el estrato1 (estrato socioeconómico bajo) y el estrato 2 (estrato socioeconómico medio), y la media del estrato 3 (estrato socioeconómico alto) es más alta, y se evidencia una mejor diferenciación de los clústeres.

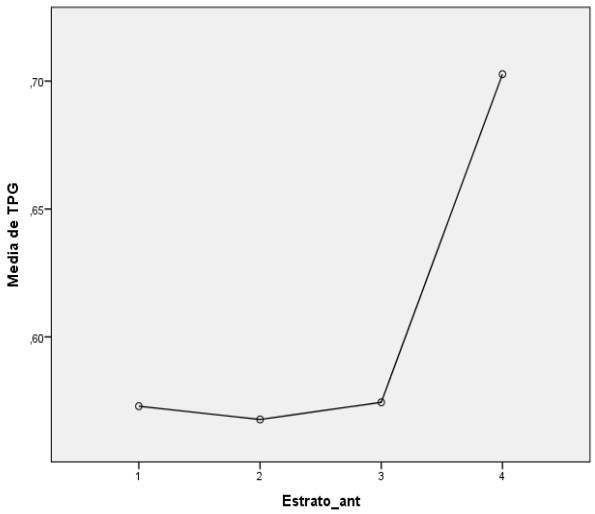


Figura 28. Gráfica de medias de la Tasa de Participación Global con estratos actuales

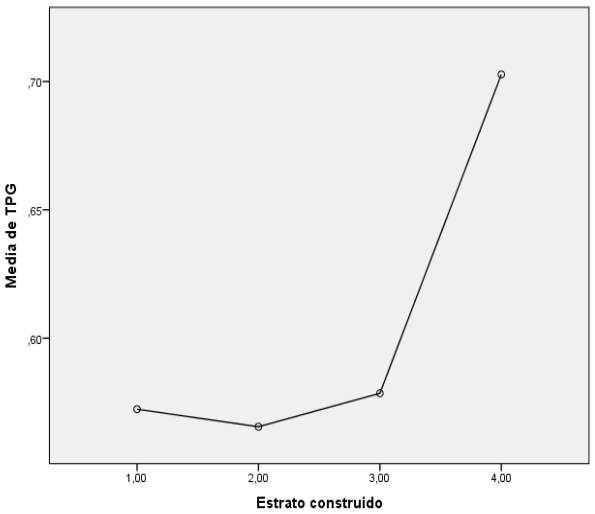


Figura 29. Gráfica de medias de la Tasa de Participación Global con estratos propuestos

La Tasa de Participación Global actual frente a la propuesta se tiene medias similares, lo cual indica que las mediciones del INEC frente a las propuestas son bastante cercanas y tienen el mismo comportamiento que la TPB.

La siguiente comparación se realizó en cuanto a la situación de servicios básicos de los distintos sectores censales con un indicador a nivel de hogares.

Se realizó el análisis ANOVA sobre Pobreza por Acceso a Servicios Básicos (PASB) a nivel nacional entre la metodología actual aplicada por el INEC y la metodología nueva propuesta.

Tabla 19
ANOVA Estratos actuales para la variable PASB

Varianza	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Entre grupos	65144,577	3	21714,859	222279,058	0,000
Dentro de grupos	364431,645	3730419	0,098		
Total	429576,223	3730422			

Tabla 20
ANOVA Estratos propuestos para la variable PASB

Varianza	Suma de cuadrados	Gl	Media cuadrática	F	Sig.
Entre grupos	121047.511	3	40349,170	487861,602	0,000
Dentro de grupos	308528.712	3730419	0,083		
Total	429576,223	3730422			

Se puede observar que la distancia entre las medias de los grupos para la metodología propuesta es 85.81% más amplia con respecto a la actual y dentro de los grupos es menor por lo tanto la metodología propuesta cuenta con mejores agrupaciones.

Seguido se revisó las diferencias significativas entre las medias de los grupos, obteniendo los siguientes resultados:

Tabla 21

Contraste de Scheffe para la variable PASB de Metodología Actual

Scheffe[a,b]		PASB_P			
Estrato anterior	**N**	**Subconjunto para alfa = 0,05**			
		1	2	3	4
4	7161	0,1126			
3	1318704		0,1808		
2	1493843			0,2422	
1	910715				0,5143
Sig.		1,000	1,000	1,000	1,000

Tabla 22

Contraste de Scheffe para la variable PASB de Metodología Propuesta

Scheffe[a,b]		PASB_P			
Estrato construido	**N**	**Subconjunto para alfa = 0,05**			
		1	2	3	4
4	7161	0,1126			
3	1296324		0,1209		
2	1652772			0,2629	
1	774166				0,6166
Sig.		1,000	1,000	1,000	1,000

En ambos análisis los grupos tienen diferencias significativas en Pobreza por Acceso a Servicios Básicos a un nivel de 0.05.

A continuación, el análisis basado en gráficas de la metodología actual frente a la propuesta

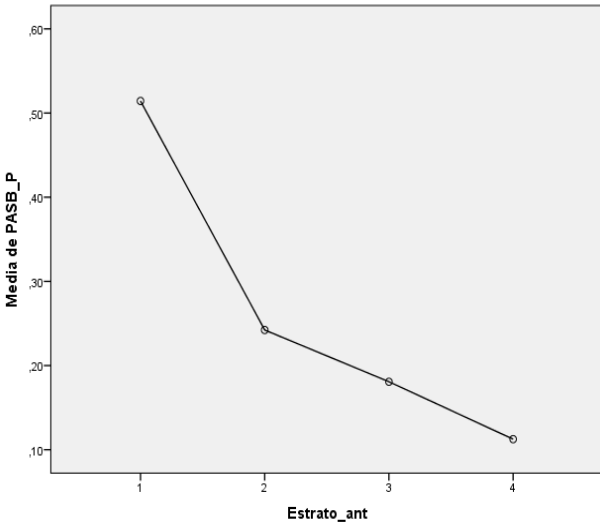


Figura 30. Gráfica de medias de la PASB con estratos actuales

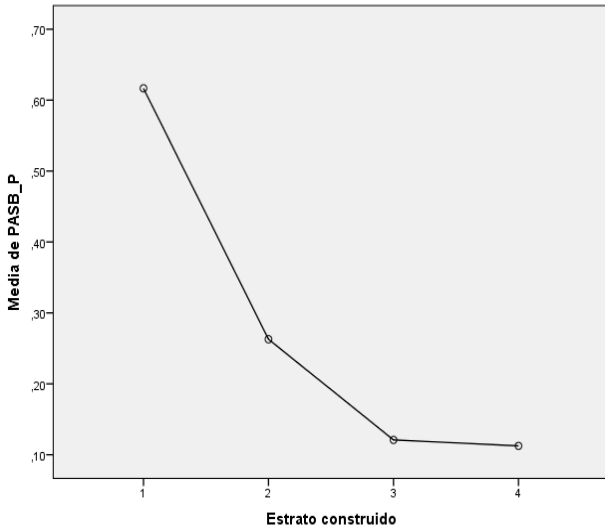


Figura 31. Gráfica de medias de la PASB con estratos propuestos

La gráfica indica que, para la variable Pobreza por Acceso a Servicios Básicos en los estratos actuales frente a los propuestos, en el estrato 1 (estrato socioeconómico bajo) la pobreza aumenta significativamente, cambiando la escala de los gráficos, mientras que en el estrato 3 (estrato socioeconómico alto) baja la PASB logrando mejor diferenciación entre estratos.

6.1.2. Evaluación Cualitativa

Otra de las formas para verificar la consistencia de los resultados es de manera gráfica, con la ayuda del sistema de información geográfica QGIS 2.18, en donde, se puede cargar los mapas publicados por el INEC en su página web, conocidos como "GeoDatabase Nacional", y unir a los resultados obtenidos a nivel de sector censal.

En primer lugar, analizamos el Ecuador en su totalidad, como se puede observar en el siguiente mapa, en su mayor extensión se puede apreciar el estrato bajo, debido a que los sectores censales de estas áreas son extensos por la dispersión que existe entre viviendas

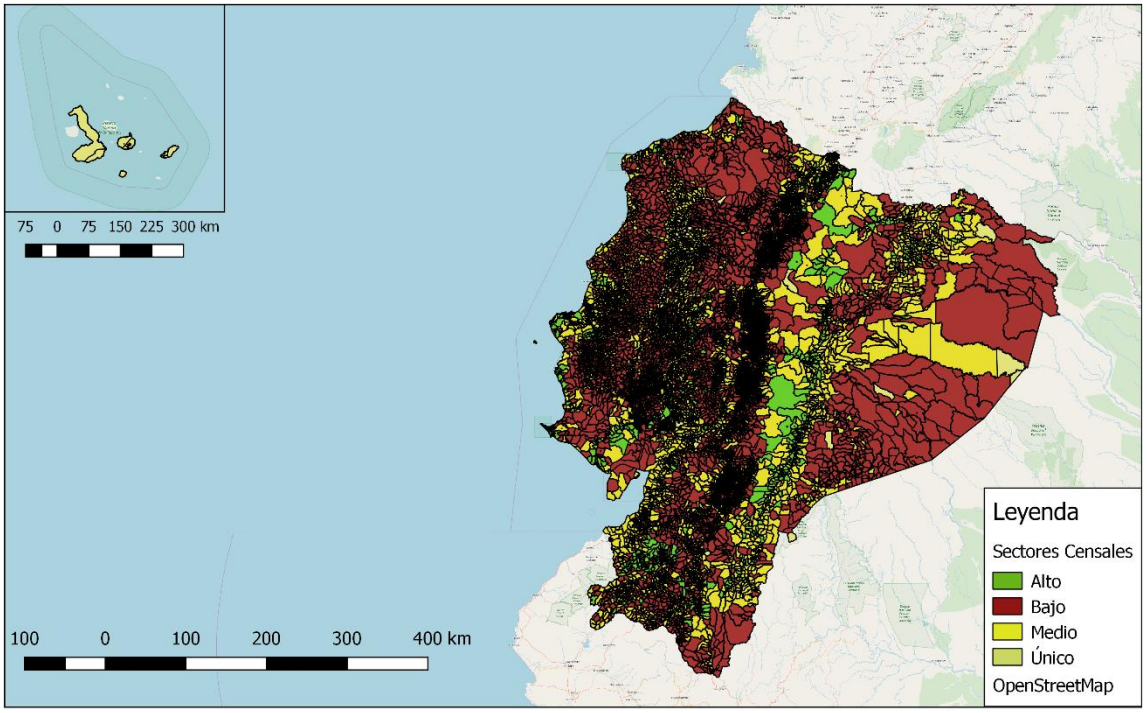


Figura 32. Sectores Censales Estratificados.- Ecuador

Para apreciar la distribución de los estratos se analiza áreas pequeñas como las cinco ciudades principales de acuerdo al INEC, Quito, Guayaquil, Cuenca, Ambato y Machala.

Como se puede observar, el estrato bajo comúnmente se puede encontrar en las periferias de los centros urbanos. Específicamente en Quito se distingue que el estrato alto se ubica en la zona centro norte y norte; y el bajo al sur, en Guayaquil se ubica claramente el estrato bajo en el sector del Fortín y el estrato alto en la parte central.

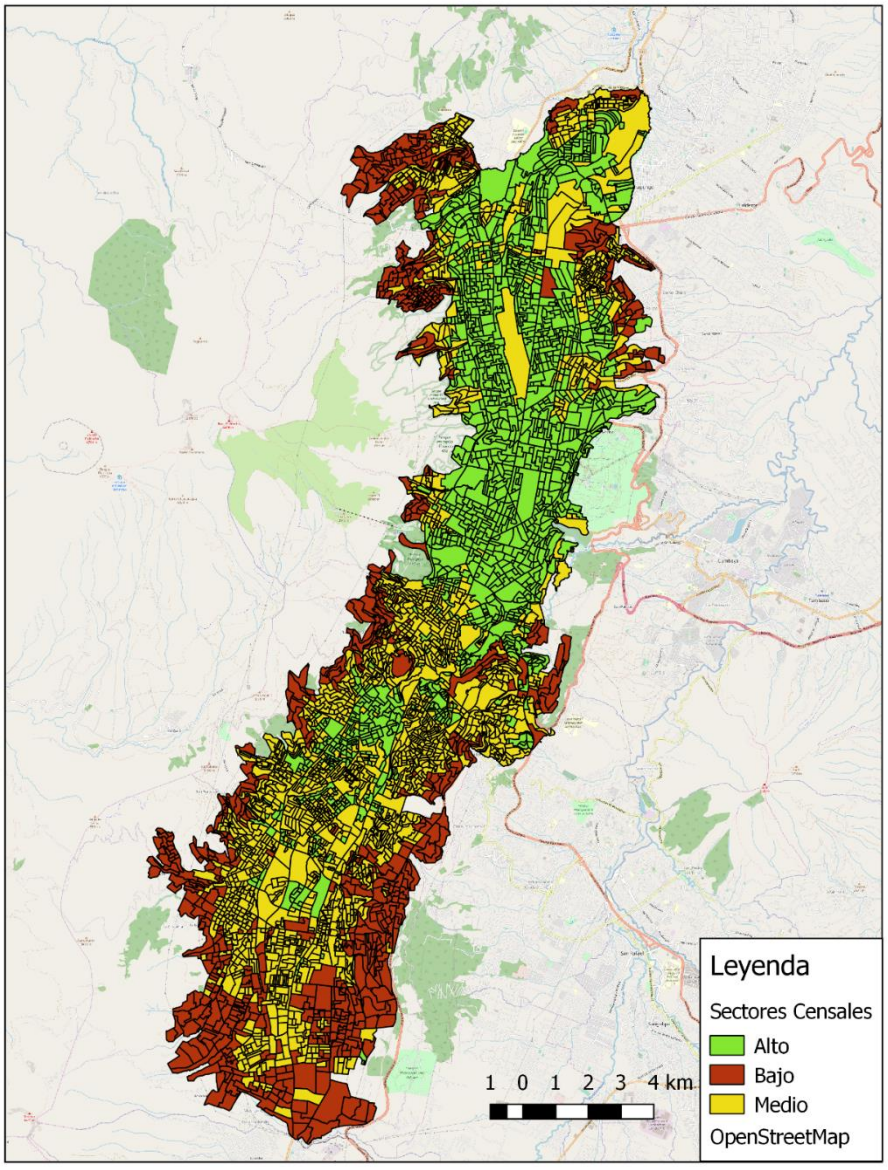


Figura 33. Estratificación Sectores Censales.- Quito

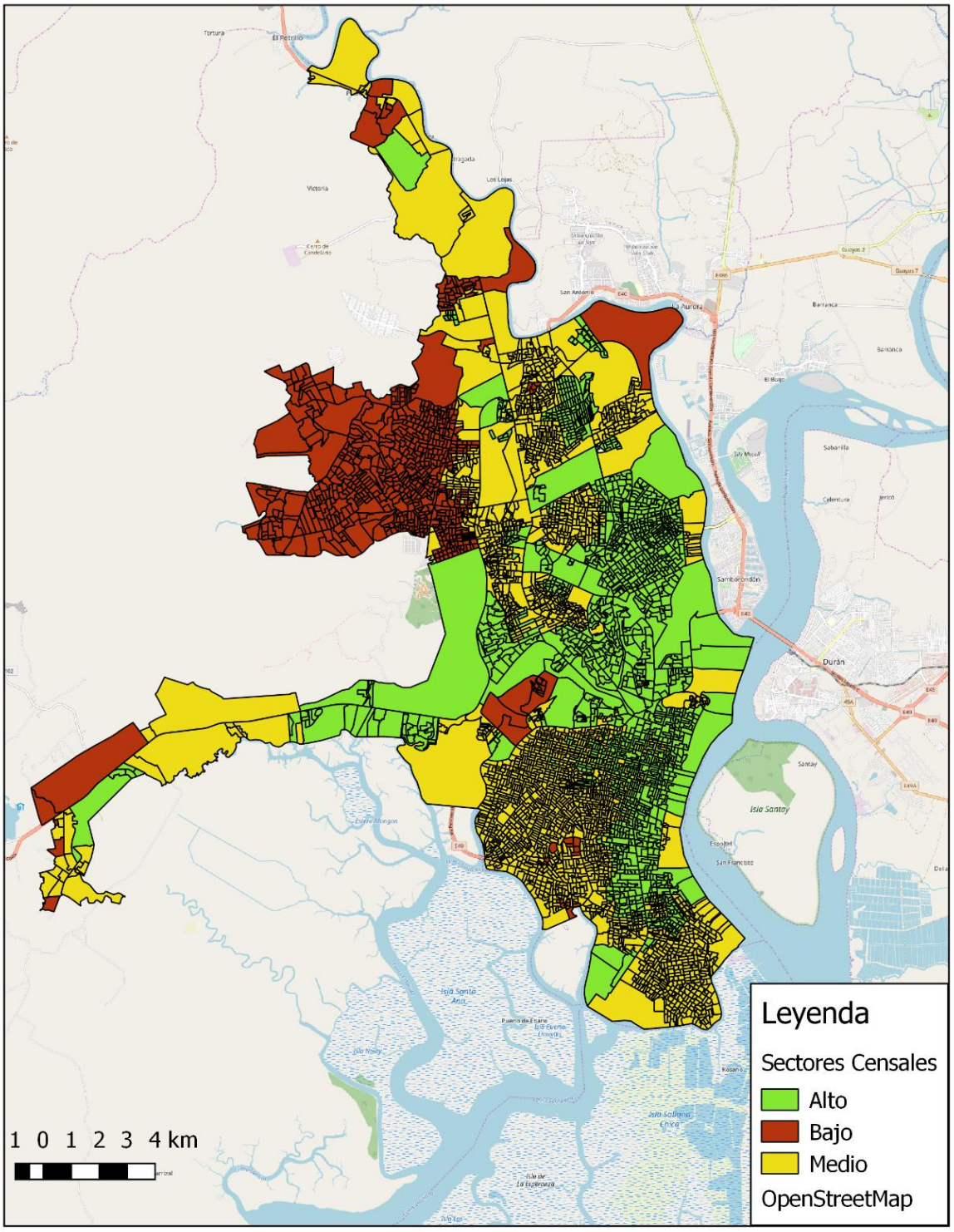


Figura 34. Estratificación Sectores Censales.- Guayaquil

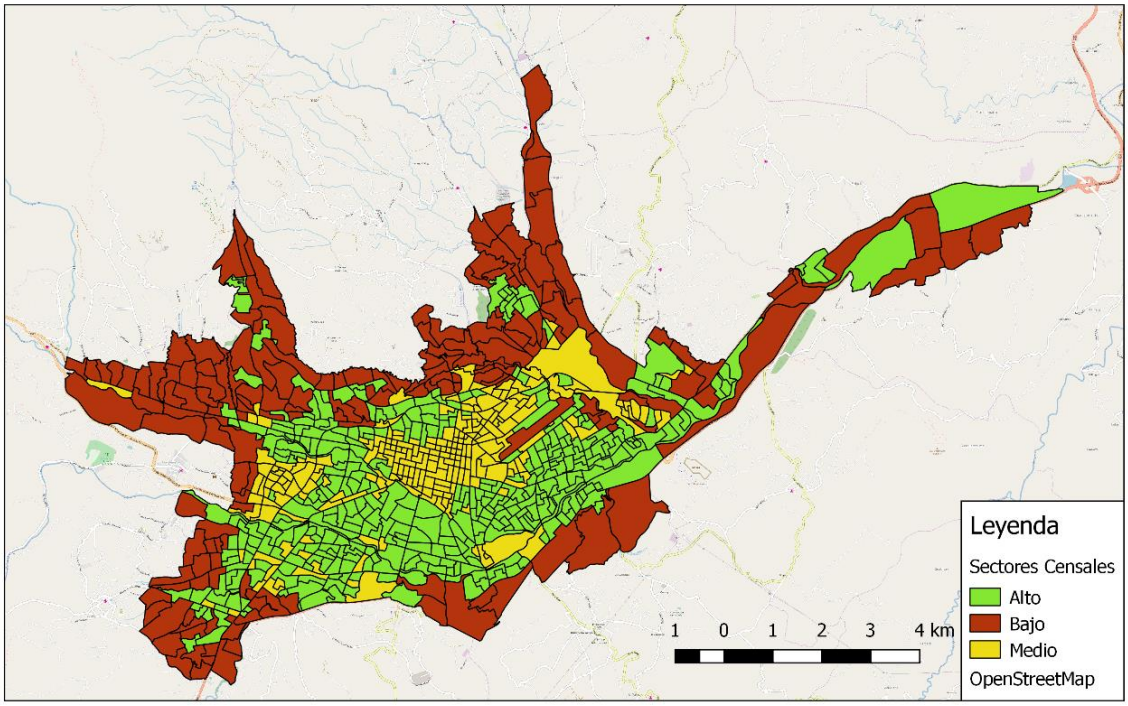


Figura 35. Estratificación Sectores Censales.- Cuenca

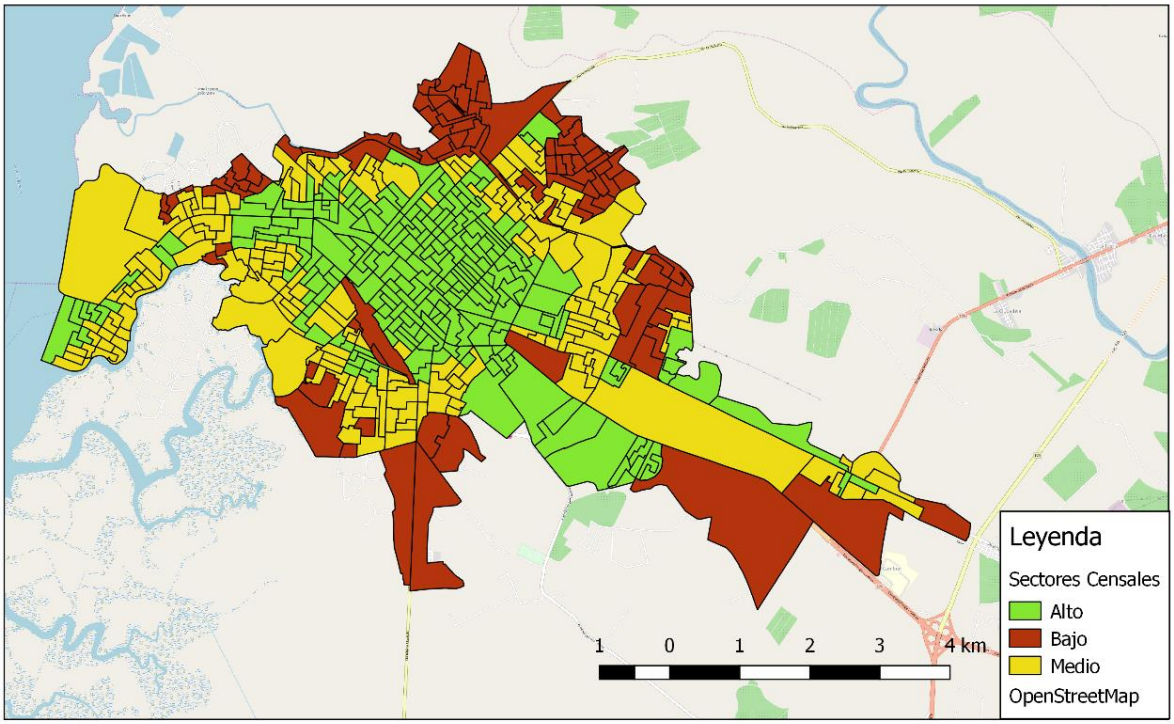


Figura 36. Estratificación Sectores Censales.- Machala

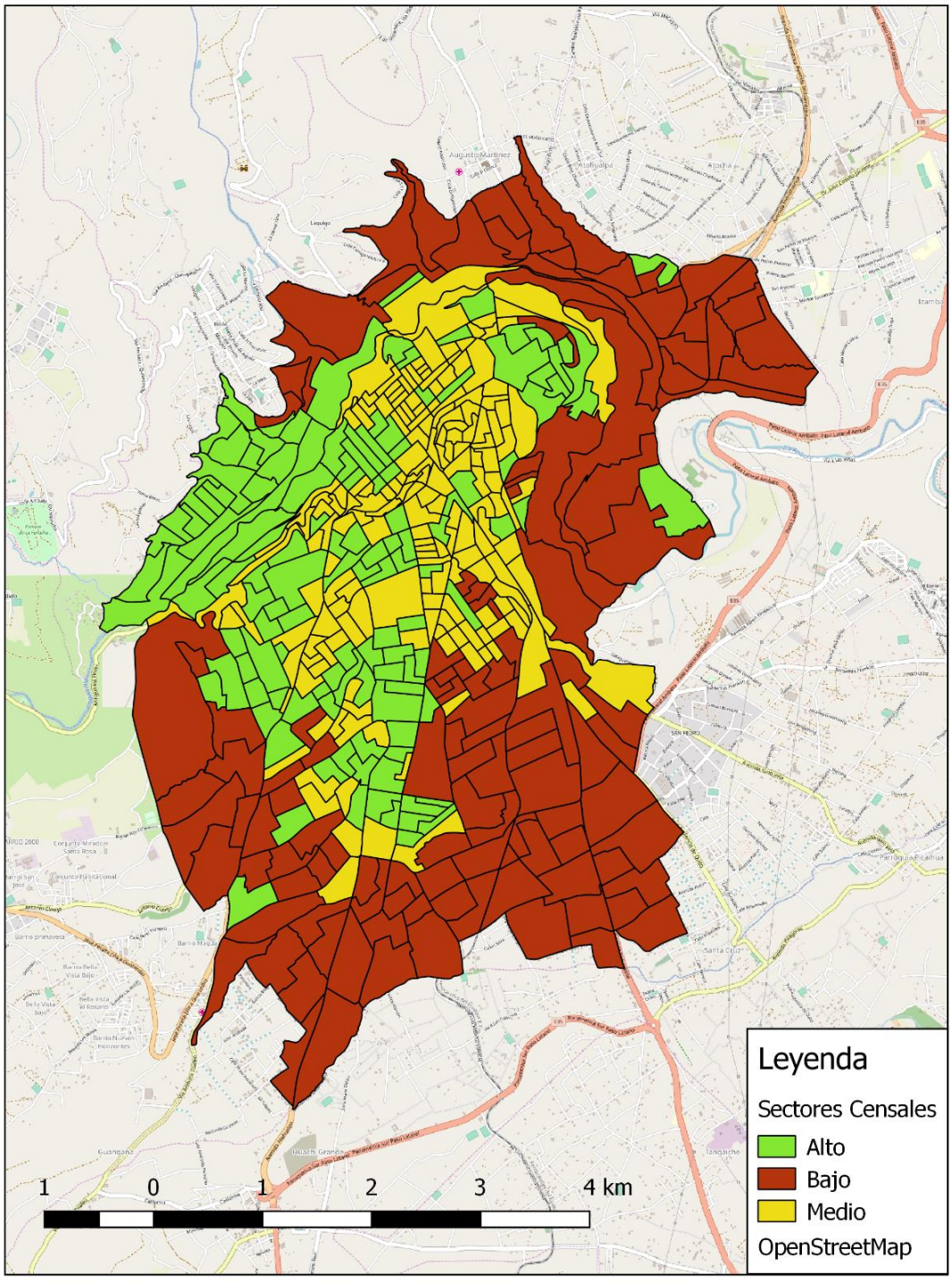


Figura 37. Estratificación Sectores Censales.- Ambato

Se ha verificado la consistencia de los clústeres en el área urbana, para el área rural se observa a continuación el mapa de Pichincha, en el que se puede distinguir que el estrato alto se encuentra en las cercanías de la urbe y el estrato bajo en las zonas más dispersas.

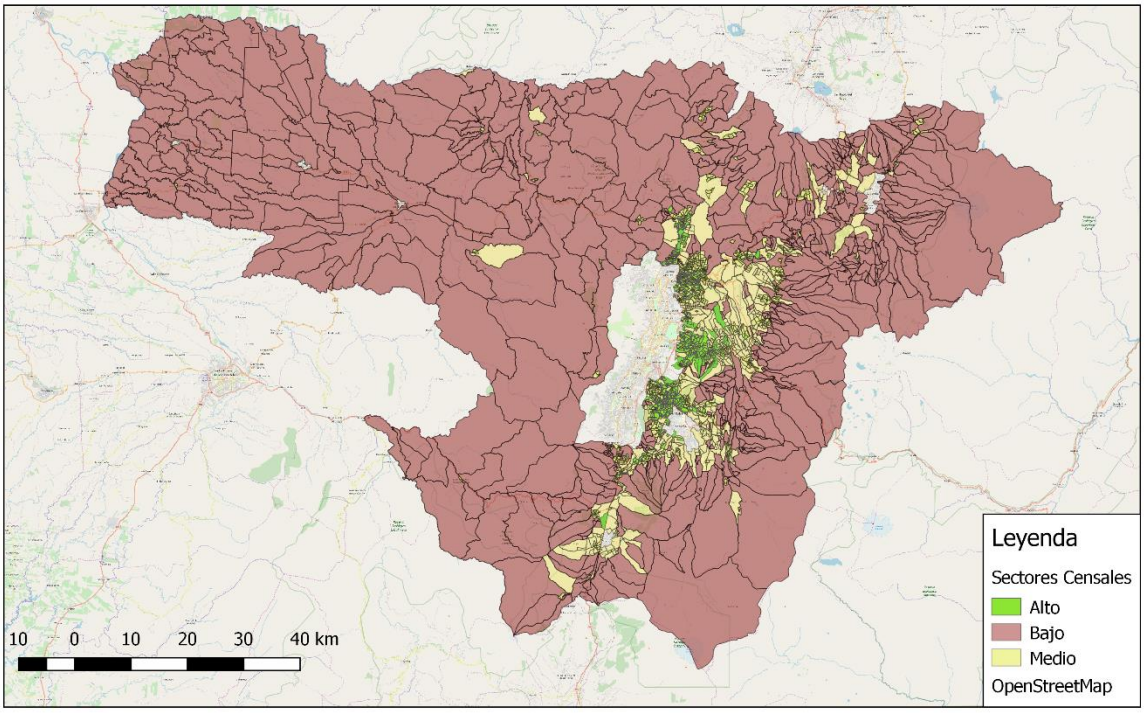


Figura 38. Estratificación Sectores Censales.- Pichincha Rural

6.2. DISTRIBUCIÓN

6.2.1. Planificación de Distribución

El presente trabajo fue desarrollado con el fin aplicarse en el Censo de Población y Vivienda 2020. Se ha demostrado que disminuye el tiempo de ejecución ya que se puede implementar en 10 semanas mientras la anterior fue efectuada en 3 años

Para la distribución de la metodología propuesta es necesario comunicar a las personas involucradas en el proceso.

Directores y coordinadores de las distintas áreas

Este personal deberá analizar las ventajas del cambio de metodología, habilitar los recursos humanos como tecnológicos y difundir e impulsar el proyecto. Además, deben mitigar los posibles riesgos en la implementación y crear planes de contingencia.

Asesores de minería de datos, analistas y desarrolladores

Son los encargados de ejecutar la metodología, para esto deberán capacitarse en las distintas herramientas a utilizar y conocer al detalle cada una de las fases del proceso, como se describe a continuación:

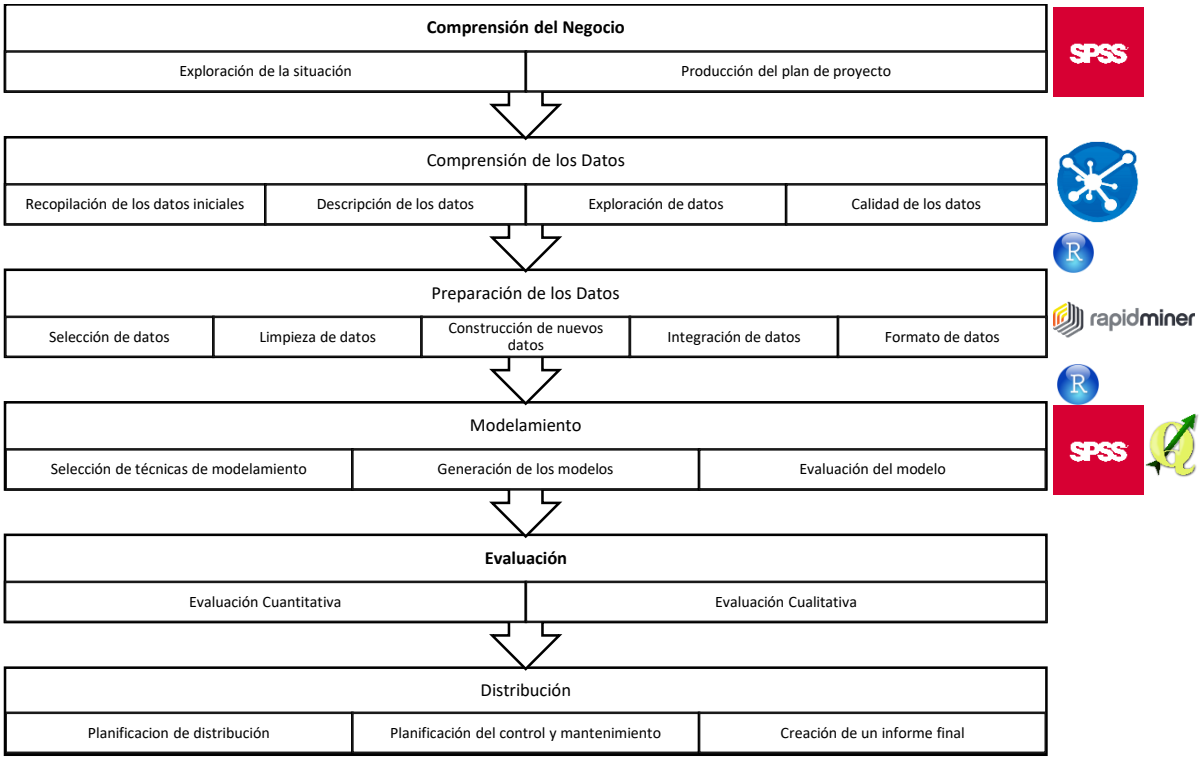


Figura 39. Metodología Propuesta

Con la participación activa de todos los actores en el proceso descrito en la presente metodología se logrará un desempeño óptimo en el próximo Censo de Población y Vivienda al momento de generar los niveles socioeconómicos del país.

CAPÍTULO VII

CONCLUSIONES Y RECOMENDACIONES

7.1. CONCLUSIONES

- Se cumplió los objetivos planteados en su totalidad. Como primer paso se revisó la metodología que usó el INEC, posteriormente se gestionó los datos provenientes del Censo de Población y Vivienda, a continuación, se desarrolló un modelo para la construcción de estratos de la población y por último se evaluó la efectividad de dicho modelo propuesto frente al desarrollado por el INEC en cuanto a tiempo de ejecución y resultados.

- Se disminuyó el tiempo de desarrollo e implementación de la estratificación en la construcción del Marco Muestral de ocho meses a diez semanas gracias a la gestión de datos organizada por una metodología definida.

- La calidad estadística de los estratos socioeconómicos fue mejorada con la metodología propuesta, logrando ampliar la distancia entre clústeres hasta en un 85% en los indicadores evaluados.

- La aplicación de la metodología CRISP-DM permite una documentación completa del proceso de estratificación de manera que pueda ser replicable en un futuro desde la fase de comprensión del negocio hasta la fase de distribución.

7.2. RECOMENDACIONES

- Se sugiere que el Instituto Nacional de Estadísticas y Censos analice la factibilidad de cambiar la metodología actual de estratificación por la propuesta en este documento para el Censo de Población y Vivienda que se llevará a cabo en el 2020, para disminuir el tiempo y mejorar la calidad estadística de los clústeres.

- Se puede realizar otras pruebas de calidad estadística de los clústeres para otros indicadores socioeconómicos que no se consideraron en este estudio.

- Debido al desarrollo social se deberá evaluar la pertinencia e inclusión de características que permitan mejorar la diferenciación entre estratos de acuerdo al desarrollo social que surja en la sociedad.

- Para futuras investigaciones con el fin de disminuir costos de licenciamiento, es importante que se considere la posibilidad de migrar los procesos realizados en SPSS al software libre R.

- Para mejorar la visualización de resultados se podría generar un tablero de control que facilite la interpretación de los clústeres creados.

Bibliografía

Azevedo, A., & Santos, M. F. (2008). *KDD, SEMMA and CRISP-DM: A parallel overview.* Obtenido de IADIS European Conference Data Mining 2008: https://pdfs.semanticscholar.org/7dfe/3bc6035da527deaa72007a27cef94047a7f9.pdf

Berzal, F. (2018). *Clustering.* Obtenido de Universidad de Granada: https://elvex.ugr.es/decsai/intelligent/slides/dm/D3%20Clustering.pdf

BINUS University. (5 de Noviembre de 2014). *Processes in Data Mining*. Obtenido de https://sisbinus.blogspot.com/2014/11/processes-in-data-mining.html

CEPAL. (Enero de 2001). *Hacia un sistema integrado de encuestas de hogares en los países de América Latina.* Obtenido de https://repositorio.cepal.org/bitstream/handle/11362/4707/1/S01010053_es.pdf

Chapman, P., Kerber, R., Clinton, J., Khabaza, T., Reinartz, T., & Wirth, R. (Marzo de 1999). *The CRISP-DM Process Model.* Obtenido de http://www.comp.dit.ie/btierney/BSI/CRISP-DM%20Process%20Model.pdf

Dynamic. (2018). *Gestión de datos (Data Management)*. Obtenido de http://www.dynasolutions.com/servicios-GestionDatos.aspx

García Cambronero, C., & Gómez Moreno, I. (Febrero de 2012). *Algoritmos de aprendizaje knn y kmeans.* Obtenido de http://blogs.ujaen.es/barranco/wp-content/uploads/2012/02/Algoritmos-de-aprendizaje-knn-y-kmeans.pdf

Garre, M., Cuadrado, J. J., & Sicilia, M. A. (9 de Septiembre de 2005). *Comparación de diferentes algoritmos de clustering en la estimación de coste en el desarrollo de software.* Obtenido de http://www.sc.ehu.es/jiwdocoj/remis/docs/GarreAdis05.pdf

IBM. (2012). *Manual CRISP-DM de IBM SPSS Modeler.* Obtenido de ftp://ftp.software.ibm.com/software/analytics/spss/documentation/modeler/15.0/es/CRISP-DM.pdf

IBM SPSS Inc. (14 de Diciembre de 2011). *Manual CRISP-DM de IBM SPSS.* Obtenido de ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/15.0/es/CRISP-DM.pdf

INEC. (2010). *Cuestionario Censal.*

INEC. (Junio de 2012). *Estadística Demográfica en el Ecuador: Diagnóstico y Propuestas.* Obtenido de Ecuador en cifras: http://www.ecuadorencifras.gob.ec/wp-content/descargas/Libros/Demografia/documentofinal1.pdf

INEC. (5 de Noviembre de 2013). *Actualización del diseño muestral de la Encuesta Nacional de Empleo y Dedempleo - ENEMDU.* Obtenido de Ecuador en cifras: http://www.ecuadorencifras.gob.ec//documentos/web-inec/EMPLEO/archivos_ENEMDU/Actualizacion_dise%f1o_muestral_ENEMDU.pdf

INEC. (5 de Noviembre de 2013). *Metodología del diseño muestral de la Encuesta Nacional de Empleo y Desempleo* . Obtenido de http://www.ecuadorencifras.gob.ec/documentos/web-inec/EMPLEO/archivos_ENEMDU/Metodologia_Disenio_Muestral-ENEMDU.pdf

INEC. (2013). *Metodología del diseño muestral de la Encuesta Nacional de Empleo y Desempleo ENEMDU.* Obtenido de http://www.ecuadorencifras.gob.ec/diseno-muestral-2/

INEC. (16 de Octubre de 2013). *Metodología del diseño muestral de la Encuesta Nacional de Empleo y Desempleo ENEMDU.* Obtenido de Ecuador en cifras: http://www.ecuadorencifras.gob.ec//documentos/web-inec/EMPLEO/archivos_ENEMDU/DisenoMuestra.pdf

INEC. (Abril de 2015). *Metodología de la Encuesta de Condiciones de Vida ECV.* Obtenido de http://www.ecuadorencifras.gob.ec/documentos/web-inec/ECV/ECV_2015/documentos/Metodologia/Documento%20Metodologico%20ECV%206R.pdf

INEC. (2015). *Pobreza por necesidades basicas insatisfechas*. Obtenido de Ecuador en Cifras: http://www.ecuadorencifras.gob.ec/pobreza-por-necesidades-basicas-insatisfechas/

INEC. (Marzo de 2016). *Encuesta Nacional de Empleo, Desempleo y Subempleo*. Obtenido de Indicadores Laborales: http://www.ecuadorencifras.gob.ec/documentos/web-inec/EMPLEO/2016/Marzo-2016/Presentacion%20Empleo_0316.pdf

INEC. (2018). *Ecuador en Cifras*. Obtenido de http://www.ecuadorencifras.gob.ec/objetivos-politicas/

INEC. (2018). *Empleo (Encuesta Nacional de Empleo, Desempleo y Subempleo-ENEMDU).* Obtenido de http://www.ecuadorencifras.gob.ec/empleo-encuesta-nacional-de-empleo-desempleo-y-subempleo-enemdu/

INEC. (2018). *Encuesta Nacional de Ingresos y Gastos de los Hogares Urbanos y Rurales*. Obtenido de http://www.ecuadorencifras.gob.ec/encuesta-nacional-de-ingresos-y-gastos-de-los-hogares-urbanos-y-rurales/

León Guzmán, E. (Septiembre de 2016). *Métricas para la validación de Clustering.* Obtenido de Minería de datos: http://www.disi.unal.edu.co/profesores/eleonguz/cursos/md/presentaciones/Sesion13_validacion_Clustering.pdf

Mayo, M. (Marzo de 2016). *The Data Science Process, Rediscovered*. Obtenido de https://www.kdnuggets.com/2016/03/data-science-process-rediscovered.html/2

Ochoa, C. (27 de Febrero de 2015). *Muestreo probabilístico o no probabilístico*. Obtenido de https://www.netquest.com/blog/es/blog/es/muestreo-probabilistico-o-no-probabilistico-ii

Pérez López, C., & Santín Gonzáles, D. (2007). *Minería de datos: técnicas y herramientas.* Madrid: Thomson.

Piatetsky-Shapiro, G. (1996). From data mining to knowledge discovery in databases. *AI magazine*, 37.

Salas L., E. (Mayo de 2018). *Análisis de la estratificación de niveles socioeconómicos de Ecuador.* Obtenido de Revista Contribuciones a las Ciencias Sociales: https://www.eumed.net/rev/cccss/2018/05/niveles-socioeconomicos-ecuador.html

SIISE. (2013). *Tasa de participacion Laboral Bruta*. Obtenido de Sistema Integrado de indicadores sociales del Ecuador: http://www.siise.gob.ec/siiseweb/PageWebs/Empleo/ficemp_T05.htm

SIISE. (2013). *Tasa global de participación laboral*. Obtenido de Sistema Intergado de Indicadores Sociales del Ecuador: http://www.siise.gob.ec/siiseweb/PageWebs/RES/Empleo/ficemp_T07.htm

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a Standard Process Model for Data. 11.